Preventing the potential discriminatory effects of the use of artificial intelligence in local services

Policy Brief

October 2020

ePaństwo Foundation

INTERCULTURAL CITIES
CITÉS INTERCULTURELLES

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Table of Contents

Written by Krzysztof Izdebski

Intercultural Cities Unit,
Council of Europe©

Council of Europe, October 2020

Krzysztof Izdebski. Board Member and Policy Director of ePanstwo Foundation (EPF) and Board Member of Consul Democracy Foundation. He is a lawyer specialized in access to public information and re-use of public sector information. He is the author of publications on freedom of information, technology and public administration including "Transparency and Open Data Principles: Why They Are Important and How They Increase Public Participation and Tackle Corruption" and recently published "alGOVrithms. The State of Play. Report on Algorithms Usage in Government-Citizens Relations in Czechia, Georgia, Hungary, Poland, Serbia, and Slovakia.

This policy brief was produced as a background paper based on the webinar organized by the Intercultural Cities Programme of the Council of Europe and Epaństwo Foundation on 21 September 2020.

1. Introduction

1.1 Background

Municipalities provide a wide range of public services to their citizens and increasingly this is supported by technologies including Automated Decision Making (ADM) tools and Artificial Intelligence (AI) solutions. The deployment of IT tools in public services has brought new challenges and potential risks of bias, prejudice towards certain categories of citizens, and discrimination. Such risks were, for example, detected in the Dutch SyRI system used by national and local authorities to detect housing or social security fraud, smart water meters in several cities in Europe or AI applications used in staff recruitment.

Some cities - like New York - have already implemented measures to prevent such irregularities, others are only starting to consider what steps should they take. Intercultural cities develop policies and expertise in social inclusion and equality, prevention of discrimination, and raising awareness around important societal challenges. It is useful for decision-makers to also understand the potential biases and risks of AI and learn about ways of mitigating such risks. The experience of advanced cities could help build trustworthy and ethical AI.

The Intercultural Cities Programme held a webinar about the challenges Artificial Intelligence and Algorithmic Decision-making present for local authorities, in particular in relation to (anti-) discrimination, inclusion, and the fight against hate speech. The webinar was prepared and led by Krzysztof Izdebski*, Policy Director of ePaństwo Foundation.

The report reflects the substantial content of the webinar and serves as a short guideline on preventing the potential discriminatory effects of the use of artificial intelligence in local services.

1.2 Glossary

Artificial Intelligence (AI): *Information technology that performs tasks that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours, or solving problems.*
- Directive on Automated Decision Making (Canada)

AI is only a type of algorithm which may cause discriminatory risk. As was stated in the Algorithm Charter For Aotearoa New Zealand, *the risks and benefits associated with algorithms are largely* ***unrelated to the types of algorithms*** *being used. Very simple algorithms could result in just as much benefit (or harm) as the most complex algorithms depending on the content, focus and intended recipients of the business processes at hand.*

Therefore, a better term to use is ***Automated Decision [Making] System*** *which according to* the Directive on Automated Decision Making (Canada) *includes any technology that either assists or replaces the judgement of human decision-makers. These systems draw from fields like statistics, linguistics, and computer science, and use techniques such as rules-based systems, regression, predictive analytics, machine learning, deep learning, and neural nets.*

To put it even simpler, after David Harel and his work Algorithmics - The Spirit of Computing (1987), we can compare an algorithm with a cooking recipe. While ingredients can be compared to input data, and a finished dish is a result, many activities such as selecting appropriate proportions at the right time or applied

methods of thermal processing are just an algorithm. From life experience, one can easily deduce that one mistake at the stage of preparing a dish can lead to failure in its taste and appearance.

According to C. Orwat in (2020) Risks of Discrimination through the Use of Algorithms:

Discrimination *is disadvantageous, unjustified unequal treatment of persons in connection with a protected characteristic.* Such characteristics can include "race" or ethnic origin; ancestry, home country, origin; gender; language; political opinion or viewpoint; religion and belief; disability; trade union affiliation; genetic characteristics or dispositions and health status; biometric characteristics; sex life, sexual identity or orientation.

To differentiate between the "traditional" and AI/ADM discriminatory it is important to take into account the below.

Taste-based discrimination is unequal treatment based on the personal, prejudiced dislikes or preferences of the decision-makers against or for a certain group of people or on dislikes or preferences for certain products.

Statistical discrimination is the unjustified unequal treatment of persons on the basis of surrogate information.

However, it is crucial to understand that, as algorithms are created by humans with all their biases included, the statistical discrimination can originate from the taste-based discrimination. These two phenomena are therefore very rarely independent of each other.

2 Consequences and examples of algorithmic discrimination

A few stories illustrate the discriminatory impact of some AI/ADM tools starting from the recent problem of A-level assessment algorithm in the United Kingdom where figures show 39.1% of 700,000 teacher assessments were lowered by at least one grade and it was especially visible among pupils from the lowest socio-economic background.

A-level results: almost 40% of teacher assessments in England downgraded

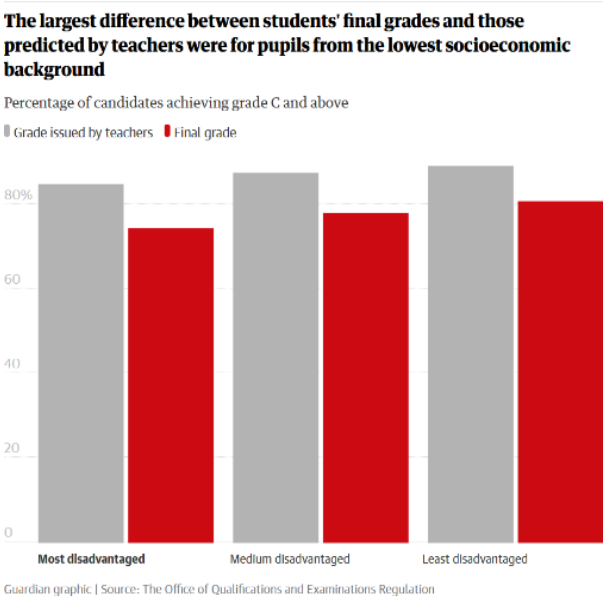
Ofqual figures show 39.1% of 700,000 teacher assessments were lowered by at least one grade

A-level results - live updates

2:09

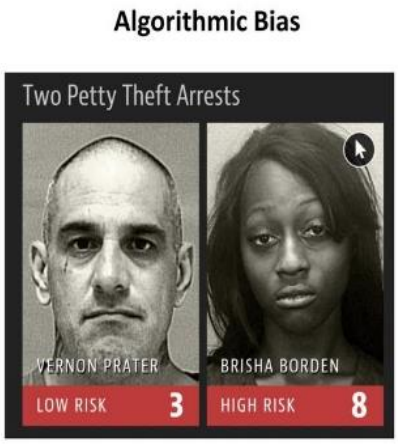
'I put my heart and soul into them': A-level students on downgraded results - video report

Teachers in England had nearly 40% of their A-level assessments downgraded by the exam regulator's algorithm, according to official figures published on Thursday morning as sixth-formers around the UK received their results.



According to some judges, the most frequent reason behind such problems evolves from the fact that *automatic decisions often fail to include an extensive evaluation of the circumstances of the case. By contrast with automatic decisions, civil servants can explain the background of a decision better and therefore delimit any dispute during the course of a review. Context is crucial to avoiding unwillingly biased decisions.*

A similar approach was shared by Eric Holder, former US Attorney General, who said referring to sentencing determination based on algorithms that *although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice.* This was said just after the scandal connected with the COMPAS, an AI/ADM tool used in the United States to predict the likelihood of committing a future crime. The Brisha Borden and Vernon Prater examples revealed that *data-driven, decision-making technologies used in the justice system to inform decisions about bail, parole, and prison sentencing are biased against historically marginalized groups*



Further, the European Commission sees that *certain algorithms, when exploited for **predicting criminal recidivism**, can display gender and racial bias, demonstrating different recidivism prediction probability for women versus men or for nationals versus foreigners.* The other example refers to *certain AI programmes for **facial analysis** which display gender and racial bias, demonstrating low errors for determining the gender of lighter-skinned men but high errors in determining gender for darker-skinned women.*

This led participants to discuss further the issue of statistical discrimination based on the following example.

> An energy supplier in Belgium refuses to supply electricity to persons living within a certain postcode area. For the energy supplier, this postal code area represents an area with many people with poor payment habits. Even solvent potential buyers are excluded from supply without taking into account their individual solvency.

In this case the **surrogate variable** is a "place of residence"

3 How discrimination in AI works

Based on the report by F. Z. Borgesius (2018) Discrimination, artificial intelligence, and algorithmic decision-making, Krzysztof Izdebski explained how AI/ADM can lead to discrimination in several ways:

> (i) how the "target variable" and the "class labels" are defined; (ii) labelling the training data; (iii) collecting the training data; (iv) feature selection; and (v) proxies as well as (vi), AI systems can be used, on purpose, for discriminatory ends.

Target variable and class labels *"by exposing so-called "machine learning" algorithms to examples of the cases of interest (previously identified instances of fraud, spam, default, and poor health), the algorithm "learns" which related attributes or activities can serve as potential proxies for those qualities or outcomes of interest." Such an outcome of interest is called a "target variable".* Class labels are connected with target variables. *Suppose a company wants an AI system to sort job applications to find good employees. How is a "good" employee to be defined? In other words: what should the "class labels" be? Is a good employee one who sells the most products? Or one who is never late at work?* Borgesius writes *that discrimination can creep into an AI system because of how an organisation defines the target variables and class labels.*

Labelling the training data An *AI system might be trained on biased data* [or] *problems may arise when the AI system learns from a biased sample.* Borgesius gives examples of the system created to sort out applications for University. *The training data for the computer programme where the admission files from earlier years* were gender and ethnicity biased, leading to fewer women and persons with immigrant background being accepted.

Collecting the training data *The sampling procedure can also be biased. For instance, when collecting data about crime, it could be the case that the police stopped more persons from an immigrant background in the past, leading the AI system to disproportionately identify persons of colour as potential perpetrators.*

Feature selection *Suppose that an organisation wants to automatically predict which job applicants will be good employees. It is not possible, or at least too costly, for an AI system to assess each job applicant completely. An organization could focus, for instance, on certain features, or characteristics, of each job applicant. By selecting certain features, the organisation might introduce bias against certain groups.*

Proxies: *Some data that are included in the training set may correlate with protected characteristics*. (…) *The training data do not contain information about protected characteristics such as skin colour. The AI system learns that people from a certain postal code were likely to default on their loans and uses that correlation to predict defaulting. Hence, the system uses what is at first glance a neutral criterion (postcode) to predict defaulting on loans. But suppose that the postcode correlates with racial origin. In that case, if the bank acted on the basis of this prediction and denied loans to the people in that postcode, the practice would harm people from a certain racial origin. The organisation could also intentionally use proxies to discriminate on the basis of racial origin.*

4 How to prevent against discrimination in AI/ADM tools

There are some methods which may help tackle or to minimize the risk of discrimination while using AI/ADM tools.

Examples are **human–centered solutions** embedded in public procurement procedures and **algorithmic impact assessments**.

The World Economic Forum Guidelines for AI procurement put forward the following **10 principles to prevent bias or harm via AI/ADM.**

"Trust-worthy" AI/ADM as defined by the European Commission High-Level Expert Group on AI includes the following principles: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountability. While planning the procurement these principles should also be taken into account.

1. Use procurement processes that focus not on prescribing a specific solution but rather on outlining problems and opportunities and allow room for iteration.
2. Define the public benefit of using AI while assessing risks.
3. Align your procurement with relevant existing governmental strategies and contribute to their further improvement.
4. Incorporate potentially relevant legislation and codes of practice in your RFP.
5. Articulate the technical and administrative feasibility of accessing relevant data
6. Highlight the technical and ethical limitations of intended uses of data to avoid issues such as historical data bias.
7. Work with a diverse, multidisciplinary team.
8. Focus throughout the procurement process on mechanisms of algorithmic accountability and of transparency norms.
9. Implement a process for the continued engagement of the AI provider with the acquiring entity for knowledge transfer and long-term risk assessment.
10. Create the conditions for a level and fair playing field among AI solution providers.

For example, to secure the transparency of the tool one of the requirements described in the contract notice could include an open-source solution, which means that external experts have the possibility to review software code to reveal potential risks of corruption.

A **practical example** of an introduction to Algorithmic Impact Assessment can be found in the Algorithm Charter For Aotearoa New Zealand risk matrix. The **Key elements of a public agency algorithmic impact assessment (AIA)** as described in AI Now Institute Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability and can be seen below.

This matrix should be used before applying the actual AIA questionnaire which helps to identify risks in more details and is useful to describe concrete discriminatory impact. Participants were presented with an example from Canada. The AIA questionnaire consists of questions such as:

▪ Does the recommendation or decision made by the system include elements of discretion? ▪ -Describe what is discretionary about the decision ▪ -Is the system used by a different part of the organization than the ones who developed it? ▪ -Are the impacts resulting from the decision reversible	▪ Does the recommendation or decision made by the system include elements of discretion? ▪ Is the system used by a different part of the organization than the ones who developed it? ▪ Are the impacts resulting from the decision reversible? ▪ How long will impacts from the decision last?	▪ Will the Automated Decision System use personal information as input data? ▪ What is the highest security classification of the input data used by the system? (Select one) ▪ Who controls the data? ▪ Who collected the data used for training the system? ▪ Who collected the input data used by the system?

1. Agencies should conduct a self-assessment of existing and proposed automated decision systems, evaluating potential impacts on fairness, justice, bias, or other concerns across affected communities.

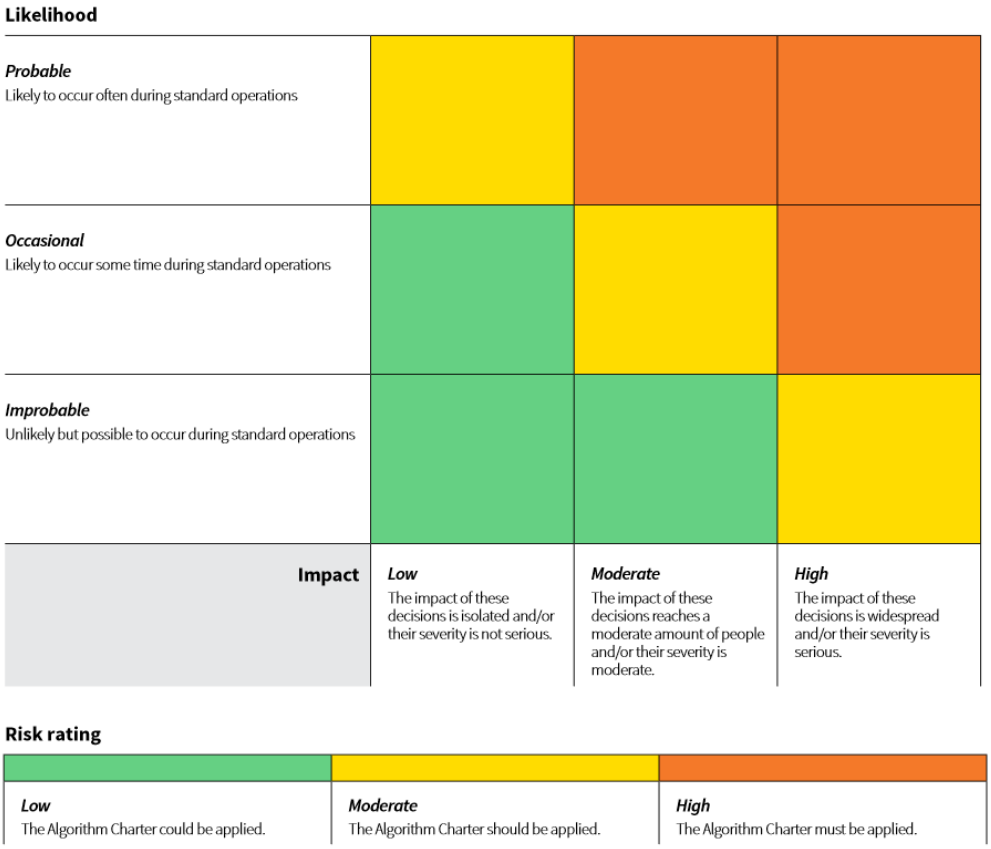
2. Agencies should develop meaningful external researcher review processes to discover, measure, or track impacts over time;

3. Agencies should provide notice to the public disclosing their definition of "automated decision system," existing and proposed systems, and any related self-assessments and researcher review processes before the system has been acquired;

4. Agencies should solicit public comments to clarify concerns and answer outstanding questions; and

5. Governments should provide enhanced due process mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased, or otherwise harmful system uses that agencies have failed to mitigate or correct.

Risk matrix



Risk rating

Low	*Moderate*	*High*
The Algorithm Charter could be applied.	The Algorithm Charter should be applied.	The Algorithm Charter must be applied.

5 Summary

Municipalities which want to prepare for wider implementation of AI/ADM solutions to prevent potential risks of discrimination should:

- Introduce policies on algorithms implementation which described the process and people responsible (ideally multi-disciplinary and diverse team).
- Introduce Algorithmic Impact Assessments.
- Introduce transparency clauses in contracts with companies delivering the software and open access to the source code, if not among the wide public at least among external experts.
- Issue guidelines explaining the operation of algorithms to those who are directly impacted.
- Elaborate on the system of reviewing AI/ADM solutions, again including the multi-disciplinary and diverse team).
- Engage citizens and experts in planning procurement and implementation of AI/ADM which will help to identify potential risks of discrimination.
- Involve knowledge and competencies building schemes for public officials and other municipality employees involved directly or indirectly in using AI/ADM solutions.