

Fair AI in public administration: Estonia and EU law

Archil Chochia

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

Abstract: This article examines fair and trustworthy artificial intelligence in public administration, presenting Estonia as the main case and EU law as the wider legal frame. It builds on the EquiTech project, which has been successfully co-implemented by the TalTech Department of Law and its project partners. The project materials, including training materials for public servants and data scientists, the project proposal, and related academic work, form an important empirical and normative basis for the article. The central claim is that fairness in public-sector artificial intelligence (AI) cannot be reduced to technical accuracy, general transparency, or compliance with a single legal instrument. Wachter, Mittelstadt and Russell's work on why fairness cannot be automated is the paper's main theoretical anchor. Their concept of contextual equality helps explain why EU non-discrimination law does not translate easily into fixed fairness metrics, even though statistical evidence remains necessary. In the Estonian setting, this problem is visible in the gap between an advanced digital state and a still developing framework for assessing bias and discrimination risks across the full lifecycle of automated decision-making. The article argues that trustworthy public-sector AI requires legal authority, data governance, bias assessment, meaningful human oversight, administrative contestability, and institutional capacity. The EU *Artificial Intelligence Act* is a major step in this direction, but it cannot by itself solve the legal and practical problems raised by algorithmic discrimination in public administration.

Keywords: *algorithmic fairness, artificial intelligence, automated decision-making, EquiTech, Estonia, EU AI Act, non-discrimination, public administration, trustworthy AI*

1. Introduction

Estonia is often treated as a reference point for digital government. This is not merely because all public services are already available online but also because the Estonian state has built much of its administrative identity around digital identity, interoperable registers, secure data exchange, and the idea that public administration should be available without unnecessary physical contact or repeated submission of the same data. The movement from e-government to data-driven government makes artificial intelligence (AI) and automated decision-making a logical next stage. It also changes the legal problem (Smuha, 2021). Digital public administration is no longer only about access to online services or secure databases but also about how administrative judgement is supported, shaped, accelerated, and sometimes partially replaced by algorithmic systems.

This article derives from the work carried out under the EquiTech project (no date) of the European Commission. The project was implemented by the Gender Equality and Equal Treatment Commissioner's Office of Estonia, TalTech Department of Law, the Office of the Equal Opportunities Ombudsperson of Lithuania, the Estonian Ministry of Economic Affairs and Communications, the Estonian Ministry of Justice, and Tallinn University of Technology. The Department of Law at TalTech has successfully implemented those project activities that are directly relevant to the legal and institutional analysis developed here, including research, training materials, academic dissemination, and practical tools for public-sector actors. The project proposal describes its objective as enhancing the capacity of public authorities and cooperation partners in Estonia and Lithuania to respond to discrimination and bias risks in automated decision-making systems. It also foresees the development of guidelines, an impact assessment checklist, training materials, and an online course for public-sector employees and system developers.

The article therefore does not treat fairness in AI as an abstract ethics topic. It examines fairness as a legal and institutional problem in a public administration that already uses digital tools at scale and is moving further towards AI-supported services. In Estonian practice, AI and automated decision-making are increasingly present in the public sector, although generally not in the form of fully autonomous final decisions. The study materials, prepared in the framework of the project for public servants (Lips and Joamets, 2025b), state that no public service in Estonia is currently

entirely based on AI-based decision-making, since a human remains involved in final decisions at least to some extent. That, however, does not remove the risk. Bias can emerge in earlier stages, including data collection, feature selection, model training, validation, prioritisation, scoring, recommendation, or routing. By the time a human official sees a specific case, the relevant administrative options may already have been narrowed.

At the same time, EU non-discrimination law is not a static technical standard. It relies on contextual equality, which means that the legal assessment of discrimination depends on the protected ground, the comparator group, the type of harm, the sector, the available evidence, and the judicial understanding of justification (Bell, 2002; Ellis and Watson, 2012; Wachter, Mittelstadt and Russell, 2021). A technical fairness metric may identify a statistical disparity, but it cannot determine by itself whether a legally relevant discrimination claim exists. The law asks questions that are not purely mathematical: Which comparison matters? Which disadvantage counts? Is the apparent disparity justified by a legitimate aim and proportionate means? Should the case be understood as direct discrimination, indirect discrimination, harassment, or something else? These questions are familiar in equality law, but automated systems make it harder to ask them because the relevant rule may not be visible.

The Estonian case brings that argument into a concrete setting. Estonia's digital government creates opportunities for efficient and consistent public services. It can also hide discrimination risks in systems that are not experienced by citizens as AI systems at all. A person may only see a public-service recommendation, a risk category, a chatbot response, a prioritisation decision, or an administrative outcome. The decision may appear ordinary, even when it is influenced by complex data processing. For a claimant, the usual starting point of a discrimination case—the suspicion that a human decision-maker has acted unfairly—may be absent. This matters because non-discrimination law has often relied on people noticing unequal treatment. Algorithmic systems can produce unequal outcomes without any individual official intending discrimination and without the affected person knowing which comparison group should be used.

The argument developed above is modest in one sense. It does not claim that AI should be excluded from public administration. Nor does it claim that human decision-making is inherently fairer. Administrative discretion has always carried risks of arbitrariness, bias, inconsistency, and opacity. Rather, the claim is that standards for fair and trustworthy public-sector AI must

be institutional, not only technical. Estonia needs methods for detecting, documenting, auditing, contesting, and correcting discriminatory effects. The EU *Artificial Intelligence Act* (hereafter AI Act; ‘Council regulation (EU) 2024/1689’, 2024) supplies part of that architecture, especially for high-risk AI systems, but it does not replace equality law, data protection law, administrative procedure, constitutional principles, or sector-specific rules. The practical challenge is to bring those regimes into contact.

2. EU legal approaches to trustworthy public-sector AI

The EU’s approach to AI has developed through ethics guidelines, policy documents, data protection law, equality law, product regulation, and now the AI Act. The High-Level Expert Group on Artificial Intelligence defined trustworthy AI through three requirements: lawful AI, ethical AI, and robust AI (AI HLEG, 2019). That formulation remains influential because it refuses to treat trust as a matter of technical performance alone. Lawfulness refers to compliance with binding legal rules. Ethics concerns values and social acceptability beyond minimum legality. Robustness covers technical and social reliability, including safety, resilience, and the capacity to avoid unacceptable harm.

The AI Act gives this policy language a binding legal structure. ‘Council regulation (EU) 2024/1689’ (2024) establishes harmonised rules for AI systems and follows a risk-based approach. Certain AI practices are prohibited. High-risk AI systems are allowed, but they are subject to requirements on risk management, data governance, technical documentation, record-keeping, transparency, human oversight, accuracy, robustness, and cybersecurity. Limited-risk systems are mainly subject to transparency duties, while minimal-risk systems are largely left outside the more demanding regulation. The AI Act is the first horizontal EU regulation on AI, and it should be read within the wider EU institutional framework on artificial intelligence rather than as a self-contained code (Göksal, Solarte-Vasquez and Chochia, 2024).

Public administration occupies a difficult position under this model. Many administrative uses of AI will fall within areas that can affect rights, benefits, obligations, access to services, or exposure to state scrutiny. Some of these applications will be high-risk under the AI Act, particularly where systems are used for access to essential public and private services, law enforcement, migration, asylum and border control, education, employment,

or other sensitive domains. Other systems may not meet the definition of high-risk AI, may not be AI systems within the meaning of the AI Act, or may be treated as lower-risk because they support rather than determine final decisions. Yet those systems may still shape administrative outcomes. An algorithm used to prioritise inspections, flag applications, classify risk, route citizens to different service channels, or recommend case outcomes may create unequal effects even if the final legal decision remains formally human.

This produces a regulatory gap that the Estonian materials clearly identify. There are currently no comprehensive legal requirements that would ensure that public-sector algorithms are systematically assessed for discriminatory impact before deployment. The AI Act partially responds through life-cycle risk management and bias mitigation duties, but its scope remains limited because many public-sector algorithmic systems may fall outside its requirements (Kerikmäe, Joamets and Metin, 2025). The Act requires risk management, data governance, technical documentation, transparency, accuracy, robustness, and human oversight for high-risk systems. For systems outside that category, discrimination-related obligations remain less clear (Joamets, 2025).

EU non-discrimination law is another part of the framework. Article 21 of the *Charter of Fundamental Rights of the European Union* (2012) prohibits discrimination on grounds such as gender, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or other opinion, membership of a national minority, property, birth, disability, age, or sexual orientation. Secondary equality law includes the Racial Equality Directive, the Employment Equality Directive, and the Gender Goods and Services Directive. These instruments differ in scope. Some protect only specific grounds in particular sectors. EU equality law provides protection through a complex and uneven structure of primary and secondary law, with the level legal protection depending heavily on the relevant sector and ground (Wachter, Mittelstadt and Russell, 2021). This is relevant for AI because algorithmic systems may generate disadvantages that do not map neatly onto protected grounds, or that operate through proxies which are difficult to identify.

The General Data Protection Regulation (GDPR) also remains central. Article 22 of the GDPR restricts decisions based solely on automated processing, including profiling, where such decisions produce legal effects or similarly significant effects ('Council regulation (EU) 2016/679', 2016). Articles 13 to

15 also require information about automated decision-making in certain circumstances. The scope and strength of these rights have been debated. Scholars argue that the GDPR does not contain a broad, general right to explanation of all automated decisions (Wachter, Mittelstadt and Floridi, 2017). It has been suggested that individuals need a more practical means of obtaining useful information without requiring full access to complex models (Wachter, Mittelstadt and Russell, 2018). For public administration, the GDPR is therefore relevant, even if incomplete. It helps establish transparency and rights of intervention in specific cases, but it does not provide a full equality audit framework for public-sector algorithms.

A further issue is the meaning of trust. The AI Act risks conflating trustworthiness with acceptable risk (Laux, Wachter and Mittelstadt, 2024). This criticism is useful in the public-sector context. A system may pass conformity requirements and still fail to deserve the trust of citizens if it is not contestable, remains poorly explained, or used in a domain where the public authority cannot justify the distributional effects. Trustworthiness is not only about reducing risk below a regulatory threshold. It also concerns the relation between the citizen and the state. Public authorities do not merely offer products; they exercise public power.

3. Fairness, contextual equality, and the limits of technical metrics

Much of the technical literature on algorithmic fairness asks whether a model satisfies the fairness criterion. Different criteria exist, including demographic parity, equalised odds, equal opportunity, calibration, predictive parity, individual fairness, counterfactual fairness, and proxy non-discrimination. These measures capture different ideas of fairness and are often mathematically incompatible. A model cannot always satisfy several of them at once, especially where base rates differ across groups (Chouldechova, 2017, Kleinberg, Mullainathan and Raghavan, 2017). It is not only a technical inconvenience, but also shows that fairness metrics embed normative choices. Choosing a metric means choosing which error matters most, which comparison matters, and which distributional pattern is acceptable.

This problem should be directly connected to EU non-discrimination law. This is not to say that the technical fairness metrics are useless; they are

not. However, they cannot automate legal fairness. EU non-discrimination law does not provide a fixed mathematical test. Courts assess claims through concepts such as direct discrimination, indirect discrimination, comparators, protected grounds, disadvantage, legitimate aims, proportionality, and burden of proof. These concepts require interpretation. A statistical disparity may help establish *prima facie* discrimination, but it does not decide the case (Wachter, Mittelstadt and Russell, 2021).

At the same time, automation changes the evidential problem. Traditional discrimination claims often begin with some visible signal or action—a “red flag”, so to speak: a refusal, a hostile comment, a policy that appears exclusionary, a pattern of treatment noticed by affected persons, or a comparator who received different treatment. Algorithmic systems can remove these signals. They classify at scale, rely on unintuitive correlations, and use data features that may be far removed from protected characteristics. Machine-learning opacity may arise from secrecy, technical complexity, and a mismatch between machine-learning methods and human understanding (Burrell, 2016). If neither the citizen nor the front-line official understands the decision pathway, the practical ability to identify discrimination is greatly weakened.

Wachter, Mittelstadt and Russell (2021) propose conditional demographic disparity as a baseline statistical measure for identifying possible automated discrimination. Conditional demographic disparity is not presented as a final legal test but as a means to establish a more consistent evidential starting point while preserving judicial interpretation. This distinction is important for public administration. A public authority should not ask a model to decide whether it is legally fair. It should ask whether the system produces disparities that require legal attention, further investigation, justification, redesign, or withdrawal.

Other academic work supports this more cautious view of algorithmic fairness. Barocas and Selbst (2016) show how big data can create disparate impact through biased training data, target variables that encode historical inequality, feature selection, and feedback loops. Selbst *et al.* (2019) describe the abstraction traps that arise when technical systems ignore the social context in which they operate. Binns (2018) argues that algorithmic accountability cannot be understood only through individualised explanations, since many harms occur at group level or through systemic effects. Mittelstadt’s (2016) work on auditing content personalisation also shows that systems can produce opaque and difficult-to-contest effects even when individual decisions appear minor.

The legal literature on algorithmic discrimination under EU law arrives at similar conclusions. Hacker (2018) argues that discrimination law can respond to algorithmic harms, but only if the evidential and procedural barriers created by opaque systems are addressed. Gerards and Xenidis (2021) suggest that algorithmic discrimination challenges the categories of European equality law, including the distinction between direct and indirect discrimination and the limits of protected grounds. Adams-Prassl, Binns and Kelly-Lyth (2023) argue that algorithms can be directly discriminatory even when their discriminatory logic is embedded in statistical systems rather than explicit human prejudice. Weerts *et al.* (2023) emphasise the point strongly: EU non-discrimination law is not a decision tree. That phrase captures the problem faced by developers and public authorities who demand clear rules. Equality law requires structured judgement, not only classification.

This has consequences for Estonia. A public-sector AI system may be accurate in aggregate and still unfair in its distribution of errors. It may treat individuals consistently according to its internal logic and still reproduce structural inequality. It may avoid the explicit use of protected characteristics and still discriminate through proxies. It may be explainable to developers but not to affected citizens. It may keep a human in the loop while placing that human under organisational pressure to follow the output. These are not speculative risks but well-documented patterns in the broader literature on automated decision-making, including work on welfare administration, predictive policing, credit scoring, online advertising, employment systems, and public-sector profiling (Citron and Pasquale, 2014; Eubanks, 2018; O’Neil, 2016; Pasquale, 2015).

4. Estonia’s digital public administration and the emergence of automated decision-making systems (ADMs)

Estonian public administration provides a distinctive environment for these questions. Its digital infrastructure is older and more integrated than that of most other EU Member States. The *Digital Agenda 2030* sets an ambition for an AI-powered government and a seamless digital state. Kerikmäe and Pärn-Lee (2021) have argued that Estonia’s AI strategy must be understood against the country’s broader e-society model and its attempt to balance

national innovation with legal responsibility. This context matters because AI is not being incorporated into paper-based administration but into an already data-rich administrative environment.

There are several examples and areas of automated decision-making, including Bürokratt, OTT, ABC Gates, and tools connected to employment and public-service delivery (Kerikmäe, Joamets and Metin, 2025). Bürokratt is often presented as a virtual assistant or conversational AI layer for public services. Its legal risk differs from that posed by a fully automated benefits decision, but it still matters. If a chatbot misdirects users, provides uneven information, fails to recognise language patterns, or gives lower-quality assistance to certain groups, the consequences may affect access to rights. A user may not experience this as a formal decision, but the effect can be practical exclusion.

ABC Gates raise a different set of issues. Biometric and border-control systems are sensitive because they combine security, mobility, identity, and state coercion. Pilving and Mikiver (2020) discuss algorithmic decisions through the Estonian idea of *Kratt*, a practical application designed using AI systems, asking how administrative law principles apply when decisions are made or supported by autonomous systems. Biometric systems can also produce unequal error rates across demographic groups, especially where training data are imbalanced or capture conditions differ. In border-control settings, even small differences in false rejection or false alarm rates can create significant burden for particular groups.

The Estonian Unemployment Insurance Fund is another important example. Systems used in labour-market services may classify individuals by risk, employability, or likelihood of returning to work. These tools can help target resources. They can also reproduce disadvantage if historical labour-market inequality becomes part of the model. Age, disability, language, location, care responsibilities, education, and employment history may interact in ways that create proxy discrimination. Even if the system is intended to support rather than replace a human official, its classifications may frame the official's view of the person (Zolkin, Chochia and Hoffmann, 2023).

The EquiTech training materials (Lips and Joamets, 2025a; 2025b) describe the practical concern in plain terms. Algorithms are expected to make decisions more uniform, objective, and accurate. They may reduce some forms of human bias. Yet bias can enter through invisible stages, including data, model design, proxy features, and system use. The materials for data

scientists and developers emphasise that deep-learning systems can identify rules or correlations that developers do not fully anticipate, and that indirect discrimination may arise through proxy characteristics. The public servant materials produced in the course of the project issue the same warning to non-technical staff: a human may remain involved in the final decision, but earlier automated stages can influence the outcome (Lips and Joamets, 2025a; 2025b).

5. Bias, data, and the public-sector lifecycle

Bias in public-sector AI is sometimes discussed as if it were a defect in the model. However, that approach is too narrow. Bias can enter at each stage of the system's lifecycle. It can emerge during defining a legal or administrative problem, the selection of the target variable, the data used for training, the labelling process, the exclusion of missing or hard-to-measure populations, the choice of features, the optimisation objective, the testing environment, procurement requirements, user-interface design, organisational incentives, and the way officials interpret system outputs. Barocas and Selbst's (2016) analysis of disparate impact in big data remains a useful point of departure because it shows how apparently neutral technical choices can create discriminatory effects. If a model is trained on historical decisions, it may learn the biases of historical administration. If the target variable measures administrative success rather than citizens' welfare, it may optimise the wrong thing. If the data contain fewer observations for marginalised groups, the model may perform less reliably for them. If protected characteristics are removed, proxies may remain. But the removal of explicit protected attributes is insufficient when other variables can stand in for them (Datta *et al.*, 2017).

Public administration adds specific difficulties. Public bodies often work with data collected for legal or administrative purposes, not for machine learning. Registries may be accurate for one task but unsuitable for another. Data quality can differ across groups because some people interact with the state more often, disclose information differently, or are recorded through different institutional pathways. Historical under-reporting, over-policing, regional inequality, language barriers, and issues with access for users with disability can all become data problems. A model trained on such data may appear technically objective while reproducing patterns created by earlier administrative or social inequality.

Feedback loops are also important. If an AI system flags certain groups as higher risk, they may receive more scrutiny. More scrutiny can produce more recorded problems, which subsequently confirm the model's assumptions. This pattern has been widely discussed in policing and welfare systems, but it can arise wherever administrative attention is allocated by risk scores. The danger in public sector is not only that benefits are denied or sanctions imposed but that some people become administratively targeted mainly as risks, while others remain ordinary service users.

A life-cycle approach is now common in AI governance, but it must be made more concrete for public administration. Before a system is developed, the authority should ask whether automation is appropriate for the task. Some tasks may be too sensitive to rights, too dependent on contextual judgement, or too poorly supported by data. Where automation is appropriate, the authority should define the public objective precisely. Vague goals such as efficiency or fraud detection are not sufficient.

The authority should identify the legal basis, affected groups, possible harms, data sources, protected characteristics, proxy risks, human oversight model, explanation duties, and appeal routes. During development, bias assessment should include both statistical and legal analysis. Statistical testing should examine error rates, false positives, false negatives, allocation outcomes, calibration, and performance across relevant groups. Legal analysis should ask whether the relevant groups map onto protected grounds, whether indirect discrimination may arise, whether the disadvantage is serious, and whether any justification is legitimate and proportionate. The conditional demographic disparity proposal can help create a baseline for statistical evidence, but it should not be mistaken for the whole legal inquiry (Wachter, Mittelstadt and Russell, 2021).

6. Human oversight, administrative law, and the risk of rubber-stamping

Human oversight is often presented as the solution to automated unfairness. EU law also gives it a central role. Article 14 of the AI Act requires human oversight for high-risk AI systems ('Council regulation (EU) 2024/1689', 2024). The GDPR also treats human intervention as one of the safeguards in certain automated decision-making contexts. The idea is simple: if a human remains responsible, the system will not exercise unchecked power.

The problem is that human oversight can be formal rather than real. A human may lack the time, knowledge, authority, or organisational support to challenge an algorithmic output. If the system is presented as accurate, objective, or scientifically validated, an official may feel coerced to follow it. If deviation requires justification but agreement does not, the rational bureaucratic choice is to accept the recommendation. This phenomenon is sometimes described as automation bias. In public administration, it can become a legal problem because the formal decision-maker may no longer exercise genuine discretion.

Administrative law requires more than the presence of a person in the process. It requires lawful authority, relevant reasons, proportionality, equal treatment, the right to be heard where applicable, and the ability to challenge decisions. Pilving and Mikiver's (2020) discussion of *Kratts* as administrative bodies points to the importance of connecting algorithmic decision-making with established administrative law principles. Similarly, automated decisions in Estonian public administration must be assessed through the requirements of Estonian administrative procedure law, rather than through technology policy (Zolkin, Chochia and Hoffmann, 2023).

A meaningful human oversight model should therefore include several elements. The official must understand the function and limits of the system. The system's output must be presented in a way that supports judgement rather than replaces it. The official must be able to see relevant reasons, uncertainty, data limitations, and known bias risks. There must be an institutional right to override the system. Overrides should be monitored, not punished. The authority should record the instances and reasons as to why officials follow or depart from recommendations. Training should include instruction in equality law and administrative-law duties, not merely in system operation.

Human oversight also raises responsibility questions. If a system produces a discriminatory outcome, then who is responsible? The developer may argue that the authority made the decision to use the case and data. The authority may argue that the vendor designed the system. The official may argue that the system recommendation was authoritative. The data protection officer may argue that equality impact was outside their scope of responsibility. The equality body may lack technical access. This diffusion of responsibility is precisely one reason why the focus of the EquiTech project on training, guidelines, and impact assessment comes useful—it treats fairness as a shared institutional capacity rather than a principle that can be assigned to one office after deployment.

7. Transparency, explanation, and contestability

Transparency is another term that can easily become overly broad. In public-sector AI, transparency should be connected to a function. Citizens need to know when an AI or automated decision-support system is involved in a process that affects them. They need understandable information about the role of the system, the main factors considered, the authority responsible, and the route for challenging an outcome. They do not always need access to source code; in many cases, it would not help. Citizens need information that allows them to identify an error, raise a discrimination concern, or ask for a human review.

Developers, auditors, courts, and oversight bodies need more detailed information. They may require access to training data descriptions, data quality reports, model cards, technical documentation, logging records, validation results, bias testing, procurement documents, and post-deployment monitoring. The requirements of the AI Act on technical documentation, logging, transparency, and human oversight become relevant here, as they create a compliance structure for high-risk systems. For systems outside that category, Estonia may need national guidance or public-sector standards that require similar documentation where rights and equality risks are taken into account.

The literature on explanation under the GDPR shows why the problem with transparency cannot be solved by a simple right to explanation. The GDPR does not create a broad right to an explanation of every automated decision (Wachter, Mittelstadt and Floridi, 2017). Counterfactual explanations may sometimes be more useful because they tell a person what would need to change for them to obtain a different result, without requiring disclosure of the full model (Wachter, Mittelstadt and Russell, 2018). Yet, in discrimination cases, counterfactual explanations may still be insufficient. A person may learn that a certain factor affected their score but remain unaware whether the factor operates as a proxy for ethnicity, disability, age, language, or social origin. Individual explanation and group fairness assessment must therefore be combined.

Contestability is the practical test. If a citizen cannot challenge an AI-influenced administrative outcome, transparency has little value. Contestability requires notice, reasons, access to relevant information, an independent or at least meaningful review process, and remedies. It also

requires that the burden of proof not fall entirely on the affected individual. Wachter, Mittelstadt and Russell's (2021) concern is precisely that automated discrimination may be hard to detect because individuals do not know the rule, the comparison group, or the wider distribution of outcomes. If the public authority holds the data and controls the model, procedural law must compensate for that asymmetry.

8. EquiTech and the practical institutionalisation of fairness

The importance of the EquiTech project lies in addressing the implementation problem. EU law and academic literature identify risks, whereas public servants and developers need practical tools. The project focuses on legal and policy gaps, mapping of discrimination and bias risks, support materials, training, an impact assessment checklist, a media campaign, and academic dissemination. The training materials prepared in the course of the project are relevant. The material for public servants explains AI, algorithmic decision-making, bias, discrimination, responsibility, and provides practical examples (such as Bürokratt) in language designed for administrative users. The material for data scientists, data protection specialists, and developers focuses more on legal principles, protected grounds, proxy characteristics, and the responsibilities of those involved in the design and deployment of systems. Importantly, this impactful distinction recognises that different actors need different forms of knowledge. A public servant does not need to become a machine-learning engineer. A developer does, however, need to understand that removing a protected attribute does not remove discrimination risk.

The EquiTech project represents an effort to develop and disseminate strategies and tools for identifying and preventing bias in Estonia's public-sector automated decision-making processes (Joamets, 2025). There is, however, an important limitation: the work discusses tools and strategies, but further empirical evaluation is needed to assess their effectiveness in practice. That limitation is not necessarily a weakness of the project; rather it points to the next research agenda. Tools of fairness should themselves be evaluated. A checklist can become superficial if it is used only after a system has already been selected. Training can fade if it is not connected to procurement rules and managerial responsibility. Also, a bias assessment can be ignored if no one has authority to halt deployment.

9. Towards a governance model for fair AI in Estonian public administration

A fair and trustworthy AI governance model for Estonia should proceed from a simple rule: no public authority should deploy an AI or algorithmic decision-support system affecting rights, benefits, obligations, access to services, or administrative scrutiny without a documented assessment of legality, necessity, equality risk, data quality, human oversight, and contestability. This should apply even where the system is not classified as high-risk under the AI Act. The reason is practical: discrimination risk depends not only on the legal classification of the technology but also on how the system is used and who is affected.

The model should have several practical parts. First, the authority should identify the legal basis for the system, the administrative task, the public interest pursued, and the alternatives to automation. Second, it should document data sources, data quality, missing data, representativeness, retention rules, and likely proxy variables. Third, it should carry out bias and discrimination assessment that combines statistical testing with legal analysis. The assessment should consider protected grounds, intersectional disadvantage, error distribution, and the possibility that relevant groups are missing from the data. Fourth, human oversight must be designed rather than assumed. Officials need training, authority to override outputs, information about uncertainty, and a working environment in which deviation from the system is possible. Fifth, citizens should receive understandable notice when automated tools in practice affect a decision or service pathway, together with information about review and complaint routes. Sixth, procurement contracts should require documentation, audit access, bias testing, performance reporting, and cooperation with oversight bodies. Claims of trade secrecy cannot be allowed to prevent a public authority from meeting its equality and administrative-law duties. Finally, Estonia should avoid fragmented oversight. Data protection, equality, procurement, cybersecurity, and administrative-law issues often emerge in the same system and should be reviewed together where the system affects rights or access to public services.

10. The EU AI Act and Estonia's implementation challenge

The AI Act will require substantial implementation efforts across Member States. For Estonia, the issue is not whether the state has the capability to adopt another digital governance framework. It likely does. The more difficult question is whether AI governance will be integrated into everyday administrative practice rather than treated as a specialist compliance exercise. If AI governance sits only with technical teams, legal equality concerns may be missed. If it sits only with lawyers, technical risks may be misunderstood. If it sits only with data protection officers, group discrimination and administrative-law issues may receive insufficient attention.

The AI Act's high-risk obligations provide a useful template. Risk management, data governance, documentation, transparency, human oversight, and accuracy are irrelevant only to high-risk systems and reflect a broader public-sector discipline. Estonia could use these obligations as a baseline for all algorithmic systems that affect administrative outcomes in practice, adjusting the depth of assessment to the level of risk. This would prevent the identified problem, whereby systems outside the scope of the AI Act may nevertheless pose discrimination risks (Kerikmäe, Joamets and Metin, 2025). At the same time, the product-safety logic of the AI Act should not dominate the whole discussion. Public administration is not a market in which citizens choose among providers. Administrative decisions carry coercive authority and often affect people who cannot opt out. A welfare applicant, an unemployed person, a cross-border traveller, a student, a patient, or a taxpayer may have no realistic alternative to engaging with the public system. This means that fairness must be judged by public-law standards as well as by AI compliance standards.

The Estonian constitutional and administrative-law framework should therefore remain central. Equal treatment, legality, proportionality, reasoning, the right to be heard, and access to remedies are not made obsolete by AI. Rather, they become harder to operationalise. An AI system may supply a recommendation, but the administrative act must still be legally attributable to a public authority. The reasons must still be understandable enough to allow review. The affected person must be able to challenge the decision. If a public authority cannot explain how a system contributed to the outcome, it should not rely on that system in a case sensitive to rights.

This approach is consistent with Wachter, Mittelstadt and Russell's (2021) warning that fairness cannot be automated because equality law requires contextual judgement. However, fairness can be made administratively assessable. Public authorities can create records, tests, procedures, and institutional responsibilities that allow equality questions to be asked in time. That is the realistic aim for Estonia and other Member States.

11. Conclusion

Fair and trustworthy AI in public administration is not achieved by declaring that a system is accurate, transparent, or human-centric. It requires a legal and institutional setting in which automated systems can be questioned before and after deployment. Estonia is a strong case for this analysis because its digital state is mature, its public administration is open to AI-supported services, and its legal institutions now face the task of translating EU AI governance into administrative routines.

The EquiTech project securely bridges legal principle and practice. Its research, training materials, impact assessment checklist, and dissemination activities address the knowledge gap among public authorities, developers, data specialists, and legal actors. The project also shows that fairness cannot be left to the responsibility of technical teams or courts after harm has occurred but it must be built into procurement, design, training, oversight, monitoring, and review. (EquiTech project, no date)

The concept of contextual equality explains why EU non-discrimination law cannot be converted into a single fairness metric. Automated systems make discrimination harder to detect because they operate through scale, opacity, proxies, and unintuitive classifications. Yet the answer is not to abandon statistical analysis. Conditional demographic disparity and other fairness measures can help produce evidence and should support legal judgement, not replace it. The EU AI Act is an important step in this direction, especially for high-risk systems. The Act gives public authorities and providers a more concrete set of obligations, whereas it does not remove the need for national administrative-law safeguards, equality monitoring, data protection rights, and public-sector accountability. For Estonia, the main task is to ensure that systems outside the strict high-risk category do not escape scrutiny where they affect people in practice.

A workable Estonian model would treat fairness as a lifecycle obligation. It would require a legal basis, data governance, bias assessment, meaningful human oversight, transparency, contestability, procurement controls, and institutional coordination. Such a model would not make AI perfectly fair. That is not a realistic legal or technical goal. However, it would make public-sector AI more visible, more reviewable, and more accountable. In a digital state, that may be the difference between administrative innovation that earns trust and automation that quietly reproduces inequality.

Dr **Archil Chochia** is the director of and a senior researcher at TalTech Law School of Tallinn University of Technology. Dr Chochia has more than 100 academic publications and is a co-editor of the books *Russian Federation in the global knowledge warfare: influence operations in Europe and its neighborhood* (Springer, 2021), *Brexit: history, reasoning and perspectives* (Springer, 2018), and *Political and legal perspectives of the EU Eastern Partnership Policy* (Springer, 2016). He has extensive experience in managing projects funded by different European programmes, such as Horizon2020, Erasmus+, Europe for Citizens, etc.

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*. European Commission High-Level Expert Group on AI.
- Barocas, S. and Selbst, A.D. (2016) 'Big data's disparate impact', *California Law Review*, 104(3), pp. 671–732. Available at: <https://doi.org/10.2139/ssrn.2477899>
- Bell, M. (2002) *Anti-discrimination law and the European Union*. Oxford University Press. Available at: <https://doi.org/10.1093/acprof:oso/9780199244508.001.0001>
- Binns, R. (2018) 'Algorithmic accountability and public reason', *Philosophy & Technology*, 31(4), pp. 543–556. Available at: <https://doi.org/10.1007/s13347-017-0263-5>
- Burrell, J. (2016) 'How the machine "thinks": understanding opacity in machine learning algorithms', *Big Data & Society*, 3(1), pp. 1–12. Available at: <https://doi.org/10.1177/2053951715622512>
- Charter of Fundamental Rights of the European Union* (2012) *Official Journal C* 326, pp. 391–407.

- Chouldechova, A. (2017) 'Fair prediction with disparate impact: a study of bias in recidivism prediction instruments', *Big Data*, 5(2), pp. 153–163. Available at: <https://doi.org/10.1089/big.2016.0047>
- Citron, D.K. and Pasquale, F. (2014) 'The scored society: due process for automated predictions', *Washington Law Review*, 89(1), pp. 1–33.
- Datta, A., Fredrikson, M., Ko, G., Mardziel, P. and Sen, S. (2017) 'Proxy non-discrimination in data-driven systems: theory and experiments with machine learnt programs', *ArXiv*. Available at: <https://doi.org/10.1145/3133956.3134097>
- Ellis, E. and Watson, P. (2012) *EU anti-discrimination law*. 2nd edn. Oxford University Press. Available at: <https://doi.org/10.1093/acprof:oso/9780199698462.001.0001>
- EquiTech project (no date) *Improving response to risks of discrimination, bias and intolerance in automated decision-making systems to promote equality*. European Commission, CERV-2023-EQUAL, Grant Agreement No. 101144709.
- Eubanks, V. (2018) *Automating inequality: how high-tech tools profile, police, and punish the poor*. Picador, St Martin's Press.
- Gerards, J. and Xenidis, R. (2021) *Algorithmic discrimination in Europe: challenges and opportunities for gender equality and non-discrimination law*. Publications Office of the European Union.
- Göksal, S.I., Solarte-Vasquez, M.C. and Chochia, A. (2024) '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>
- Hacker, P. (2018) 'Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law', *Common Market Law Review*, 55(4), pp. 1143–1185. Available at: <https://doi.org/10.54648/COLA2018095>
- Joamets, K. (2025) 'Towards fair AI in Estonia's public service: discussing and disseminating bias prevention in automated decision-making', *TalTech Journal of European Studies*, 15(3), pp. 201–224. Available at: <https://doi.org/10.2478/bjes-2025-0037>
- Kerikmäe, T. and Pärn-Lee, E. (2021) 'Legal dilemmas of Estonian artificial intelligence strategy: in between of e-society and global race', *AI & Society*, 36, pp. 561–572. Available at: <https://doi.org/10.1007/s00146-020-01009-8>
- Kerikmäe, T., Joamets, K. and Metin, E. (2025) 'Automated decision-making in Estonia's public sector: innovation, regulation, and challenges about non-discrimination', in F. Decarolis, B. Marchetti and L. Torchia (eds) *European digital regulation and its impact on Member States, European Union and its neighbours in a globalized world*, 27, pp. 149–186. Springer. Available at: https://doi.org/10.1007/978-3-032-06490-5_6
- Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017) 'Inherent trade-offs in the fair determination of risk scores', *Proceedings of Innovations in*

- Theoretical Computer Science*, pp. 1–23. Available at: <https://doi.org/10.48550/arXiv.1609.05807>
- Laux, J., Wachter, S. and Mittelstadt, B. (2024) ‘Trustworthy artificial intelligence and the European Union AI Act: on the conflation of trustworthiness and acceptability of risk’, *Regulation & Governance*, 18(1), pp. 3–32. Available at: <https://doi.org/10.1111/regg.12512>
- Lips, S. and Joamets, K. (2025a) *Educational material for data scientists, data protection specialists and developers*. Project 101144709, EquiTech. Tallinn University of Technology. Unpublished training material.
- Lips, S. and Joamets, K. (2025b) *Study material for public servants*. Project 101144709, EquiTech. Tallinn University of Technology. Unpublished training material.
- Mittelstadt, B. (2016) ‘Automation, algorithms, and politics: auditing for transparency in content personalization systems’, *International Journal of Communication*, 10, pp. 4991–5002.
- O’Neil, C. (2016) *Weapons of math destruction: how big data increases inequality and threatens democracy*. Crown Publishing Group.
- Pasquale, F. (2015) *The black box society: the secret algorithms that control money and information*. Harvard University Press. Available at: <https://doi.org/10.4159/harvard.9780674736061>
- Pilving, I. and Mikiver, M. (2020) ‘A Kratt as an administrative body: algorithmic decisions and principles of administrative law’, *Juridica International*, 29, pp. 47–61. Available at: <https://doi.org/10.12697/JI.2020.29.05>
- ‘Council regulation (EU) 2016/679 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)’ (2016) *Official Journal* L 119. Available at: <http://data.europa.eu/eli/reg/2016/679/oj> (Accessed: 5 August 2026).
- ‘Council regulation (EU) 2024/1689 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. Available at: <http://data.europa.eu/eli/reg/2024/1689/oj> (Accessed: 5 August 2026).
- Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J. (2019) ‘Fairness and abstraction in sociotechnical systems’, *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 59–68. Available at: <https://doi.org/10.1145/3287560.3287598>
- Smuha, N.A. (2021) ‘From a “race to AI” to a “race to AI regulation”: regulatory competition for artificial intelligence’, *Law, Innovation and Technology*, 13(1), pp. 57–84. Available at: <https://doi.org/10.1080/17579961.2021.1898300>

- Wachter, S. (2020) 'Affinity profiling and discrimination by association in online behavioural advertising', *Berkeley Technology Law Journal*, 35(2), pp. 367–430. Available at: <https://doi.org/10.2139/ssrn.3388639>
- Wachter, S., Mittelstadt, B. and Floridi, L. (2017) 'Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation', *International Data Privacy Law*, 7(2), pp. 76–99. Available at: <https://doi.org/10.1093/idpl/ix005>
- Wachter, S., Mittelstadt, B. and Russell, C. (2018) 'Counterfactual explanations without opening the black box: automated decisions and the GDPR', *Harvard Journal of Law & Technology*, 31(2), pp. 841–887. Available at: <https://doi.org/10.2139/ssrn.3063289>
- Wachter, S., Mittelstadt, B. and Russell, C. (2021) 'Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI', *Computer Law & Security Review*, 41, article 105567. Available at: <https://doi.org/10.1016/j.clsr.2021.105567>
- Weerts, H., Xenidis, R., Tarissan, F., Palmer Olsen, H. and Pechenizkiy, M. (2023) 'Algorithmic unfairness through the lens of EU non-discrimination law: or why the law is not a decision tree', *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 805–816. Available at: <https://doi.org/10.1145/3593013.3594044>
- Zolkin, V., Chochia, A. and Hoffmann, T. (2023) 'Automated decision-making in the EU Member State's public administration: the compliance of automated decisions of the Estonian Unemployment Insurance Fund with Estonian administrative procedure law', *European Studies: The Review of European Law, Economics and Politics*, 10(2), pp. 178–202. Available at: <https://doi.org/10.2478/eustu-2023-0017>