POLICY **BRIEF** 2022:25

Perspectives into topical issues in society and ways to support political decision making.

This publication is part of the implementation of the 2021 Government plan for analysis, assessment and research (**tietokayttoon.fi/en**). The producers of the information are responsible for its content and it does not necessarily represent the views of the Government.

Promoting equality in the use of Artificial Intelligence – an assessment framework for non-discriminatory AI

Atte Ojanen, Anna Björk, Johannes Mikkonen

Demos Helsinki

Algorithmic discrimination threatens citizens' fundamental rights and poses a challenge to public governance. However, AI systems can also be used to **promote equality**, in line with the Finnish Non-Discrimination Act. This policy brief summarises the main findings of the Avoiding AI Biases -project. It introduces an assessment framework for non-discriminatory AI systems and associated policy recommendations, which support its application to public governance.

Algorithmic discrimination needs to be recognized and governed urgently

Increasingly advanced artificial intelligence (AI) systems and algorithmic decision-making are promised to revolutionize societies for the better. In the last five years, AI has developed at an unprecedented pace. The current wave of AI is characterised by the abundance of data, increasing computing power and, in particular, machine learning algorithms such as neural networks and deep learning.

Although the application of AI is still relatively moderate in countries such as Finland, these systems are already having major societal impacts, which also extend to people's fundamental rights. The use of machine learning algorithms, which process massive amounts of data can come to reproduce and reconfigure existing forms of inequality and discrimination, especially indirect and intersectional discrimination. For example, automated decision-making systems trained on biased and unrepresentative data can lead to discrimination on unprecedented scale. Discriminatory structures created, maintained and exacerbated by AI systems are therefore one of the major challenges that need to be addressed by public administration and governance.

Non-discrimination is a fundamental right that has an established role in both the EU and the Finnish legislation, referring to the unjustified treatment of individuals or groups based on protected characteristics such as age, ethnic origin, disability, sexual orientation, religion or belief. It is, therefore, necessary to raise awareness of both direct and indirect forms of algorithmic discrimination. However, the government also urgently needs tools to act upon the risk of discrimination, if it wishes to exploit the potential of AI for advancing equality without risking the fundamental rights of citizens.

In practice, algorithmic bias can arise from**:**

1. Problem formulation and design: Questionable basis for developing the system and poorly defined objectives without consulting the people affected by the system can result in biases. Equality considerations should be factored into the start of the design and problem formulation phase of the AI lifecycle, with the active participation of affected and vulnerable communities.

2. Data: Data always reflects existing power relations in a society. Algorithms trained on historical data can replicate and reinforce existing social inequalities, even if the

data is representative and protected attributes are excluded from it. This happens because AI systems learn to use ‘proxy variables’ — such as postcodes — to replicate historic biases.

3. Algorithmic model: The algorithms themselves can also be biased due to misaligned target variables, fairness metrics and comparison classes. While these can be partly alleviated by different debiasing techniques and organisational practices like diversity and participatory design, a lack of transparency and explainability of AI systems remains an additional challenge and a risk factor for discrimination.

4. Deployment of the system in practice: Implementation of an AI system in a context where it’s not intended to be used can lead to major discriminatory outcomes. Moreover, changes in the target population, data drift and other incremental changes should be monitored with eye to their equality impacts.

As bias does not merely lie in technical aspect such as data or the algorithm, algorithmic discrimination is unlikely to be tackled successfully by only relying on statistical and technical notions of fairness. Discovering and mitigating algorithmic discrimination requires an interdisciplinary socio-technical perspective, that considers the societal context around the use of AI alongside its technical components. This includes diversity of development and participatory design practices with affected stakeholders and minority groups. As algorithms affect society at large, public oversight and deliberation is essential.

Algorithmic impact assessments have recently gained traction as a way of ensuring ethical application of AI in a transparent manner, but so far they have only provided limited focus on equality and non-discrimination. The assessment framework developed for the Prime Minister’s Office in Finland in 2022, now introduced here, combines the evaluation of discriminatory risks of AI systems with the promotion of equality. Thereby it seeks to allow governments and public officials to steer technological innovation and development while protecting the fundamental rights of citizens.

The assessment framework for non-discriminatory AI systems

Building the assessment framework

The assessment framework was developed within the "Avoiding AI biases: a Finnish assessment framework for non-discriminatory AI applications" project of the Finnish Government's analysis, assessment and research activities. The team consisted of experts and researchers from Demos Helsinki, the University of Turku and the University of Tampere. The aim of the research project was to identify the possible risks to fundamental rights and non-discrimination posed by AI and machine learning systems in use within Finland, in light of the Non-Discrimination Act.

First, the University of Turku conducted a national mapping of AI applications in use within public sector with possible impacts on fundamental rights. The mapping shows that the adoption of AI systems in Finland is still at a modest level. While the public sector is reasonably aware of the discriminatory risks of AI systems, there is no clear model of cooperation between authorities to tackle them. The research also highlighted the different responsibilities of private and public sectors in fighting discrimination and concluded that the global production chains of AI systems pose a severe challenge to non-discrimination due to their lack of transparency.

In the second stage, the Tampere University overtook an in-depth analysis of the discriminatory risks of AI systems, the methods developed to identify and prevent them, and the potential challenges of using these methods. The research shed light on the risk profile of algorithmic discrimination in socio-technical contexts, its technical and non-technical causes and the challenges related to the identification and prevention of discrimination whether by debiasing or organizational practices.

The final part of the research focused on developing an assessment framework for non-discriminatory AI systems, led by Demos Helsinki. The framework helps to identify and manage risks of discrimination, especially in public sector AI systems,

and to promote equality in the use of AI. Furthermore, policy recommendations for utilising the framework were developed and co-produced with stakeholders.

Format and stakeholders

The main intended audience of the framework is the public sector and civil servants who can use it to assess possible discriminatory effects of AI systems when planning, procuring or deploying them. Indirectly it also acts as a tool for AI developers to assess their systems and processes, especially if intended for public use. The framework takes into account the Finnish context, namely the Non-Discrimination Act that obligates authorities, education providers and employers not only to prevent discrimination, but also to promote equality. In practice, this can mean assessing the impacts of AI systems on equality and associated equality planning.

As per the lifecycle model, the assessment framework emphasises that the discriminatory impacts of AI systems can arise at different stages of development. The lifecycle model allows for a holistic approach to managing the risks of algorithmic discrimination, rather than merely assessing the system against a set of static criteria. As such, the framework serves as an algorithmic impact assessment process for addressing risks of discrimination and promoting equality throughout the lifecycle of an AI system. The framework and its questions are structured according to three phases, following a lifecycle model:

1) Design, i.e. the initial assessment and definition of the AI system's objectives, motivations for use, necessity and equality impacts. The importance of the planning and design stage is highlighted when a public actor commissions or deploys an AI application for which it has limited access to the development process. From the perspective of equality and its promotion, the design must focus on the objectives of the system, the participation of vulnerable stakeholders and how the application interacts with prevailing social inequalities.

2) Development, covering three areas: data and its preparation, training of the algorithmic model and validation of the model. Unrepresentative and mislabeled training data is one of the most important sources of algorithmic discrimination. Similarly, biases can manifest in the model training phase. In addition, discrimination can emerge during the testing and validation phase of the model. Within the development phase, technical pre-, in- and post-processing methods and fairness metrics can be applied to correct algorithmic biases to a degree.

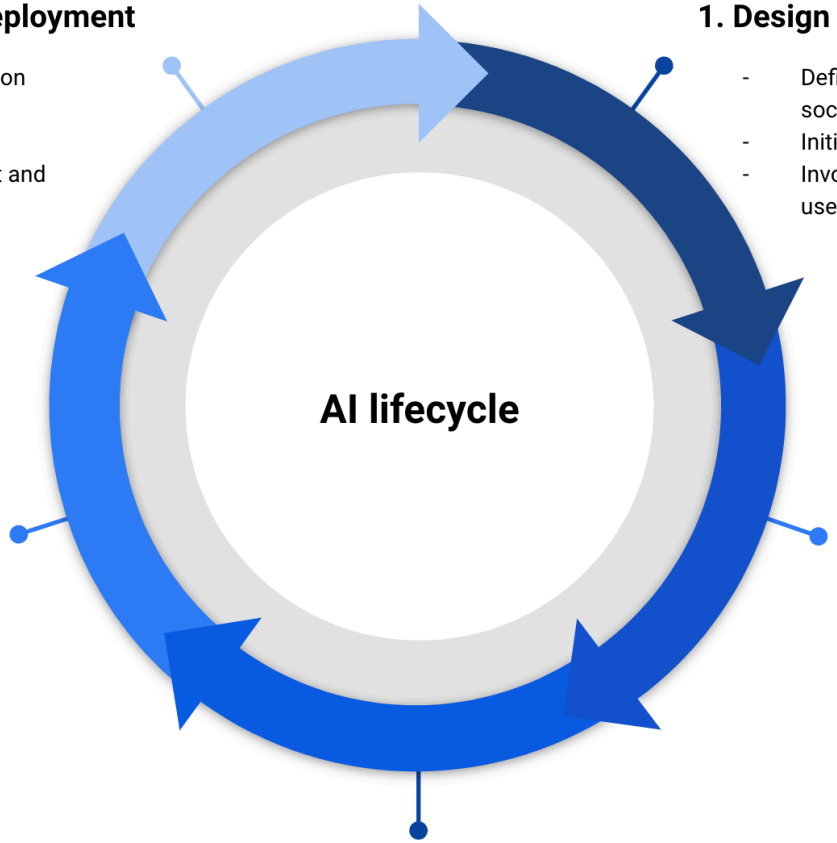
3) Deployment, where important issues include ensuring human oversight, transparency and monitoring the system's equality impacts in practice. The reliability and fairness of the AI system is likely to be undermined, leading to discrimination, if it is deployed without sufficient oversight, monitoring and maintenance. It is noteworthy that the AI lifecycle does not end at deployment, but the monitoring of its performance and equality impacts should feed back to the development.

Using the framework

The assessment framework divides each stage of the AI lifecycle into questions, which describe the discrimination risks throughout the stages of the AI system. they address possibilities for mitigating the risks as well as opportunities to promote equality. The respondent evaluates, how well the issues are taken into account in the development of the AI system; good (0 pts), partially (0.5 pts) or not at all (1 pts). Black triangles indicate essential questions, which should be at least partially taken into account in the development. Each section also has questions about promoting equality, which can lessen the risk. Based on the responses, a risk score is calculated both for each section and their total score to guide whether the AI system's development or deployment should be continued as is (to continue development, no section or the total score should be in the red, i.e. bottom 50%). We encourage users to tailor the assessment framework to specific use cases by modifying it as the risks of discrimination are always contextual, and the framework is not fit for every context. Easy-to-use Excel version is also available (see read more -section at the end). The evaluation framework can be used for example:

- To develop procurement practices and guide procurement so that the risks of discrimination in the AI system being procured are minimised.
- Clarifying roles and responsibilities: the assessment framework can be used as a tool to clarify the roles and responsibilities of different actors in tackling risks of discrimination, both within and between organisations
- Training and awareness-raising: using and communicating the assessment framework's results will raise understanding of the discriminatory risks of AI. It can also be used to train and develop the skills of public sector employees, create common practices and reinforce good governance principles.
- Responding to and correcting errors: The assessment framework can help both public administrations and developers identify the source of discrimination and inequality in the system, enabling corrective measures.
- To identify measures to promote equality in the development and deployment of the AI application.

Design | Development | Deployment



1. Design

- Defining the objectives of the AI system, taking into account the social and legal context of the use case
- Initial assessment of the system's equality impacts
- Involvement of stakeholders such as minorities in defining the use of the AI system

2.1. Data

- Data collection, verification of origin and quality
- Data preparation and validation to improve representativeness
- Pre-processing of data to reduce bias

2.2. Model training

- Assessing the fairness of the model using fairness metrics and selecting model features
- Diversity in development and co-development with stakeholders
- In-processing such as regularisation and optimisation constraints to correct biases

2.3. Model testing

- Testing and evaluating the fairness and equality of the model in relation to different individuals and groups
- Documenting the characteristics and functioning of the model
- Post-processing methods to promote equality

3. Deployment

- Defining the scale of use and the level of automation
- Monitoring and tracking the equality impacts and discriminatory risks of the system in practice
- System maintenance, transparency, external audit and ensuring redress mechanisms

1. DESIGN

The process of conceptualization and design is one of the most important steps from the perspective of equality and non-discrimination, especially in public sector procurement settings where the commissioning party lacks direct access to the AI development process. This phase defines the rationale, objectives and motivations for the use of the AI system, taking into consideration the context of system use.

Responsibility: The primary responsibility for the design lies with the organisation commissioning and deploying the AI system, subject to collaboration with the technical developers.

CONCEPT AND PRE-DESIGN		
Topic	**Questions**	**Considered**
Goals for system development	Have the objectives and grounds for developing the AI system been assessed as being in the public interest and generally acceptable? What societal problem does it seek to address?	▲
System objectives	Are the goals and objectives of using the AI system clearly defined, including problem formulation, operation, use case, users and people affected by the system?	
Necessity and proportionality	Is the envisioned system or the means of using AI necessary and proportionate to achieve the stated objective?	
Social context	Have the historical and social inequalities related to the social context of the use case been taken into account so that the system does not reproduce these inequalities?	
Sectoral risks	Is the AI system used in a public or semi-public sector to support decision-making or activities that directly or indirectly affect citizens' fundamental rights (e.g., health, safety, access to education and work)?	▲
EQUALITY IMPACTS		
Resourcing equality assessment	Has the organisation designated a person or a team responsible for assessing the system's equality impacts, with knowledge of the requirements and possibilities stated in non-discrimination law? Can this person be contacted?	
Differential treatment	Can the use of the system lead to different treatment on the basis of prohibited grounds for discrimination (e.g. age, nationality, language, religion, disability or other personal grounds)? How has this been taken into account as part of the design?	▲
Group impacts	Which group or groups may be disproportionately affected, either directly or indirectly, by the use of the AI system? Are these individuals in positions of particular or intersecting vulnerabilities? Is the disadvantage related, for example, to the accuracy, availability, or concrete effects of the service?	
STAKEHOLDER PARTICIPATION		
Planning participation	Is there a plan and a designated person responsible for the active involvement of stakeholders and citizens in the system design and development process?	
Stakeholder consultation	Will affected communities and groups be consulted on the use case of AI, its benefits and the appropriate definition of its objectives? How will it be ensured that their concerns and experiences are effectively taken into account in the design of the system?	
Participation of the most vulnerable	Do the stakeholders consulted and involved in the planning process include vulnerable groups, for example through local and non-governmental organisations?	▲
ACCESSIBILITY		

The needs of minorities	Has accessibility been taken into account in considering the needs of different minorities when designing the system?	
Reasonable accommodations	Has the system been designed to ensure that people with disabilities receive appropriate reasonable accommodations?	▲
Equal opportunities	Does the system benefit people equally and ensure equal opportunities to use and access the service for all?	
ACCOUNTABILITY		
Procurement policies and communication	Are procurement policies, such as requirements concerning non-discrimination and equality, clearly communicated and passed on to the technical developers of the AI system? How will continuous communication throughout the development phases be ensured?	
Possible misuse	Have the risks associated with potential misuse of the system been documented and prepared for as part of the design? Has both the unintentional misuse by users as well as intentional abuse by malicious actors been considered?	
Roles and responsibilities	Has the division of responsibilities, roles and transparency of the development been defined both between the participating organisations and within their teams at this stage? Are employees from different departments involved in the development, including leadership, domain experts, legal and technical?	
PROMOTING EQUALITY		
Obligation to promote equality	Does the design take into account the broad duty of public authorities to promote equality? Is the system explicitly designed to promote equality, for example by better considering the needs of people with disabilities?	
Identifying discrimination	Is the AI being used to detect, prevent or tackle discrimination in current decision-making or services?	
Positive action	Will AI be used for positive action as a means to promote substantive or de facto equality (e.g., to implement language quotas in schools)?	
RISK SCORE (0-4, 4,5-9, 9,5-17)		**X / 17**

2.1. DEVELOPMENT - DATA

Training data that reflects past discrimination and inequalities is one of the main causes of discrimination in the use of AI systems. Actions to take during the data collection and pre-processing phase include 1) data collection activities mindful of the origins, quantity and quality of the data, 2) data preparation measures to ensure representativeness and lack of harmful content, and finally 3) data validation and balancing.
Responsibility: The technical developer of the AI system is mainly responsible for the development phase, but it also requires continuous cooperation with the organisation that commissioned it.

DATA COLLECTION		
Topic	**Questions**	**Considered**
Origin of data	Is the training data the commissioning or developing organisation's own, pre-validated data? If external data sources are used, is there certainty about their purpose, sampling, quality, ownership, access or storage constraints?	
Volume of data	Does the system make use of large amounts of data from multiple sources? How are their origin, quality and compatibility ensured?	
Data quality	How has it been taken into account that the quality of data may not be uniform across groups or sources? For example, has quality been assessed in terms of the distribution of target variables, characteristics of population groups, effects of aggregation or missing data fields?	

Personal data	Does the training data include special categories of personal data or other (sensitive) personal characteristics such as age, religion or ethnic origin? If so, is this necessary for the functioning of the system and its equality? Are data protection issues such as privacy and fairness of data processing considered?	▲
Informing the data subjects	Is the collection and use of data (including its justification, purpose and categorizations) communicated clearly and accessibly to the people concerned? Have individuals given their informed consent to the use of the data?	
Data storage and documentation	Is the data stored and documented securely, covering the above points, so that training data can be recreated, verified and updated if necessary? Is there a designated person responsible for data management, updating and compatibility?	
DATA PREPARATION		
Ensuring data quality	Is it ensured that the data classes and labels used are appropriate for the purposes of the system?	▲
Responsible data annotation	Is the person or team responsible for labeling data points qualified to prepare and annotate the data used by the system, especially sensitive data (e.g., demographic or behavioural data)?	
Precedent use	Has the applicable data or dataset been used in similar applications before? For example, are there successful trials or precedents for the use of this or similar data?	
Causality behind data	Is the choice of data justified in the light of the causal structure of the use context? For example, is the information used relevant from a causal perspective in the decision-making context or is it missing relevant factors?	
Historical inequalities	What kind of aims, assumptions, beliefs and possible forms of historical or social inequality does the data reflect?	▲
DATA VALIDATION		
Representativeness	Is the training data representative of the different population groups affected by the system, especially minorities? Has the representativeness of the data been compared with relevant reference population (e.g., demographic statistics)?	▲
Sample and label biases	Have potential sampling and labeling biases between groups in the training data been taken into account? Are data points labelled in a reliable, consistent and non-discriminatory manner?	
Rebalancing data	Has the training data been rebalanced if necessary, for example, by oversampling to improve representativeness and predictive accuracy? Has the need for oversampling in the case of minority groups been considered?	
Proxy variables	Have clear proxy variables for personal characteristics that may lead to discriminatory outcomes been identified and cleaned from the data?	▲
Offensive content	Has any discriminatory, offensive or psychologically harmful content (e.g., derogatory language or imagery that reinforces discriminatory stereotypes) been identified and removed from the data if necessary?	
PROMOTING EQUALITY		
Inclusive data collection	Has the data been specifically collected inclusively with and for the underrepresented groups to promote equality, while considering the privacy and data subject rights as set out in the GDPR?	
Non-binary data labels	Has the potential harm that binary labels and classifications can cause for minorities been accounted for, for example by applying more inclusive, non-binary variables or labels in the data?	

Data pre-processing	Has any other pre-processing of the data been performed at the development stage to correct for sampling, measurement and label biases and to promote equality (e.g., resampling and reweighing the data)?	
RISK SCORE (0-4, 4,5-8,5, 9-16)		**X / 16**

2.2. DEVELOPMENT - MODEL TRAINING

Discriminatory biases can also occur during the training phase of the model and algorithm due to choices of metrics and models, for example. The diversity of organisations and developers is also important, as is training on equality

ASSESSING MODEL FAIRNESS		
Topic	**Questions**	**Considered**
Defining baselines	As part of the modeling, have internal baselines been defined for the phenomenon or characteristic to be modelled (cf. values of the target variable) within demographic groups?	
Use of fairness metrics	Has the development process assessed equality impacts of algorithms? For example, have fairness metrics such as conditional statistical parity or equal error rates been used to identify potential biases and risks of discrimination in the AI system?	▲
Choice of appropriate fairness metrics	Is the selection of appropriate equality and fairness metrics justified, taking into account the use case and applicable legislation? Are the methodological assumptions and limitations of the applied fairness metrics identified?	
Selecting comparison classes	Are the comparison classes (e.g., ethnicity, age) that are used to assess the fairness of the model relevant for the purpose of identifying risks of direct and indirect discrimination? Have analyses been carried out with respect to, for example, prohibited grounds of discrimination or proxy variables that correlate with them?	
Assessing multiple discrimination	Have analyses been conducted intersectionally between subgroups to account for multiple discrimination?	
MODEL AND FEATURE SELECTION		
Model operation	Is it clearly documented how the model works, what it is intended to predict, classify or recommend (including choice of target variables, classification task, performance objectives) and what its intended effects are? How will its performance, fairness, and other relevant factors be measured and evaluated?	▲
Model limitations	Has the suitability of the chosen AI model (e.g., linear regression, neural network) been assessed? Have its potential limitations been identified and documented? Is the least complex, previously validated and transparent model possible used to perform the task?	
Model vulnerabilities	Have the system features which may predispose the model to generate discriminatory results (e.g., lack of transparency or use of potentially discriminatory proxy variables) been identified and documented? How are these vulnerabilities addressed, especially in the case of an algorithm that is continuously learning from data in its operating environment?	
Selection of target variables	Is it ensured that the model does not use prohibited grounds of discrimination as target variables without a legally justified and legitimate reason?	▲
Target values for fairness	Have target values for the fairness of the model been defined and documented as part of the development of the model (e.g., acceptable	

	differences between groups in terms of false positives and false negatives)? Is a fairness objective explicitly included in the objective function of the algorithm?	
DIVERSITY IN DEVELOPMENT		
Interdisciplinarity and diversity of developers	Is the team developing the system inclusive and diverse in terms of demographics, values, education and acquired knowledge and skills? Does the team include people who have received training in equitable and ethical technology design?	
Cooperation with the deployer	Is there ongoing cooperation between the deploying organisation or other non-technical experts and the technical developers, in particular to improve equality impacts of the system?	
Co-development with stakeholders	Does the development process involve end-users and people affected by the system, for example through stakeholder co-development and workshops? How are the interests of particularly vulnerable people or under-represented groups taken into account?	
PROMOTING EQUALITY		
Knowledge and training on bias mitigation	Are resources, training or other measures in place to ensure that technical developers are able to incorporate equality considerations into their work? For example, is it ensured there are required competences to use the technical tools to identify discrimination and for debiasing algorithms?	
Use of bias transforming fairness metrics	Is the use of bias transforming metrics considered as part of the model training to promote de facto equal access to goods and services, provided the use of those metrics as target values is deemed justified in the use case context?	
In-processing methods	Have in-processing methods such as regularization, adversarial debiasing or optimisation constraints been used to advance equality in the model training phase?	
RISK SCORE (0-3, 3,5-7, 7,5-13)		**X / 13**

2.3. DEVELOPMENT - MODEL TESTING

The trained model must be tested and validated to ensure that it works as intended and that its outputs do not reflect discriminatory biases. Based on the results of validation and testing, the model may need to be revised and corrected to prevent discrimination, ensure equal opportunities and to improve equitability in outcomes. The functioning of the revised model should also be carefully documented.

MODEL TESTING PROCESS		
Topic	**Questions**	**Considered**
Model test and evaluation data	What kind of test data is used to evaluate the performance of the model (e.g., prediction quality, accuracy, error rates)? Is it sufficiently different from the training data to avoid model overfitting and to identify potential biases?	
Model fairness testing	Has the model been tested for discriminatory bias? For example, has the model been evaluated and compared using appropriate fairness metrics?	▲
Testing for direct discrimination	Has the model and its performance been assessed on individual level for risks of direct discrimination?	
Testing for indirect discrimination at group-level	Has model performance and fairness been tested across different population groups for risks of indirect discrimination (including intersectional groups)? Are the errors by the model evenly distributed across population groups and do they appear more harmful to certain	

	groups?	
MODEL VALIDATION		
Justifiable differences	If there are differences in model performance or outcomes for different individuals or groups (e.g., protected groups), are the differences necessary, proportionate and justified? Have the differences in model scores or predictive accuracy been assessed in relation to the case law related to the use case?	▲
Bias mitigation	Have different strategies been identified, compared and applied to avoid and mitigate discriminatory bias?	
Trade-offs in target values	Have potential trade-offs been identified (e.g., between the overall accuracy and fairness of the model or between different target values of fairness)? Is the approach to resolving potential conflicts and trade-offs documented in a transparent manner?	
Long-term effects	Have the longer-term impacts of the model and the use of the system been assessed? Have processes been designed or implemented to monitor long-term impacts, such as running model scenarios or simulations? How has the potential model or data drift been addressed?	
MODEL DOCUMENTATION		
Alternative approaches	Has the need and use of alternative approaches such as different modelling approaches or complementary decision-making practices been assessed from a non-discrimination perspective?	▲
Changes in model performance	Has the behaviour of the model (e.g., overall and relative predictive accuracy) or estimated group-specific effects changed as a result of testing and potential corrective actions?	
Documenting model functioning	Have the changes in the functioning (including the expected effects on equality), training, testing and use case of the AI model been comprehensively documented on the basis of the testing and evaluation process?	
Communication and training with deployer	Has the organization that deploys the AI system and its users been provided with documentation, licenses and sufficient training to enable them to evaluate the system in practice and to avoid misuse?	
PROMOTING EQUALITY		
Improving the position of the disadvantaged	Has the possibility of resolving trade-offs between the model's target values (e.g., accuracy and fairness) in favour of those belonging to underrepresented or marginalised groups been considered? Similarly, has consideration been given to using different models or decision criteria for different population groups to promote equality?	
Documenting model equality impacts publicly	Are model development and evaluation activities, especially fairness and equality assessments properly documented, for example by using model cards? Is this information also available for external audit?	
Post-processing methods	Have post-processing methods (e.g., changing decision thresholds or balancing representation within prediction classes) been used to make model's scores more fair and equitable, while mindful of the legal justification for these methods?	
RISK SCORE (0-3, 3,5-6,5, 7-12)		**X / 12**

3. DEPLOYMENT

Deployment is a key part of the AI lifecycle and a potential source of risks for discrimination. As an example, if an AI system is deployed in a different environment and population than it was designed for, its accuracy and behaviour may be compromised, leading to potentially discriminatory effects. Attention must also be paid to the monitoring, transparency, maintenance and regular auditing of the system's equality impacts.
Vastuu: Deployment is the joint responsibility of the organisation that commissioned the AI system and its technical developer.

Topic	Questions	Considered
HUMAN-IN-THE-LOOP		
Level of automation	Are the decisions made by the system always subject to human control and validation (i.e., not fully automated)? Are the operator's responsibilities and obligations clearly defined, such as to what extent the AI-generated recommendation or prediction should guide actual decisions?	▲
Trust in the system	How likely are end-users and citizens to trust and delegate decisions to an AI system, and how will this be taken into account in the deployment? For example, will operators be able to interpret the system's output correctly?	
Effects related to the user interface	Could the features of the system's user interface (e.g., how results are presented to users) have an unexpected and negative impact on its use, for example by contributing to careless use or exacerbating users' prejudices about people? How has this been addressed?	
MONITORING		
Scale of use	Will the system be deployed on a smaller scale or in a pilot setting before moving to full-scale deployment to safeguard against possible discriminatory results, especially in cases where an automated decision-making system is used in a critical area of society?	
Target population and use	Has the system been deployed within the same target group and societal use case for which it was designed and trained for? Have there been changes in the social context that affect the need for the system, its suitability or its equality impacts?	
Changes in the input data	Is there active monitoring of whether the data the AI system encounters in the real-world environment sufficiently matches the training and test data? Does the system notify if significant changes occur either in the data or the model (e.g., model learns new categories, classes or the observed probability distributions change)?	
Monitoring equality impacts	Have timeframes and indicators been established to monitor and evaluate the performance of the system in terms of equality and discrimination throughout its life cycle? Have users of the system been trained to monitor these indicators?	▲
Defining the monitoring indicators	Will the deployment and monitoring metrics (e.g., for predictive accuracy and error rates) include specified standards that, if exceeded or not met, will trigger an alert and a system review? Does the system then generate an error log?	
Monitoring discriminatory impacts	Have any discriminatory impacts or risks for such impacts across individuals or groups been observed during the deployment of the AI system? How are they suspected to have arisen? Have local stakeholders and NGOs been involved in monitoring and reporting on discrimination, for example through surveys or feedback systems?	▲
MAINTENANCE		
System update and feedback loop	Will the generated data and/or insights be fed back into earlier stages of planning and development to improve equality impacts of the system? Are the risks associated with feedback loops considered at an operational and strategic level?	
Maintenance and up-keep of the system	Have maintenance and management responsibilities been established for the AI system's life cycle together with the technical developer of the AI system? How will the system's accuracy, equality impacts and operational safeguards be maintained over time, especially as	

	the societal context and demographics change?	
Decommissioning the system	Has a timeline for system use and support been defined, including how it will be replaced or decommissioned if necessary? Has it been considered how the decommissioning affects the fundamental rights of affected people? For example, what happens to the collected data when the application is no longer used?	
TRANSPARENCY		
Responsibility for system errors	Are the parties and organisations responsible for using the system and responding to any errors clearly defined? Are errors transparently reported to appropriate external parties?	
Explainability to affected individuals	Are the people affected by the system, informed about the use of AI, the decision-making process and the reasons behind it in a transparent and accessible way? Is the functioning of the system and the decisions taken explainable and understandable to affected persons?	▲
Documenting the decision-making process	Is the functioning of the system and the decisions taken comprehensively recorded throughout the process, including changes made to the system during or after deployment?	
PROMOTING EQUALITY		
Right to legal remedies	Have affected persons' rights to effective remedies (e.g., access to court and legal aid) been taken into account and ensured during the deployment of the AI system or service?	
Open auditing	Is the system and its decisions openly available for external auditing (e.g., to public authorities, researchers, civil society), to the extent that is possible under existing legislation?	
Rectifying decisions	Have feedback mechanisms been established to allow affected persons and communities to seek redress, compensation or challenge decisions connected to the system, when necessary? How will individuals be compensated for any damage caused by incorrect decisions?	
RISK SCORE (0-3,5, 4-8, 8,5-15)		**X / 15**

TOTAL RISK SCORE	**X / 73**

0-18	**19-38**	**39-73**
Relatively safe to proceed with use	Reconsider use	Do not use before changes

Policy recommendation to support the implementation of the framework:

1) Raising the public awareness of algorithmic discrimination: Knowledge and awareness of the nature and societal relevance of algorithmic discrimination is growing, but this trend needs to be actively supported through communication, education and capacity building, and the identification of responsibilities. Under the Finnish Non-Discrimination Act public authorities, education and training providers and employers have a responsibility to promote equality. Therefore, policy guidance should ensure that these key groups have a good understanding of the root causes and risks associated with algorithmic discrimination. Furthermore, their understanding of how AI can be used to promote equality should be strengthened.

2) Increasing cooperation between different stakeholders in the responsible development of AI systems: The assessment framework builds on a lifecycle approach in identifying the responsible parties involved in the different stages of AI development. Clarifying the roles, responsibilities and cooperation between the deployer, developer and affected stakeholders across different parts of the AI lifecycle is essential for ensuring non-discrimination. In addition to the immediate cooperation related to the use of the framework itself, coordination across organizational boundaries is needed to mainstream its use among private and public actors using AI.

3) Promoting equality in the use of AI through proactive regulation and tools: As awareness of algorithmic discrimination grows and the use of AI systems becomes more widespread, regulation of digitalisation and AI is becoming increasingly pertinent. It is recommended that ministries mandate government agencies and public procurers to use the framework or similar tools to tackle algorithmic discrimination within their area of governance. National and EU-level AI policy should focus increasingly on the use of AI to promote equality and associated tools.

Read more

Project report:
Ojanen, A., Sahlgren, O., Vaiste, J., Björk, A., Mikkonen, J., Kimppa, K., Laitinen, A., Oljakka, N. (2022): *Algoritminen syrjintä ja yhdenvertaisuuden edistäminen. Arviointikehikko syrjimättömälle tekoälylle*. Valtioneuvoston tutkimus- ja selvitystoiminnan julkaisusarja 2022: 54 (in Finnish). https://urn.fi/URN:ISBN:978-952-383-404-0

Project tool:
The assessment framework for non-discriminatory AI systems (Excel tool).

More information:

Atte Ojanen (Demos Helsinki) is a research coordinator at Demos Helsinki. His expertise includes the democratisation of AI systems, ethics and deliberative democracy. Contact: atte.ojanen@demoshelsinki.fi

Anna Björk (Demos Helsinki) is the research area lead at Demos Helsinki. She holds a PhD in political science and has expertise in democracy, citizenship and inclusion. Contact: anna.bjork@demoshelsinki.fi

Johannes Mikkonen (Demos Helsinki) is a senior policy expert at Demos Helsinki. He has extensive expertise in digital governance and the societal impacts of emerging technologies such DLTs and AI. Contact**:** johannes.mikkonen@demoshelsinki.fi

We wish to thank our project partners from the Tampere University (Otto Sahlgren, Arto Laitinen) and the University of Turku (Juho Vaiste, Nea Oljakka, Kai Kimppa) for their invaluable contribution to the project and the assessment framework.

Avoiding AI biases: a Finnish assessment framework for non-discriminatory AI applications was implemented as part of the 2021 Government plan for analysis, assessment and research.

Chair of the steering group:
Katriina Nousiainen, Senior specialist
Ministry of Justice Finland, katriina.nousiainen@gov.fi

DEMOS HELSINKI Tampere University

UNIVERSITY OF TURKU Government's analysis, assessment and research activities

POLICY **BRIEF** is a series of articles for government analysis, assessment and research. It gives perspectives on topical issues in society and ways to support political decision-making. The articles are published on our web pages at **tietokayttoon.fi/en**.
© Prime Minister's Office