sciendo

Towards fair AI in Estonia's public service: discussing and disseminating bias prevention in automated decision-making

Kristi Joamets

Department of Law
Tallinn University of Technology
Ehitajate tee 5
Tallinn 19086, Estonia
kristi.joamets@taltech.ee

Abstract: Estonia is recognized as a pioneer state in digitalizing its public services, particularly through its integration of artificial intelligence (AI) systems. However, the lack of tailored strategies and tools to address biases in AI-based decision-making in Estonian public services processes poses significant ethical challenges. This study explores the strategies and tools available to identify and prevent bias in automated decision-making (ADM) processes and ways of their dissemination. It draws from the European Commission-funded EquiTech project, which seeks to ensure fairness in algorithmic decision-making in public services. Employing mainly qualitative methodology, the study draws on literature review, document analysis and discussions conducted in the framework of the EquiTech project. Through this exploration and dissemination, the study contributes to the conceptualization of tailored strategies and tools in Estonia, ensuring fairness in ADM processes within public services, promoting equitable digital governance and mitigating discriminatory outcomes.

Keywords: *AI bias, automated decision-making, bias mitigation strategies, bias mitigation tools, EquiTech project, Estonian public services, fairness in AI*

https://doi.org/10.2478/bjes-2025-0037 *TalTech Journal of European Studies*, 15(3(43))
Tallinn University of Technology (ISSN 2674-4619)

1. Introduction

Estonia is widely recognized as a pioneer in the digitalization of public services, with its nearly full online access to government services and high-performance scores in the European Union's Digital Decade report—98.9 for digital public services for businesses and 95.8 for citizens in 2024 (Vatsa and Chhaparwal, 2021; European Commission, 2024). The integration of AI in Estonian public services represents a rapidly growing trend, with AI adoption among businesses more than doubling from 5.19% in 2023 to 13.89% in 2024, and over 130 AI applications deployed in the public sector by 2025 (European Commission, 2024; Thelle *et al*., 2024), including initiatives such as the Bürokratt virtual assistant for streamlined citizen interaction (Dreyling III *et al*., 2024), personalized medicine programmes in healthcare (Metsallik and Ross, 2021) and smart traffic management systems in transportation (Edler, 2019; Kalda *et al*., 2021).

This digitalization of government services involves the strategic use of AI systems in administrative processes and decision-making, as outlined in Estonia's *Digital Agenda 2030* (Ministry of Economic Affairs and Communications (Estonia), 2021). These systems are designed to promote decisions that are free from stereotypes, more uniform, objective and accurate, thereby fostering greater fairness, as evidenced by aiming for "zero bureaucracy" and proactive, personalized services. At present, no state service in Estonia relies entirely on AI-based decision-making; in all cases, a human remains involved, at least to some extent, in every final decision (Dreyling *et al*., 2023; Nõmmik, 2025), aligning with ethical principles, policies and legislative rules from EU's regulatory framework on AI (AI HLEG, 2019; European Commission, 2020; 'Regulation (EU) 2024/1689', 2024). Although, Estonia's AI systems are designed to ensure AI governance in human-centric manner by aligning with EU AI regulatory framework, challenges remain in fully mitigating potential biases within the automated decision-making (ADM) process.

This process encompasses multiple interconnected stages—including data collection, algorithm design, model training, testing, validation and deployment—where biases can inadvertently infiltrate less visible components, such as imbalanced datasets or flawed feature selection, subtly skewing final outcomes toward unfair or discriminatory results (Singhi and Liu, 2006; Amri *et al*., 2018; Vetrò *et al*., 2021). Detecting bias in these ADM decisions is further compounded by the inherent lack of transparency

in many AI systems, often characterized as "black boxes" due to their complex neural architectures, which can obscure decision pathways and leave developers, deployers and end-users unaware of embedded risks or specific biases, thereby perpetuating unfairness in areas like public service allocation (Vorras and Mitrou, 2021; Franzoni, 2023; Pappachan *et al*., 2024). To date, Estonia lacks bespoke strategies and tools to systematically identify and prevent bias in ADM processes, though ongoing efforts align with the objective of EU's regulatory framework on AI, which is to institutionalize and conceptualize trustworthiness as an attribute of AI. The EquiTech project[1] is one of them. Funded by the European Commission, it aims to develop strategies and tools for identifying and preventing bias in ADM processes, with a focus on Estonia's public sector, and to disseminate these approaches.

First, it is essential to have the foundational understanding of how AI systems function and how decisions are made through encompassing processes such as data ingestion, algorithmic computation via machine learning models (e.g., neural networks or decision trees and probabilistic or rule-based output generation, which can inadvertently amplify human-like errors unless scrutinized (Bose, 2019; Mian *et al*., 2024; Dave *et al*., 2025; Chen *et al*., 2025). This understanding must extend to the concepts of bias defined as inherent assumptions or imbalances in data and systems that distort results toward unfairness and bias that involves disparate treatment or impact on protected groups (e.g., based on race, gender or socioeconomic status), with AI often exacerbating these through reflected societal inequalities, flawed training datasets or opaque algorithmic designs (Lobo, 2022; Jain and Menon, 2023; Ali *et al*., 2025). To combat these issues, effective tools and strategies should be conceptualized, including bias detection frameworks (DeAlcala *et al*., 2023; Tellez *et al*., 2023; Marripudugala, 2025), bias mitigation techniques (González-Sendino *et al*., 2024; Choi *et al*., 2025) and adversarial debiasing (Senthil *et al*., 2025). In addition, regulatory-driven approaches, such as risk assessments, are crucial (DeAlcala *et al*., 2023). Along with strategies and tools, the responsibility of the system's range stakeholders is also critical. It is to address AI bias spanning multiple stakeholders: designers and developers of AI systems must ensure fairness and transparency during design, development and testing (Radanliev, 2025); deployers are accountable for lawful, non-discriminatory deployment (Constantino, 2025); lawmakers should establish ethical regulations and

[1] Further on the project see EquiTech – Ensuring Fairness in Public Sector Algorithmic Decision-Making from the European Commission programme 'Citizens, Equality, Rights and Values' (CERV-2023-EQUAL) for the period May 1, 2024 – April 30, 2026) at https://www.volinik.ee/en/activities/ec-projects.html.

oversight mechanisms (Göksal and Solarte-Vásquez, 2024), as seen in the EU AI regulatory framework (AI HLEG, 2019; European Commission, 2020; 'Regulation (EU) 2024/1689', 2024); and end-users bear the responsibility to interpret AI outputs objectively and intervene when necessary (Porter *et al.*, 2025).

These principles constitute the conceptual foundation of the EquiTech project, which employs mainly qualitative methodological approach, incorporating a literature review of the technical underpinnings of AI systems, the challenge of bias, Estonia's application of AI in the public sector and document analysis on EU AI regulatory framework. This study focuses on discussing the outputs of the EquiTech project, including analyses of bias in AI-based decision-making—such as assumptions in data leading to unfair outcomes, classifications of bias types, reviewing the case studies in the public sector related to AI systems, the concepts of bias in AI and the sociotechnical nature of bias. Furthermore, the study discusses and disseminates strategies and tools to prevent bias in ADM processes such as representative training data, bias verification tests and continuous monitoring alongside compliance and conformity with regulatory requirements.

The subsequent sections are structured as follows: the second section presents the technical foundation of AI technology, the third explores bias in AI-based decision-making, providing a detailed explanation with real-life cases alongside the concepts of bias in ADM process. The fourth section analyzes accountability aspect, identifying parties responsible for bias mitigation and outlining potential strategies and tools to prevent discriminatory outcomes. The final section concludes the discussion, highlights key limitations and proposes directions for future research.

2. The technical foundation of AI technology

AI encompasses a broad array of technologies designed to simulate human-like cognitive processes, such as perception, reasoning, learning and decision-making, through computational systems (Russell and Norvig, 2021). At its core, AI uses algorithms—step-by-step procedures for calculations—to process data and generate outputs, often in the form of predictions, classifications or recommendations. Within the domain of ADM process, the most relevant subset is machine learning (ML), a data-driven approach in which systems improve performance on tasks by learning patterns from large datasets

without explicit programming for every scenario (Dineva and Atanasova, 2020). ML models, including decision trees, support vector machines and deep neural networks, constitute the backbone of these systems, enabling applications such as Estonia's virtual assistant Bürokratt or predictive analytics in healthcare (Dreyling III *et al*., 2024; Metsallik and Ross, 2021). The technical foundation of AI technology can be understood through its interconnected development pipeline, which includes data collection, algorithm design, model training, testing, validation and deployment (Kim, 2019; Issaro *et al*., 2023; Steidl *et al*., 2023).

The process begins with data collection, where raw inputs—such as from structured datasets (e.g., tabular records of citizen interactions) or unstructured sources (e.g., text from public queries or images from surveillance)—are gathered to train models (Kim, 2019). Data serves as the foundational "fuel" for AI, with features (input variables such as age, income or location) and labels (desired outputs, such as eligibility for services) defining the learning objectives (Woolf *et al*., 2023). Following collection, algorithm design involves selecting and configuring the model's architecture. For instance, supervised learning algorithms, commonly used in ADM processes, in involving map inputs to outputs using labeled data (e.g., neural networks trained to classify loan applications based on historical approvals). Unsupervised methods, in contrast, cluster data without labels, such as segmenting citizen needs for personalized services. Design choices, like feature selection—deciding which variables to include—can inadvertently embed flaws; excluding or mishandling proxies for protected attributes (e.g., zip codes correlating with ethnicity) risks amplifying societal inequalities (Zemmari and Benois-Pineau, 2020; Sen and Das, 2023).

Model training follows, during which algorithms iteratively adjust internal parameters (e.g., weights in neural networks) to minimize errors between predictions and actual outcomes, often using optimization techniques such as gradient descent (Heo and Kim, 2022). Testing and validation then evaluate the model's performance on held-out data, employing metrics such as accuracy, precision, recall or fairness-specific indicators such as demographic parity to detect disparities across groups (Raja *et al*., 2024). Cross-validation techniques, such as k-fold splitting, help ensure robustness, but inadequate testing datasets may fail to capture edge cases, allowing subtle biases in feature interactions to persist (Fontanari *et al*., 2022).

Deployment represents the final step in AI technology. It integrates the model into operational systems, often with real-time inference engines that

process new inputs and generate decisions, supported by monitoring tools for drift detection, where the model's performance degrades over time due to changing data distributions (Kore *et al.*, 2024).

3. Bias in AI-based decision-making, its conceptual foundation and real-life cases

Bias in ADM processes refers to systematic assumptions or flaws embedded within AI systems, their underlying algorithms or the datasets used for training, which produce unfair, distorted or discriminatory outcomes (Cross *et al.*, 2024; Bhattacharya *et al.*, 2024; Hofmann, 2025). These biases often stem from incomplete, unrepresentative or historically skewed data, as well as from human influences during system design and implementation, such as unconscious prejudices held by developers or the perpetuation of outdated societal norms (Bhattacharya *et al.*, 2024; Ruschemeier and Hondrich, 2024). For instance, if training data predominantly features certain demographics, the resulting AI model may prioritize those groups. This phenomenon mirrors broader societal inequalities, where power imbalances and historical injustices are encoded into technology, causing disadvantages for marginalized or underrepresented groups based on protected attributes like race, gender, socioeconomic status, nationality, skin color, age, disability, sexual orientation or neurodiversity. In the context of ADM, bias can manifest when algorithms analyze vast amounts of data to inform high-stakes decisions across various sectors, including recruitment (Soleimani *et al.*, 2022), public services (Alrawahna *et al.*, 2025) or law enforcement (Zhaltyrbayeva *et al.*, 2025). If these are left unaddressed, they also create feedback loops, where discriminatory outputs generate new biased data (Yang *et al.*, 2023), perpetuating a cycle of inequity that undermines fairness in AI systems, erodes social justice and potentially violates ethical standards and legal rules.

3.1 Conceptual foundation of bias

Bias can lead to discrimination. Accordingly, discussions often address not only different forms of discrimination but also different types of biases. At its core, bias involves treating individuals or groups less favorably based on protected characteristics under relevant laws. Different countries define these protections variably; in Estonia, for instance, the *Constitution of*

the Republic of Estonia (1992) prohibits unequal treatment on unlimited grounds, while the *Equal Treatment Act (Estonia)* (2008) and *Gender Equality Act (Estonia)* (2004) specify protections against discrimination based on ethnicity, race, colour, religion, age, disability, sexual orientation and gender.

Bias in an ADM process can take several forms: direct bias, indirect bias and proxy bias. Direct bias occurs when a person is explicitly treated less favorably than another individual or group in a comparable situation due to a protected characteristic (Adams-Prassl *et al*., 2023; Carter, 2024). For example, this might involve an algorithm programmed to outright reject loan applications based solely on the applicant's gender, ethnicity or age. In such cases, the system ignores other qualifying factors like credit history or income stability. This form of bias is overt and intentional in its design. It directly violates anti-discrimination laws by embedding prejudicial rules into the decision-making logic. As a result, it can lead to immediate legal challenges. Direct bias lacks any veneer of neutrality, which makes it easier to identify. However, it is no less harmful as it perpetuates stereotypes and excludes qualified candidates from opportunities in sectors like finance, employment or housing.

Indirect bias arises from neutral rules, criteria or practices that disproportionately disadvantage certain groups unless they can be justified by a legitimate aim achieved through appropriate and necessary means (Aytekin, 2023; Carter, 2024). For instance, this could manifest in a loan approval model that prioritizes factors such as income levels and educational attainment. In doing so, it excludes individuals from lower-income communities, which are often correlated with specific ethnic or socioeconomic groups. This type of bias is subtler than direct bias because it appears objective on the surface. However, it still results in systemic unfairness by amplifying existing disparities without explicit reference to protected characteristics. Therefore, it can evade immediate detection, requiring deeper analysis of outcomes across demographics to uncover. Indirect bias often stems from well-intentioned designs that fail to account for broader social contexts.

Proxy bias, arguably the most common in AI, involves using non-protected attributes that correlate with protected ones, leading to biased outcomes (Sogancioglu *et al*., 2023; Daish *et al*., 2023), such as relying on zip codes to indicate poverty or ethnicity, occupations or first name with cultural implications or employment history in terms of length or interruptions that

disadvantage women due to parental leave or family responsibilities. This form of bias operates indirectly by substituting innocuous variables for sensitive traits, allowing discrimination to persist under the guise of data-driven neutrality (Sogancioglu *et al*., 2023). It is particularly insidious in AI systems because algorithms may learn these correlations from historical data without explicit programming (Vetrò, 2021). As a result, it can perpetuate cycles of inequality, such as in hiring algorithms that favor certain cultural names or in insurance pricing models that charge higher rates based on residential proxies for race. Detecting proxy bias requires scrutiny of feature selections and correlations in models (Singhi and Liu, 2006; Gonçalves *et al*., 2025).

In addition to these forms of bias in ADM, it is useful to categorize bias into broader types that reflect its origins and mechanisms, providing a framework for understanding how it infiltrates AI systems at various levels.

Statistical bias in an ADM process arises from fundamental errors in the processes of data collection, utilization, interpretation or validation that can skew AI models toward inaccurate or unfair outcomes. This type of bias often occurs when datasets are imbalanced, lacking diversity or sufficient representation across different groups, leading the algorithm to prioritize patterns from overrepresented segments of the population (Kulkarni *et al*., 2020; Gonuguntla *et al*., 2025). For instance, if a facial recognition system is trained predominantly on images of light-skinned individuals, it may perform poorly on darker skin tones, resulting in higher error rates for underrepresented demographics and favoring predictions that align with the majority data. Such errors are not necessarily intentional but stem from methodological flaws, such as inadequate sampling techniques or failure to account for outliers, which can amplify disparities in applications ranging from healthcare diagnostics to predictive analytics. Addressing statistical bias requires rigorous data auditing, balanced sampling strategies and validation methods to ensure that models generalize effectively across all relevant populations (Taylor *et al*., 2008; Bakker, 2010; Debray *et al*., 2017; Seraj *et al*., 2022; Gonuguntla *et al*., 2025).

Human bias refers to the individual or collective prejudices introduced by developers, data scientists or stakeholders during the design and implementation of AI systems, often manifesting unconsciously through subjective decisions in feature selection, weighting or model tuning. This can happen when personal assumptions influence how factors are prioritized (Pulivarthy and Whig, 2024; Aninze and Bhogal, 2024); for example, an

algorithm might use vocabulary style as a proxy indicator for the socioeconomic background, inadvertently disadvantaging certain racial or gender groups based on linguistic patterns correlated with cultural differences. Developers may embed their own worldviews into the system without realizing it, such as overemphasizing certain metrics that reflect societal stereotypes, leading to biased outputs in hiring tools or recommendation engines. This form of bias highlights the human element in AI, where cognitive shortcuts or groupthink in teams can perpetuate inequities, emphasizing the need for diverse development teams, bias-awareness training and ethical guidelines to mitigate these influences from the outset (Xivuri and Twinomurinzi, 2023; Verma *et al*., 2024).

Systemic bias, also known as institutional or historical bias, is deeply rooted in the broader organizational practices, historical injustices or prevailing social norms that shape the data and frameworks used in ADM, often embedding long-standing inequalities such as racism or sexism into technological systems. This occurs when historical data, which captures past discriminatory patterns—such as the underrepresentation of women in professional roles—is used to train models, thereby propagating those same disadvantages into future predictions and decisions (Fountain, 2022; Vitek *et al*., 2023). For example, a credit scoring algorithm drawing from decades-old financial records might penalize groups historically excluded from economic opportunities, reinforcing cycles of poverty or exclusion in modern lending practices. Unlike more isolated forms of bias, systemic bias reflects institutional structures and cultural contexts, making it pervasive and challenging to eradicate without systemic reforms, including diversifying data sources, incorporating historical context in model evaluations and aligning AI deployments with equity-focused policies to break the chain of inherited prejudices.

3.2 Real-life cases of bias in ADM

Numerous real-life cases illustrate the issues of bias in an ADM process, demonstrating how flaws in AI systems can lead to discriminatory outcomes across various sectors. These examples highlight the real-world consequences of unaddressed biases, often caused by imbalanced training data, flawed algorithmic designs or systemic inequalities, and underscore the need for mitigation strategies and tools.

One frequently cited technology which is affected by bias is facial recognition, particularly concerning disparities in skin color. AI systems frequently

struggle with darker skin tones because their training datasets are dominated by images of lighter-skinned individuals, resulting in higher error rates for underrepresented groups. A study by the American Civil Liberties Union (ACLU) tested Amazon's Rekognition software by matching photographs of US Congress members against criminal databases. The results were alarming: the system misidentified 28 members as criminals, with error rates exceeding 40% for Black and Hispanic individuals, compared to near-100% accuracy for those with white skin. This revelation prompted widespread calls for Amazon to halt sales of the technology to police forces, citing severe risks of racial bias and potential violations of human rights (Snow, 2018).

Bias also manifests in employment-related ADM processes, such as recruitment systems and job advertising platforms. For instance, in the mid-2010s, Amazon developed an AI-based hiring tool that was trained on historical resumes predominantly from male applicants. The system learned to downgrade women's applications by penalizing gender-specific terms like "women's chess club", ultimately leading to its discontinuation in 2018 due to inherent gender discrimination (Dastin, 2022). This issue extends to large language models such as OpenAI's GPT-4 and Microsoft's Copilot, which, trained on male-dominated open data, can perpetuate gender biases in generated content, further perpetuating stereotypes in professional opportunities (Bano *et al*., 2025).

Financial and welfare systems further exemplify bias in ADM processes, where decisions can have economic impacts. In Finland, an AI loan approval system denied applications based on factors like nationality (viewing Finns as poorer than Swedes), residence in rural areas (associated with lower wealth) and gender, leading to unfair exclusions (Matzat, 2019). Beyond lending, the 2020 Dutch child benefits fraud detection scandal involved an AI system with erroneous data assumptions that falsely accused thousands of families of fraud, causing significant confusion, reputational damage and financial hardship; the government later apologized and initiated reforms to prevent future errors (Amnesty International, 2021).

Other diverse cases reveal the broad scope of bias in ADM processes, affecting vulnerable populations in everyday and high-stakes scenarios. Emotion recognition AI often misinterprets neurodiverse individuals, leading to inaccurate assessments in contexts such as hiring or security (Said *et al*., 2026). Airport verification systems frequently disadvantage transgender, intersex, non-binary or gender-non-conforming individuals by failing to effectively handle non-traditional data inputs (Reddy-Best and Olson, 2020).

Insurance pricing models may inflate rates based on country of birth as a proxy for risk, while deepfake technologies enabled by AI can facilitate fraud with biased implications (Arku *et al*., 2020). Biometric systems for public services sometimes exclude users with darker skin tones, compounding access inequalities (Kolberg *et al*., 2023). Collectively, these examples demonstrate how bias in ADM processes leads to unfair outcomes, often in critical areas such as justice, employment, finance and public welfare, emphasizing the urgent need for ethical AI development to safeguard equity and trust.

4. Accountability and mitigation of bias in AI systems

In legal contexts, determining responsibility for violations, such as bias arising from AI-based decisions, is crucial. The inherent opacity and complexity of AI systems—stemming from dynamic interactions between hardware, software and evolving components—complicate the establishment of accountability rules. Since users typically procure rather than develop these systems, accountability must be linked to specific stages of data processing where discriminatory elements may originate.

Accountability for AI systems spans multiple levels:

- *Designers and developers* must prioritize fairness and transparency to prevent biases, such as correcting inaccuracies in facial recognition algorithms that disadvantage dark-skinned individuals (Radanliev, 2025);
- *Deployers* bear the responsibility for ensuring lawful, non-discriminatory application, including reevaluating lending algorithms that unfairly impact ethnic groups (Constantino, 2025);
- *Lawmakers* are tasked with enacting regulations, exemplified by the EU AI regulatory framework, which mandates transparency and accountability for high-risk systems (AI HLEG, 2019; European Commission, 2020; 'Regulation (EU) 2024/1689', 2024).
- *End-users* must objectively interpret AI outputs and apply decisions without perpetuating bias (Porter *et al*., 2025).

Mitigating AI bias requires proactive management during implementation, incorporating effective strategies and tools such as bias detection frameworks, bias mitigation techniques and adversarial debiasing. Bias detection

frameworks (DeAlcala *et al*., 2023; Tellez *et al*., 2023; Marripudugala, 2025) are essential for identifying and quantifying biases in AI systems before they lead to discriminatory outcomes. They typically involve systematic audits of datasets, models and outputs to uncover imbalances, such as underrepresented groups in training data or skewed predictions favoring certain demographics. For instance, they might employ metrics like disparate impact ratios or fairness-aware evaluations to measure how an algorithm performs across protected attributes such as gender or ethnicity. By integrating them early in the development pipeline, developers can pinpoint sources of bias—whether from historical data prejudices or algorithmic assumptions—and generate reports that guide corrective actions, ultimately fostering more equitable AI deployments.

Bias mitigation techniques (González-Sendino *et al*., 2024; Choi *et al*., 2025) encompass a range of techniques designed to reduce or eliminate identified biases during the model's training or post-processing stages. Common approaches include data augmentation, where underrepresented samples are synthetically generated to balance datasets or reweighting algorithms that assign higher importance to minority groups in the learning process. Other techniques involve fairness constraints in optimization objectives, ensuring that the model minimizes accuracy disparities without sacrificing overall performance. They are proactive, allowing organizations to refine AI systems iteratively, such as adjusting a hiring algorithm to prevent gender-based favoritism, thereby promoting both transparency and accountability in real-world applications.

Adversarial debiasing (Senthil *et al*., 2025) represents a more advanced approach that uses adversarial training to make AI models robust against biases. In this approach, a primary model is trained alongside an adversarial network that attempts to predict protected attributes (e.g., race or age) from the model's outputs, forcing the main model to produce representations that obscure these sensitive features while maintaining utility. This game-theoretic method helps in creating "bias-blind" embeddings, which are particularly useful in scenarios such as facial recognition or credit scoring where subtle correlations could perpetuate inequalities. By continuously challenging the model during training, adversarial debiasing enhances generalization and fairness, though it requires careful tuning to avoid over-correction or reduced efficiency.

Additionally, regulatory-driven approaches such as risk assessments are crucial for identifying and addressing potential biases (Hartung *et al*.,

2025). Development must align with ethical principles and legal frameworks to avoid unfair outcomes, while ongoing monitoring and updates prevent emerging biases (Ricaurte and Zasso, 2023; Lacmanovic and Skare, 2025). Effective supervision further reinforces these efforts (Henzinger *et al*., 2023). To operationalize regulatory elements, states play a key role in raising awareness among decision-makers about AI's potential for discrimination, including how to detect and prevent it (Goel and Nelson, 2025). In Estonia, this could involve legal mandates, such as requiring notification to individuals affected by algorithmic decisions about AI's involvement and extent, along with guidance on recourse if discrimination is suspected. Developers should adhere to state guidelines throughout the AI lifecycle to preempt biases at every stage.

Central EU instruments, including the *General Data Protection Regulation* (*GDPR*) ('Regulation (EU) 2016/679', 2016) and the EU *AI Act* ('Regulation (EU) 2024/1689', 2024), emphasize human control. Article 22 of the *GDPR* mandates human intervention in automated decisions, while Article 14 of the EU *AI Act* requires human oversight, encompassing system-wide controls (design, implementation, testing auditing) and case-specific reviews. Additionally, the EU *AI Act* outlines requirements for high-risk systems under Articles 9 (risk management), 10 (data governance), 11 (technical documentation), 13 (transparency) and 15 (accuracy and robustness) (Göksal *et al*., 2025). For discrimination contexts, rules for medium- or low-risk systems remain unclear.

Organizations must fulfill human oversight duties by understanding high-risk AI operations, responding to issues, recognizing automated biases in decisions affecting individuals and being ready to suspend usage if risks emerge. While aligning with EU AI regulatory framework across EU Member States, including in Estonia, is still evolving, additional national regulations and guidelines are essential to clarify the rights, obligations and responsibilities for developers, deployers and end-users in preventing bias and discrimination in automated decision-making.

5. Conclusion

Estonia's pioneering role in digitalizing public services, with nearly full online accessibility and high EU Digital Decade scores, has accelerated the integration of AI. While these AI systems enhance objectivity and uniformity in administrative decision-making, they remain vulnerable to biases introduced at various stages, from data collection to deployment. Estonia currently lacks tailored strategies and tools to systematically identify and prevent bias in the ADM process. The objective of the EquiTech project is to explore strategies and tools for identifying and preventing bias in ADM processes, with a focus on Estonia's public sector, and to disseminate these approaches to promote trustworthy AI governance. The purpose of this paper was to discuss the outputs of the EquiTech project, thereby contributing to the efforts to conceptualize these strategies and tools specifically for the Estonian AI ecosystem in the public sector.

This article discussed the technical foundation of AI technology and analyses of bias in AI-based decision-making, including the concept of bias in AI, how assumptions in data led to unfair outcomes, classifications of bias types and relevant case studies. It also examined strategies and tools such as bias detection frameworks, bias mitigation techniques, adversarial debiasing, continuous monitoring and compliance with regulatory requirements, thereby contributing to a human-centric and equitable AI ecosystem in Estonia. However, a key limitation of this study is that it primarily discusses these elements without fully conceptualizing strategies and tools tailored to the Estonian ecosystem, and it lacks empirical evaluation to assess the success or effectiveness of any such conceptualization in practice.

Future research should prioritize conceptualization studies that develop tailored strategies and tools for bias mitigation within the Estonian AI ecosystem, particularly in the public sector, by integrating sociotechnical perspectives that account for local cultural, regulatory and infrastructural contexts. This could involve designing framework/sociotechnical AI models that blend technical innovations—such as advanced bias detection algorithms—with social elements, such as stakeholder engagement and ethical training for public administrators, to ensure holistic implementation. Furthermore, empirical studies, including longitudinal evaluations and controlled experiments in Estonian public AI systems, are essential to assess the practical success, scalability and impact of these conceptualized approaches on reducing biases and enhancing trustworthiness.

Acknowledgment

This research was funded by the European Commission EquiTech project (Grant agreement number: 101144709).

Kristi Joamets, PhD is a senior lecturer at TalTech Law School, Tallinn University of Technology. Dr Joamets earned her doctoral degree in public administration and has published more than 20 academic works, including articles in peer-reviewed journals, book chapters and conference papers. Her main areas of research include labour law, equality law, education law and social law.

References

Adams-Prassl, J., Binns, R. and Kelly-Lyth, A. (2023) 'Directly discriminatory algorithms', *Modern Law Review*, 86(1), pp. 144–175. Available at: https://doi.org/10.1111/1468-2230.12759

AI HLEG (2019) *Ethics guidelines for trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. Brussels: European Commission. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed: 3 October 2025).

Ali, S.I., Albadoo, S.F. and Al Mubarak, M. (2025) 'Ethical ramifications and remedial approaches on bias in artificial intelligence', in S.I. Ali *et al.* (eds) *Studies in Systems, Decision and Control*, 237. Springer Science and Business Media Deutschland GmbH, pp. 95–108. Available at: https://doi.org/10.1007/978-3-031-86708-8_8

Alrawahna, A.S., Alzghoul, A. and Awad, H. (2025) 'The impact of artificial intelligence on public sector decision-making: benefits, challenges, and policy implications', *International Review of Management and Marketing*, 15(5), pp. 125–138. Available at: https://doi.org/10.32479/irmm.19419

Amnesty International (2021) 'Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms', *Amnesty International*, 25 October. Available at: https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/ (Accessed: 3 October 2025).

Amri, A.A., Ismail, A.R. and Zarir, A.A. (2018) 'Comparative performance of deep learning and machine learning algorithms on imbalanced handwritten data', *International Journal of Advanced Computer Science and Applications*, 9(2), pp. 258–264. Available at: https://doi.org/10.14569/IJACSA.2018.090236

Arku, D., Doku-Amponsah, K. and Howard, N.K. (2020) 'A Markov-modulated tree-based gradient boosting model for auto-insurance risk premium pricing', *Risk and Decision Analysis*, 8(1–2), pp. 1–13. Available at: https://doi.org/10.3233/RDA-180050

Aytekin, A.B. (2023) 'Algorithmic bias in the context of European Union anti-discrimination directives', *CEUR Workshop Proceedings*, 3442.

Bakker, D. (2010) 'Language sampling', in J.J. Song (ed) *The Oxford handbook of linguistic typology*. Oxford: Oxford University Press, pp. 100–128. Available at: https://doi.org/10.1093/oxfordhb/9780199281251.013.0007

Bano, M., Gunatilake, H. and Hoda, R. (2025) 'What does a software engineer look like? Exploring societal stereotypes in LLMs', in *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pp. 173–184. Available at: https://doi.org/10.1109/ICSE-SEIS66351.2025.00023

Bhattacharya, A., Stumpf, S. and Verbert, K. (2024) 'Representation debiasing of generated data involving domain experts', in *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24)*, pp. 516–522. Available at: https://doi.org/10.1145/3631700.3664910

Bose, B.K. (2019) 'Artificial intelligence applications in renewable energy systems and smart grid: some novel applications', in B.K. Bose (ed) *Power electronics in renewable energy systems and smart grid: technology and applications*. Hoboken, NJ: Wiley, pp. 625–675. Available at: https://doi.org/10.1002/9781119515661.ch12

Carter, C. (2024) 'Why the algorithmic recruiter discriminates: the causal challenges of data-driven discrimination', *Maastricht Journal of European and Comparative Law*, 31(3), pp. 333–359. Available at: https://doi.org/10.1177/1023263X241248474

Chen, W., Cauteruccio, F., Li, Y., Zheng, J. and Liu, W. (2025) 'Three technical routes of AI', in W. Hussain, L.M. López, M. Sahni, Z.-S. Chen and J. Liu (eds) *Cutting-edge artificial intelligence: advances and implications in real-world applications*. Boca Raton: CRC Press, pp. 35–54. Available at: https://doi.org/10.1201/9781032632483-3

Choi, Y., Hong, J., Lee, E., Kim, J. and Kim, S. (2025) 'Enhancing fairness in financial AI models through constraint-based bias mitigation', *Journal of Information Processing Systems*, 21(1), pp. 89–101. Available at: https://doi.org/10.3745/JIPS.01.0111

Constantino, J. (2025) 'Accountable AI: it takes two to tango', in R. Gsenger and M.-T. Sekwenz (eds) *Digital Decade: how the EU shapes digitalisation research*, Vol. 3. Baden-Baden: Nomos Verlagsgesellschaft mbH und Co., pp. 95–114.

Constitution of the Republic of Estonia (1992) *RT*, 26, 349. Available at: https://www.riigiteataja.ee/en/eli/521052015001/consolide (Accessed: 3 October 2025).

Cross, J.L., Choma, M.A. and Onofrey, J.A. (2024) 'Bias in medical AI: implications for clinical decision-making', *PLoS Digital Health*, 3(11), article e0000651. Available at: https://doi.org/10.1371/journal.pdig.0000651

Daish, P., Roach, M. and Dix, A. (2023) 'Raising user awareness of bias-leakage via proxies in AI models to improve fairness in decision-making', in *Proceedings of the AISB Convention*, pp. 86–88.

Dave, G.S., Pandhare, A.P., Kulkarni, A.P. and Khankal, D.V. (2025) 'Innovative data techniques for centrifugal pump optimization with machine learning and AI model', *PLoS One*, 20(6), article e0325952. Available at: https://doi.org/10.1371/journal.pone.0325952

Dastin, J. (2022) 'Amazon scraps secret AI recruiting tool that showed bias against women', in K. Martin (ed) *Ethics of data and analytics: concepts and cases*. Boca Raton: Auerbach Publications.

DeAlcala, D., Serna, I., Morales, A., Fierrez, J. and Ortega-Garcia, J. (2023) 'Measuring bias in AI models: a statistical approach introducing N-Sigma', in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1167–1172. Available at: https://doi.org/10.1109/COMPSAC57700.2023.00176

Debray, T.P.A., Damen, A.A.A.G., Snel, K.I.E., Ensor, J., Hooft, L. Reitsma, J.B., Riley, R.D. and Moons, K.G.M. (2017) 'A guide to systematic review and meta-analysis of prediction model performance', *BMJ (Online)*, 356. Available at: https://doi.org/10.1136/bmj.i6460

Dineva, K. and Atanasova, T. (2020) 'Systematic look at machine learning algorithms: advantages, disadvantages and practical applications', in *International Multidisciplinary Scientific GeoConference on Surveying Geology Mining Ecology Management (SGEM)*, (2.1), pp. 317–324. Available at: https://doi.org/10.5593/sgem2020/2.1/s07.041

Dreyling, R.M., Tammet, T. and Pappel, I. (2023) 'Digital transformation insights from an AI solution in search of a problem', in T.K. Dang, J. Küng and T.M. Chung (eds) *Future data and security engineering. Big data, security and privacy, smart city and Industry 4.0 applications. FDSE 2023. Communications in computer and information science*, 1925, pp. 341–351. Available at: https://doi.org/10.1007/978-981-99-8296-7_24

Dreyling III, R., Tammet, T. and Pappel, I. (2024) 'Technology push in AI-enabled services: how to master technology integration in case of Bürokratt', *SN Computer Science*, 5(6), article 738. Available at: https://doi.org/10.1007/s42979-024-03064-0

Edler, D. (2019) 'Digitisation and urban development in Estonia: smart city solutions in "e-Estonia"—the example of Tartu', *Geographische Rundschau*, 71(4), pp. 46–51.

Equal Treatment Act (Estonia) (2008) *RT I*, 56, 315. Available at: https://www.riigiteataja.ee/en/eli/530102013066/consolide (Accessed: 3 October 2025).

European Commission (2020) *White paper on artificial intelligence: a European approach to excellence and trust*. Available at: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (Accessed: 3 October 2025).

European Commission (2024) *Estonia 2024 Digital Decade country report*. Available at: https://digital-strategy.ec.europa.eu/en/factpages/estonia-2024-digital-decade-country-report (Accessed: 3 October 2025).

Fountain, J.E. (2022) 'The moon, the ghetto and artificial intelligence: reducing systemic racism in computational algorithms', *Government Information Quarterly*, 39(2), article 101645. Available at: https://doi.org/10.1016/j.giq.2021.101645

Fontanari, T., Fróes, T.C. and Recamonde-Mendoza, M. (2022) 'Cross-validation strategies for balanced and imbalanced datasets', in J.C. Xavier-Junior and R.A. Rios (eds) *Intelligent systems. BRACIS 2022. Lecture notes in computer science*, 13653, pp. 626–640. Available at: https://doi.org/10.1007/978-3-031-21686-2_43

Franzoni, V. (2023) 'From black box to glass box: advancing transparency in artificial intelligence systems for ethical and trustworthy AI', in O. Gervasi *et al.* (eds) *Computational science and its applications—ICCSA 2023 Workshops. ICCSA 2023. Lecture notes in computer science*, 14107, pp. 118–130. Available at: https://doi.org/10.1007/978-3-031-37114-1_9

Gender Equality Act (Estonia) (2004) *RT I*, 27, 181. Available at: https://www.riigiteataja.ee/en/eli/530102013038/consolide (Accessed: 3 October 2025).

Goel, R.K. and Nelson, M.A. (2025) 'Awareness of artificial intelligence: diffusion of AI versus ChatGPT information with implications for entrepreneurship', *Journal of Technology Transfer*, 50(1), pp. 96–113. Available at: https://doi.org/10.1007/s10961-024-10089-3

Göksal, Ş.İ. and Solarte-Vásquez, M.C. (2024) 'The blockchain-based trustworthy artificial intelligence supported by stakeholders-in-the-loop model', *Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration*, 32(2), article 2. Available at: https://doi.org/10.46585/sp32022083

Göksal, Ş. İ., Solarte-Vásquez, M.C. and Chochia, A. (2025) 'The EU AI Act's alignment within the European Union's regulatory framework on artificial intelligence', *International and Comparative Law Review*, 24(2), pp. 25–53. Available at: https://doi.org/10.2478/iclr-2024-0017

Gonçalves, R., Gouveia, F., Lynce, I. and Santos, J.F. (2025) 'Proxy attribute discovery in machine learning datasets via inductive logic programming', in

A. Gurfinkel and M. Heule (eds) *Tools and algorithms for the construction and analysis of systems. TACAS 2025. Lecture notes in computer science*, 15697, pp. 343–363. Available at: https://doi.org/10.1007/978-3-031-90653-4_17

Gonuguntla, K., Shaik, A. and Agrawal, P. (2025) 'Statistical bias', in A.E.M. Eltorai, J.A. Bakal and C.M. Gibson (eds) *Translational cardiology*. Amsterdam: Elsevier, pp. 179–180. Available at: https://doi.org/10.1016/B978-0-323-91790-2.00039-3

González-Sendino, R., Serrano, E. and Bajo, J. (2024) 'Mitigating bias in artificial intelligence: fair data generation via causal models for transparent and explainable decision-making', *Future Generation Computer Systems*, 155, pp. 384–401. Available at: https://doi.org/10.1016/j.future.2024.02.023

Hartung, T., Hoffmann, S. and Whaley, P. (2025) 'Assessing risk of bias in toxicological studies in the era of artificial intelligence', *Archives of Toxicology*, 99(8), pp. 3065–3090. Available at: https://doi.org/10.1007/s00204-025-03978-5

Henzinger, T., Karimi, M., Kueffner, K. and Mallik, K. (2023) 'Runtime monitoring of dynamic fairness properties', in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 604–614. Available at: https://doi.org/10.1145/3593013.3594028

Heo, J. and Kim, J.-Y. (2022) 'PIM for ML training', in J.-Y. Kim, B. Kim and T.T.-H. Kim (eds) *Processing-in-memory for AI: from circuits to systems*. Cham: Springer International Publishing, pp. 121–142. Available at: https://doi.org/10.1007/978-3-030-98781-7_6

Hofmann, B. (2025) 'Biases in AI: acknowledging and addressing the inevitable ethical issues', *Frontiers in Digital Health*, 7. Available at: https://doi.org/10.3389/fdgth.2025.1614105

Issaro, S., Nilsook, P. and Wannapiroon, P. (2023) 'Artificial intelligence engineering for predictive analytics', in *2023 Research, Invention and Innovation Congress: Innovation electrical and electronic: innovation for a better life (RI2C)*, pp. 51–58. Available at: https://doi.org/10.1109/RI2C60382.2023.10355979

Jain, L.R. and Menon, V. (2023) 'AI algorithmic bias: understanding its causes, ethical and social implications', in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 460–467. Available at: https://doi.org/10.1109/ICTAI59109.2023.00073

Kalda, K., Sell, R. and Soe, R.-M. (2021) 'Use case of autonomous vehicle shuttle and passenger acceptance analysis', *Proceedings of the Estonian Academy of Sciences*, 70(4), pp. 429–435. Available at: https://doi.org/10.3176/proc.2021.4.09

Kim, C. (2019) 'A modular framework for collaborative multimodal annotation and visualisation', in *Companion Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19 Companion)*, pp. 165–166. Available at: https://doi.org/10.1145/3308557.3308730

Kolberg, J., Rathgeb, C. and Busch, C. (2023) 'The influence of gender and skin colour on the watchlist imbalance effect in facial identification scenarios', in J.J. Rousseau and B. Kapralos (eds) *Pattern recognition, computer vision, and image processing. ICPR 2022 International Workshops and Challenges. ICPR 2022. Lecture notes in computer science*, 13643, pp. 465–478. Available at: https://doi.org/10.1007/978-3-031-37660-3_33

Kore, A., Bavil, E.A., Subasri, V., Abdalla, M., Fine, B., Dolatabadi, E. and Abdalla, M. (2024) 'Empirical data drift detection experiments on real-world medical imaging data', *Nature Communications*, 15(1), article 1887. Available at: https://doi.org/10.1038/s41467-024-46142-w

Kulkarni, A., Chong, D. and Batarseh, F.A. (2020) 'Foundations of data imbalance and solutions for a data democracy', in F.A. Batarseh and R. Yang (eds) *Data democracy: at the nexus of artificial intelligence, software development, and knowledge engineering*. Amsterdam: Elsevier, pp. 83–106. Available at: https://doi.org/10.1016/B978-0-12-818366-3.00005-8

Lacmanovic, S. and Skare, M. (2025) 'Artificial intelligence bias auditing—current approaches, challenges and lessons from practice', *Review of Accounting and Finance*, 24(3), pp. 375–400. Available at: https://doi.org/10.1108/RAF-01-2025-0006

Lobo, P.R. (2022) 'Bias in hate speech and toxicity detection', in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, p. 910. Available at: https://doi.org/10.1145/3514094.3539519

Marripudugala, M. (2025) 'Leveraging large language models for bias detection and mitigation in data analytics models', in *2025 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 960–965. Available at: https://doi.org/10.1109/ICICCS65191.2025.10984507

Matzat, L. (2019) 'Finnish credit score ruling raises questions about discrimination and how to avoid it', *AlgorithmWatch*. Available at: https://algorithmwatch.org/en/finnish-credit-score-ruling-raises-questions-about-discrimination-and-how-to-avoid-it/ (Accessed: 3 October 2025).

Metsallik, J. and Ross, P. (2021) 'Testing the applicability of digital decision support on a nationwide EHR', in R. Jallouli, M.A., Bach Tobij, H. Mcheick and G. Piho (eds) *Digital economy. Emerging technologies and business innovation. ICDEc 2021 Lecture notes in business information processing*, 431, pp. 134–146. Available at: https://doi.org/10.1007/978-3-030-92909-1_9

Mian, S.M., Khan, M.S., Shawez, M. and Kaur, A. (2024) 'Artificial intelligence (AI), machine learning (ML) and deep learning (DL): a comprehensive overview on techniques, applications and research directions', in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pp. 1404–1409. Available at: https://doi.org/10.1109/ICSCSS60660.2024.10625198

Ministry of Economic Affairs and Communications (Estonia) (2021) *Estonia's Digital Agenda 2030*. Available at: https://www.mkm.ee/en/e-state-and-connectivity/digital-agenda-2030 (Accessed: 3 October 2025).

Nõmmik, S. (2025) 'Challenges with the design and use of AI-based tools in public sector in Estonia: moving beyond the potential of a great idea on paper', in N. Urs, D. Špaček and D. Nõmmik (eds) *Digital transformation in European public services. Governance and public management*, F355. Cham: Palgrave Macmillan, pp. 129–154. Available at: https://doi.org/10.1007/978-3-031-81425-9_7

Pappachan, P., Moslehpour, M., Bansal, R. and Rahaman, M. (2024) 'Transparency and accountability', in B. Gupta (ed) *Challenges in large language model development and AI ethics*. Hershey, PA: IGI Global, pp. 178–211. Available at: https://doi.org/10.4018/979-8-3693-3860-5.ch006

Porter, Z., Ryan, P., Morgan, P., Al-Qaddoumi, J., Twomey, B., Noordhof, P., McDermid, J. and Habli, I. (2025) 'Unravelling responsibility for AI', *Journal of Responsible Technology*, 23, article 100124. Available at: https://doi.org/10.1016/j.jrt.2025.100124

Pulivarthy, P. and Whig, P. (2024) 'Bias and fairness addressing discrimination in AI systems', in P. Bhattacharya, A. Hassan, H. Liu and B. Bhushan (eds) *Ethical dimensions of AI development*. Hershey, PA: IGI Global, pp. 103–126. Available at: https://doi.org/10.4018/979-8-3693-4147-6.ch005

Radanliev, P. (2025) 'AI ethics: integrating transparency, fairness, and privacy in AI development', *Applied Artificial Intelligence*, 39(1), article 2463722. Available at: https://doi.org/10.1080/08839514.2025.2463722

Raja, V.J., Dhanamalar, M., Solaimalai, G., Rani, D.L., Deepa, P. and Vidhya, R.G. (2024) 'Machine learning revolutionising performance evaluation: recent developments and breakthroughs', in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pp. 780–785. Available at: https://doi.org/10.1109/ICSCSS60660.2024.10625103

Reddy-Best, K.L. and Olson, E. D. (2020) 'Packers, dilators and the options for either male or female: navigating movement of transgender and gender non-conforming bodies, appearances and luggage through airport security', *Fashion, Style and Popular Culture*, 7(2–3), pp. 223–246. Available at: https://doi.org/10.1386/fspc_00016_1

'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) (text with EEA relevance)' (2016) *Official Journal* L 119, 4.5.2016, pp. 1–88. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

'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)' (2024) *Official Journal* L 2024/1689, 12.7.2024, pp. 1–170. Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Ricaurte, P. and Zasso, M. (2023) 'AI, ethics, and coloniality: a feminist critique', in M. Cebral-Loureda, E.G. Rincón-Flores and G. Sanchez-Ante (eds) *What AI can do: strengths and limitations of artificial intelligence*. Boca Raton: CRC Press, pp. 39–57. Available at: https://doi.org/10.1201/b23345-4

Ruschemeier, H. and Hondrich, L.J. (2024) 'Automation bias in public administration: an interdisciplinary perspective from law and psychology', *Government Information Quarterly*, 41(3), article 101953. Available at: https://doi.org/10.1016/j.giq.2024.101953

Russell, S. and Norvig, P. (2021) *Artificial intelligence: a modern approach*. 4th edn. Harlow: Pearson.

Said, Y., Saidani, T., Atri, M., Alsheikhy, A.A. and Shawly, T. (2026) 'Computational intelligence for emotion recognition in autism spectrum disorder: a systematic review of signal-based modelling, simulation, and clinical potential', *Biomedical Signal Processing and Control*, 111, article 108367. Available at: https://doi.org/10.1016/j.bspc.2025.108367

Sen, R. and Das, S. (2023) 'Unsupervised learning', in *Indian Statistical Institute Series*. Dordrecht: Springer Science and Business Media B.V., pp. 305–318. Available at: https://doi.org/10.1007/978-981-19-2008-0_21

Senthil, G.A., Geerthik, S., Jerlin Ida, J. and Ashika Jubi, S. (2025) 'Ethical AI auditor for bias detecting in AI models using adversarial debiasing', *in 2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE), pp. 1–6*. Available at: https://doi.org/10.1109/AMATHE65477.2025.11081330

Seraj, A., Mohammadi-Khanaposhtani, M., Daneshfar, R., Naseri, M., Esmaeili, M., Baghban, A., Habibzadeh, S. and Eslamian, S. (2022) 'Cross-validation', in S. Eslamian and F. Eslamian (eds) *Handbook of hydroinformatics: Volume I: Classic soft-computing techniques*. Amsterdam: Elsevier, pp. 89–105. Available at: https://doi.org/10.1016/B978-0-12-821285-1.00021-X

Singhi, S.K. and Liu, H. (2006) 'Feature subset selection bias for classification learning', in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, 148, pp. 849–856. Available at: https://doi.org/10.1145/1143844.1143951

Snow, J. (2018) 'Amazon's face recognition falsely matched 28 members of Congress with mugshots', *ACLU of Northern California, 26 July*. Available at: https://

www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28 (Accessed: 3 October 2025).

Sogancioglu, G., Kaya, H. and Salah, A.A. (2023) 'Using explainability for bias mitigation: a case study for fair recruitment assessment', in *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*, pp. 631–639. Available at: https://doi.org/10.1145/3577190.3614170

Soleimani, M., Intezari, A. and Pauleen, D.J. (2022) 'Mitigating cognitive biases in developing AI-assisted recruitment systems: a knowledge-sharing approach', *International Journal of Knowledge Management*, 18(1). Available at: https://doi.org/10.4018/IJKM.290022

Steidl, M., Felderer, M. and Ramler, R. (2023) 'The pipeline for the continuous development of artificial intelligence models—current state of research and practice', *Journal of Systems and Software*, 199, article 111615. Available at: https://doi.org/10.1016/j.jss.2023.111615

Taylor, J.M.G., Ankerst, D.P. and Andridge, R.R. (2008) 'Validation of biomarker-based risk prediction models', *Clinical Cancer Research*, 14(19), pp. 5977–5983. Available at: https://doi.org/10.1158/1078-0432.CCR-07-4534

Tellez, N., Serra, J., Kumar, Y., Li, J.J. and Morreale, P. (2023) 'Gauging biases in various deep learning AI models', in K. Arai (ed) *Intelligent systems and applications. IntelliSys 2022. Lecture notes in networks and systems*, 544, pp. 171–186. Available at: https://doi.org/10.1007/978-3-031-16075-2_11

Thelle, M.H., Lundberg, A.T., Hovmand, B.E., Woltmann, H.H., Virtanen, L., Tranholm-Mikkelsen, N., Pedersen, S.T. and Oure, A.J. (2024) *The economic opportunity of AI in Estonia: capturing the next wave of benefits from generative AI*. Copenhagen: Implement Consulting Group. Available at: https://www.justdigi.ee/sites/default/files/documents/2024-12/2024_The-economic-opportunity-of-AI-in-Estonia.pdf (Accessed: 3 October 2025).

Vatsa, V.R. and Chhaparwal, P. (2021) 'Estonia's e-governance and digital public service delivery solutions', in *2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pp. 135–138. Available at: https://doi.org/10.1109/CCICT53244.2021.00036

Verma, S., Paliwal, N., Yadav, K. and Vashist, P. C. (2024) 'Ethical considerations of bias and fairness in AI models', in *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pp. 818–823. Available at: https://doi.org/10.1109/ICDT61202.2024.10489577

Vetrò, A. (2021) 'Imbalanced data as risk factor of discriminating automated decisions: a measurement-based approach', *Journal of Intellectual Property, Information Technology and E-Commerce Law*, 12(4), pp. 272–288.

Vetrò, A., Torchiano, M. and Mecati, M. (2021) 'A data quality approach to the identification of discrimination risk in automated decision-making systems',

Government Information Quarterly, 38(4), article 101619. Available at: https://doi.org/10.1016/j.giq.2021.101619

Vitek, M., Das, A., Lucio, D.R., Zanlorensi, L.A., Menotti, D. and Khiarak, J.N. (2023) 'Exploring bias in sclera segmentation models: a group evaluation approach', *IEEE Transactions on Information Forensics and Security*, 18, pp. 190–205. Available at: https://doi.org/10.1109/TIFS.2022.3216468

Vorras, A. and Mitrou, L. (2021) 'Unboxing the black box of artificial intelligence: algorithmic transparency and/or a right to functional explainability', in T.-E. Synodinou, P. Jougleux, C. Markou and T. Prastitou-Merdi (eds) *EU internet law in the digital single market*. Cham: Springer International Publishing, pp. 247–264. Available at: https://doi.org/10.1007/978-3-030-69583-5_10

Woolf, B., Betke, M., Yu, H., Bargal, S.A., Arroyo, I., Magee, J., Allessio, D. and Rebelsky, W. (2023) 'Face readers: the frontier of computer vision and math learning', *CEUR Workshop Proceedings*, 3491, pp. 24–36.

Xivuri, K. and Twinomurinzi, H. (2023) 'How AI developers can assure algorithmic fairness', *Discover Artificial Intelligence*, 3(1), article 27. Available at: https://doi.org/10.1007/s44163-023-00074-4

Yang, M., Wang, J. and Ton, J.-F. (2023) 'Rectifying unfairness in recommendation feedback loop', in *SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 28–37. Available at: https://doi.org/10.1145/3539618.3591754

Zemmari, A. and Benois-Pineau, J. (2020) 'Supervised learning problem formulation', in *Deep learning of mining of visual content: SpringerBriefs in computer science*. Cham: Springer, pp. 5–11. Available at: https://doi.org/10.1007/978-3-030-34376-7_2

Zhaltyrbayeva, R., Jangabulova, A., Suleimenova, S., Saimova, S. and Tlembayeva, Z. (2025) 'Legal challenges of regulating artificial intelligence in law enforcement: taking into account the interdisciplinary approach to socio-legal transformations', *Social & Legal Studios*, 8(2), pp. 118–130. Available at: https://doi.org/10.32518/sals2.2025.118