# 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

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 Govemment-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.

This is part 2 of the policy brief prepared for the online course Artificial intelligence and anti-discrimination for local authorities by the Intercultural Cities programme. The full policy brief published online can be found on the Intercultural Cities webpage.

## 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

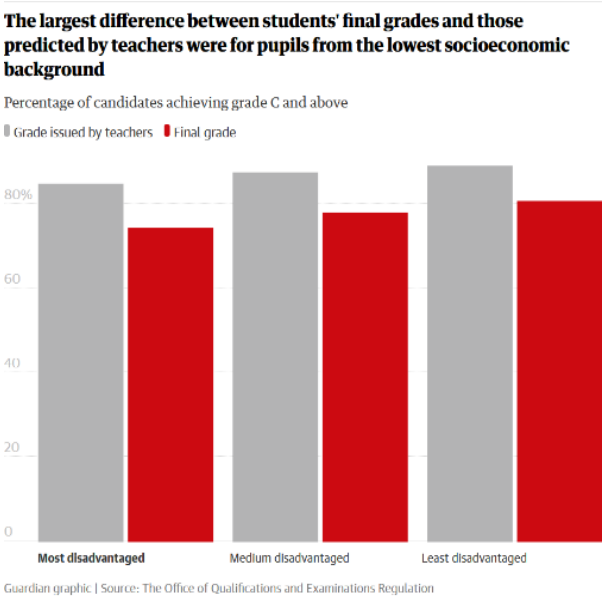
Ofqual figures show 39.1% of 700,000 teacher assessments were lowered by at least one grade

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.



The largest difference between students' final grades and those predicted by teachers were for pupils from the lowest socioeconomic background

Percentage of candidates achieving grade C and above

Grade issued by teachers | Final grade

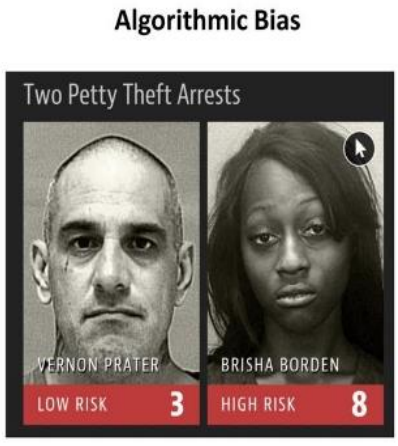
80%
60
40
20
0

Most disadvantaged | Medium disadvantaged | Least disadvantaged

Guardian graphic | Source: The Office of Qualifications and Examinations Regulation

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.*