ELA
EUROPEAN LABOUR AUTHORITY

# ARTIFICIAL INTELLIGENCE AND ALGORITHMS IN RISK ASSESSMENT

## ADDRESSING BIAS, DISCRIMINATION AND OTHER ETHICAL ISSUES

## HANDBOOK SUMMARY

**The ELA handbook, based on the online training session of 26 May 2023, in the context of the ELA programme on AI and algorithms in risk assessment, addresses the use of automation, rule-based models, and AI systems. It aims to enhance understanding about biases and legal/ethical issues associated with algorithm development and utilisation, providing insights into the legislative framework and methods to mitigate biases and discrimination.**

## Algorithms, automation and AI key concepts

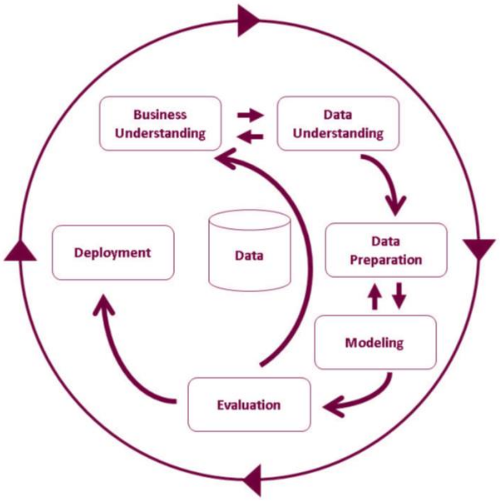
- **Automation**: The use of technology to produce and deliver goods and services with minimal human intervention, improving efficiency and reliability.
- **Algorithms**: Step-by-step procedures used to solve problems or perform computations.
- **Artificial Intelligence (AI)**: A branch of computer science that automates intelligent behaviour, often including machine learning and deep learning.
- **Machine Learning**: The process in which algorithms learn patterns from data and propose new solutions based on that learning.
- **Deep Learning**: A specialised form of machine learning that uses neural networks.
- **Data Science**: The field that combines expertise in automation, algorithms, machine learning, and deep learning to analyse and interpret data.

**Algorithmic discrimination** can occur when AI and algorithms exhibit biases or discriminatory outcomes. Decision-making algorithms rely on past data, which can amplify existing forms of discrimination found in human decision-making. If the training data used for algorithms is biased, the output of the system is likely to exhibit biases as well.

**Biased data collection, curation, labelling, modelling, problem framing, and interpretations of algorithmic output** can also contribute to algorithmic discrimination. It is important to recognise that biases can arise from both humans and machines within the socio-technical systems where algorithms operate.

## The CRISP-DM model

**Figure 1 - Process phases of the CRISP-DM model**



The **CRISP-DM model** is a standardised process for the development of AI/ML applications, in which an understanding of the data is created to solve the problem and to make a prediction for the machine learning process.

## Machine learning bias

**Machine learning bias refers to unintended or harmful deviations in algorithmic results caused by biases in data or model development**. Biases can enter through data collection or design choices.

Addressing biases is crucial to prevent discrimination and unfair outcomes. Strategies like diverse data collection and ongoing monitoring can mitigate biases and promote fairness in machine learning systems. **Several types of bias**, including:

- **Social Bias**: Existing biases in human society are reflected in the available data, leading to replication and reinforcement of bias within the model.

- **Measurement Bias**: Imperfect features and labels are used as proxies for the real variables of interest, resulting in incorrect measurements.
- **Representation Bias**: The input data does not accurately represent the real world, causing systematic errors in model predictions.
- **Label Bias**: The labeled data systematically deviates from the underlying truth, introducing bias into the model.
- **Algorithmic Bias**: Inappropriate technical considerations during modeling result in systemic deviations in the model's outcomes.
- **Evaluation Bias**: Non-representative testing populations or inappropriate performance metrics are used to evaluate the model, leading to biased evaluations.
- **Deployment Bias**: The model is used, interpreted, and deployed in a different context than it was originally built for, introducing bias.
- **Feedback Bias**: The model's outcome influences the training data, creating a feedback loop that reinforces even small biases.

## The legal framework

**EU legislation on non-discrimination** includes **Article 19 of the TFEU**, which allows the Council to adopt legislation to combat discrimination based on personal characteristics. **Article 157** ensures gender equality in work and pay, while **Article 21 of the Charter of Fundamental Rights** prohibits discrimination based on protected grounds. **Four Directives** set minimum requirements for anti-discrimination measures.

**Article 14** of the **ECHR** prohibits discrimination in the enjoyment of rights based on protected grounds.

**AI sectoral regulations**, including: **EU AI Act (proposed); AI Liability Directive (proposed); Council of Europe Framework Convention on AI; GDPR.**

### When is bias unlawful discrimination?

Algorithmic bias qualifies as unlawful discrimination if it harms a protected group, falls within the scope of anti-discrimination law, and results in differential treatment or disproportionate disadvantage.

**Direct discrimination** occurs when someone is treated unfavourably based on a protected ground. **Indirect discrimination** refers to seemingly neutral practices that disadvantage protected groups. There is in principle no justification for direct discrimination, save for certain exceptions, such as genuine and determining occupational requirements. For indirect discrimination, The **CJEU** developed a **proportionality test** to determine if a disproportionate disadvantage can be justified by a measure that serves a legitimate aim, is suitable to achieve that aim, and is necessary in the sense that no less intrusive measure could have been used for the same purpose.

## Mitigation framework

AI and ML can cause legal, ethical, and fairness-related harms by unfairly allocating opportunities, resources, or information, providing unequal service quality, reinforcing stereotypes, under- or overrepresenting groups.

**Seven key methods** that can be implemented without the need for a data scientist are:

- ✓ **Diversity in teams** can help mitigate measurement, representation, and deployment biases.
- ✓ **Collaborating with domain experts** on project objectives can address emerging measurement bias.
- ✓ **Discuss social and technical consequences** helps prevent deployment bias.
- ✓ **Data plotting** can reveal spikes or outliers that may distort empirical conclusions.
- ✓ **Rapid prototyping** enables the identification of unintended biases of different types.
- ✓ **Monitoring plan** allows for ongoing evaluation and detection of biases.
- ✓ **Human supervision in deployment** enhances objectivity and helps mitigate potential deployment and feedback biases.

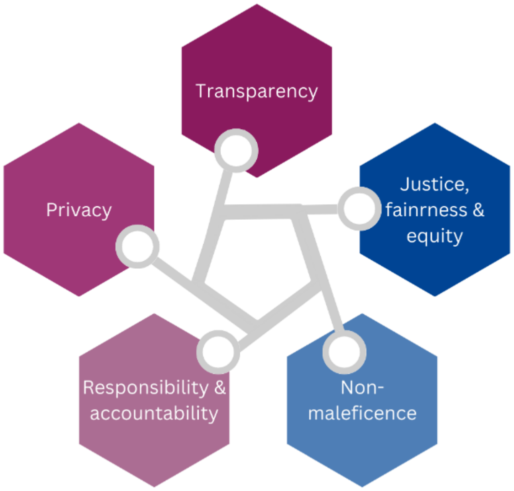
## Key ethical requirements

### Ethics Guidelines for Trustworthy AI

These guidelines have been developed by the European Commission's **High-Level Expert Group on Artificial Intelligence**, in 2019. These guidelines aim to promote trustworthy and human-centric AI by focusing on three components: **legality, ethics, and robustness**. The guidelines emphasise four principles for trustworthy AI: **respect for human autonomy, prevention of harm, fairness** and **explicability**.

### The five converging ethics principles

**Figure 2 - The five converging ethic principles**


Transparency
Privacy
Justice, fairnness & equity
Responsibility & accountability
Non-maleficence