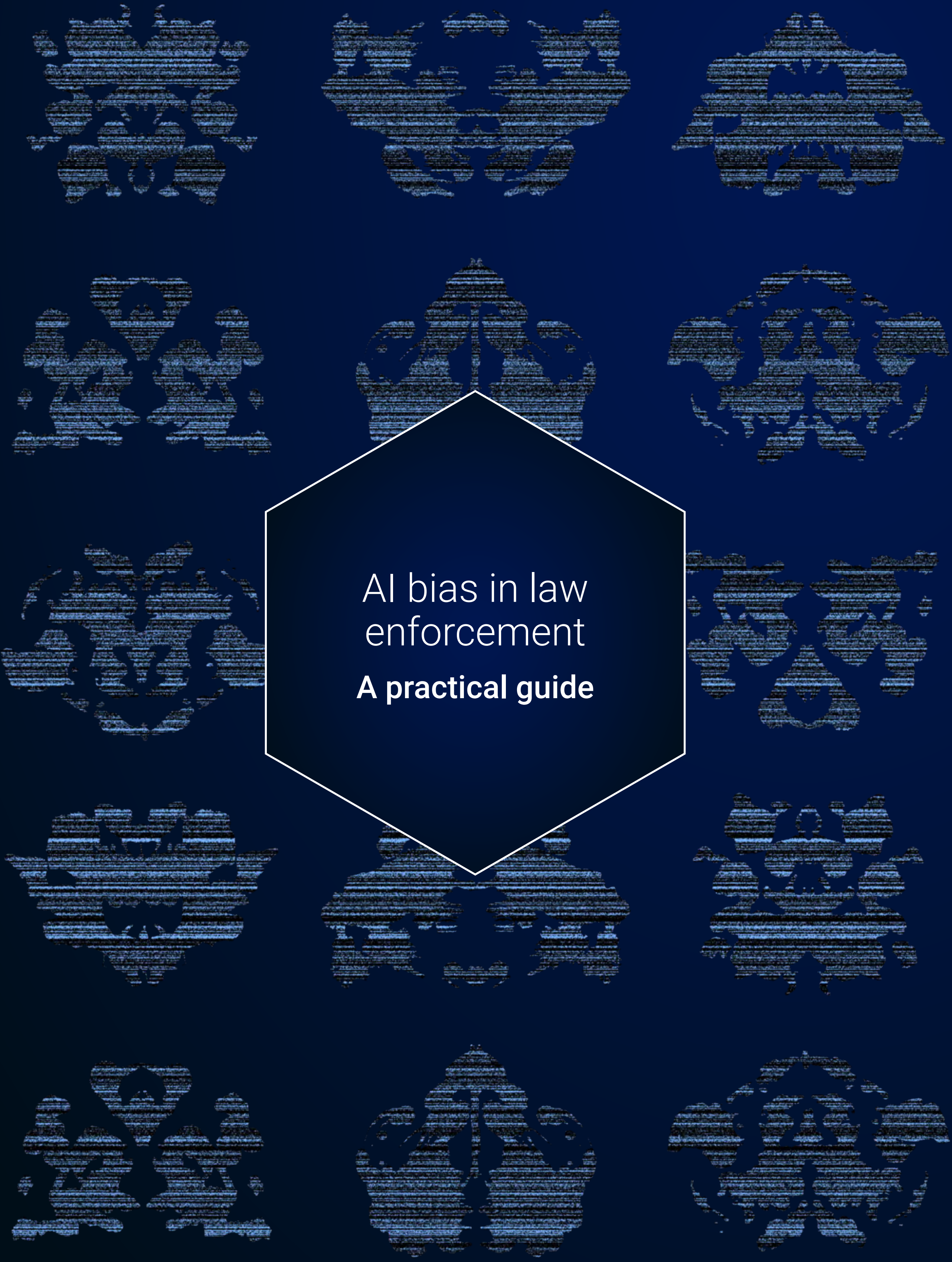
EUROPOL



AI bias in law enforcement

A practical guide

An Observatory Report from the Europol Innovation Lab

AI BIAS IN LAW ENFORCEMENT. A PRACTICAL GUIDE
An Observatory Report from the Europol Innovation Lab

PDF | ISBN 978-92-95236-94-3 | ISSN 2600-5182 | DOI: 10.2813/8081090 | QL-01-25-000-EN-N

Neither the European Union Agency for Law Enforcement Cooperation nor any person acting on behalf of the agency is responsible for the use that might be made of the following information.

Luxembourg: Publications Office of the European Union, 2025

While best efforts have been made to trace and acknowledge all copyright holders, Europol would like to apologise should there have been any errors or omissions. Please do contact us if you possess any further information relating to the images published or their rights holder.

Cite this publication: Europol (2025), AI bias in law enforcement - A practical guide, Europol Innovation Lab observatory report, Publications Office of the European Union, Luxembourg.

This publication and more information on Europol are available on the Internet.

Contents

AI Glossary

AI SYSTEM: a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

AI-BIAS: systematic errors or prejudices in data, algorithms or outcomes of an AI system that unfairly favour or disadvantage certain groups or individuals.

AI-FAIRNESS: the principle of ensuring that AI systems make decisions and predictions that are equitable, impartial, and do not result in unjustified discrimination against any group or individual.

AGGREGATE LEVEL OF ANALYSIS: involves examining crime data collectively across specific geographical areas or time periods to identify patterns and trends. This approach focuses on predicting crime hotspots or times of increased risk, rather than targeting specific individuals or incidents.

ACCURACY: a metric that measures how often a machine learning model correctly predicts the outcome. It is equal to (TP + TN)/(TP + TN + FP + FN).

BASE RATE: the proportion of actual failures, (FP + FN)/(TP+TN+FP+FN) or the proportion of actual success, which is (TP + TN)/(TP+TN+FP+FN).

BALANCE FOR POSITIVE CLASS: a classifier meets this definition if the average predicted probability score for individuals in the positive class is equal across all observed demographic groups.

BIAS: a tendency or inclination that results in unfair judgment or prejudice for or against a person, group or idea.

BLACK BOX: a system whose inputs, outputs and general function are known but whose contents or implementation are unknown or irrelevant.

CONFIRMATION BIAS: a cognitive bias where individuals favour information that confirms their existing beliefs or hypotheses while disregarding or downplaying evidence that contradicts them.

CLASSIFIER: a classifier in machine learning is an algorithm that automatically orders or categorises data into classes or categories.

CALIBRATION: a classifier satisfies this definition if individuals with the same predicted probability score s have the same probability of being classified in the positive class when they belong to any of the demographic groups.

CONDITIONAL STATISTICAL PARITY: ensures that an algorithm's decisions are equitable across different groups, given a set of relevant conditions, or factors. It adjusts the requirement of demographic parity by conditioning on these factors X: $P(Y=1|G=A,X)=P(Y=1|G=B,X)$.

CONFUSION MATRIX: a specific table layout that allows visualisation of the performance of an algorithm and contains four numbers: number of false positives, false negatives, true positives and true negatives (all defined in this Glossary)

CONDITIONAL USE ACCURACY EQUALITY: requires that both the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV) are equal across demographic groups.

DATA NORMALISATION: the process of transforming features or data attributes to a common scale without distorting differences in the ranges of values.

DATA CLEANING: the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset. Common data cleaning practices include: filling in missing values, removing duplicates, outlier detection and removal, and data formats standardisation.

DEMOGRAPHIC PARITY: also known as Statistical Parity, Group Fairness, Classification Parity, and Proportional Parity, requires that the probability of a positive prediction is the same across different groups, regardless of the ground truth. It is represented as: $P(Y=1\boxtimes G=A)=P(Y|G=B)$, where Y is the prediction outcome, 1 for positive prediction and 0 for negative prediction, while G represents demographic group membership.

DISCRIMINATION: less favourable treatments of individuals based on protected characteristics

DISPARATE IMPACT: ratio in the probability of positive outcome between minority and non-minority group. Mathematically, $P(Y=1\boxtimes G=A)/P(Y=1|G=B$, where Y is the prediction outcome, 1 for positive prediction and 0 for negative prediction, while G represents demographic group membership.

EXPLAINABILITY: the extent to which the internal workings of an AI system can be described in human terms, enabling users to understand how specific decisions or outputs are generated.

EXPLICIT BIAS: intentional, consciously articulated beliefs that result in discriminatory attitudes and behaviours toward others.

EQUAL OPPORTUNITY: requires that the True Positive Rate (TPR) is equal across demographic groups. A classifier satisfies this definition if individuals from different demographic groups have an equal chance of receiving a positive outcome, given that they belong to the positive class (i.e. they qualify for the positive outcome). Specifically, it requires that the true positive rate (the proportion of actual positive cases correctly identified by the model), TPR, is the same across all observed demographic groups. Hence, a fair equal opportunity classifier predicts positive outcomes for members of the positive class (e.g. people that are correctly identified as suspects) in both minority and non-minority groups with the same likelihood. Mathematically, equal TPRs also implies equal false negative rates (FNRs). It is also known as False negative error rate balance and True Positive Rate balance.

EQUALISED ODDS: a classifier satisfies the definition if it has equal true positive rates and equal false positive rates across different demographic groups. In other words, for a predictive model, equalised odds are achieved when the model's accuracy and error rates are consistent for all groups.

FAIRNESS: the quality of making judgments that are just, impartial and free from discrimination, ensuring equitable treatment for all.

FALSE POSITIVE (FP): a case predicted to be in the positive class when the actual outcome belongs to the negative class.

FALSE NEGATIVE (FN): a case predicted to be in the negative class when the actual outcome belongs to the positive class.

FALSE POSITIVE RATE (FPR): the fraction of negative cases incorrectly predicted to be in the positive class out of all actual negative cases, FP/FP+TN . FPR represents the probability of false alarms – falsely accepting a negative case, P(d = 1|Y = 0)).

FALSE NEGATIVE RATE (FNR): the fraction of positive cases incorrectly predicted to be in the negative class out of all actual positive cases, FN/TP+FN . FNR represents the probability of a negative result given an actually positive subject, P(d = 0|Y = 1).

FALSE DISCOVERY RATE (FDR): the fraction of negative cases incorrectly predicted to be in the positive class out of all predicted positive cases, FP/TP+FP. FDR represents the probability of false acceptance, P(Y = 0|d = 1).

FALSE OMISSION RATE (FOR): the fraction of positive cases incorrectly predicted to be in the negative class out of all predicted negative cases, FN/TN+FN . FOR represents the probability of a positive case to be incorrectly rejected, (P(Y = 1|d = 0)).

FEEDBACK LOOPS: a feedback loop in an AI system refers to the process where the outputs of the system are fed back into it as inputs, influencing future outputs. This isn't inherently harmful or biased. It can be either positive or negative, depending on how it's applied and managed. If outputs are biased, this bias is being reinforced with the existence of feedback loops.

GROUND TRUTH: information that is known to be real or true, provided by direct observation and measurement.

HUMAN COGNITIVE BIAS: occurs when humans are processing and interpreting information, showing favouritism towards some things, people, or groups, over others.

HUMAN EVALUATION: refers to the process of having people (e.g. domain experts, annotators, or end-users) directly assess, review, or rate the performance, outputs, or behaviour of an AI system.

IMPLICIT BIAS: unintended and unconscious assumptions, often based on stereotypes

LARGE LANGUAGE MODEL (LLM): an AI system that uses deep learning on vast amounts of text data to understand and generate human-like language

MACHINE LEARNING (ML): a field of study that gives computers the ability to learn without being explicitly programmed.

MINORITY OR MINORITY GROUP: a subgroup of the population with unique social, religious, ethnic, racial, and/or other characteristics that differ from those of a majority group. The term usually refers to any group that is subjected to oppression and discrimination by those in more powerful social positions, whether or not the group is a numerical minority

NEGATIVE PREDICTIVE VALUE (NPV): the fraction of negative cases correctly predicted to be in the negative class out of all predicted negative cases, TN/TN+FN . NPV represents the probability of a subject with a negative prediction to truly belong to the negative class, $P(Y = 0|d = 0)$.

OBJECTIVE FUNCTION: in the domain of AI, an objective function serves as a critical tool for quantifying the performance of a model concerning its defined goals. It encapsulates the desired outcomes and guides the learning and adaptation processes of AI algorithms.

OVERALL ACCURACY: (TP+TN)/Total Number Of Predictions.

POSITIVE PREDICTIVE VALUE (PPV): the fraction of positive cases correctly predicted to be in the positive class out of all predicted positive cases, TP/TP+FP . PPV is often referred to as precision, and represents the probability of a subject with a positive predictive value to truly belong to the positive class, $P(Y = 1|d = 1)$.

PROBABILITY OF A POSITIVE OUTCOME: NumberOfPositivePredictions/Total NumberOfCases. In the case of AI system issuing fines for speeding, probability of a positive prediction is a number of fines issued divided by number of cars tested by the AI system.

PREDICTIVE EQUALITY: a classifier satisfies this definition if the false positive rates (FPRs) are equal across different demographic groups. Specifically, it requires that the probability of incorrectly predicting a positive outcome (a false positive) is the same for all groups. It is also known as False positive error rate balance.

PREDICTIVE PARITY: focuses on achieving equal positive predictive value (PPV) across groups. PPV is the proportion of true positive outcomes among all instances that the model predicts as positive. A classifier satisfies this definition if the probability of a correct prediction (e.g. a positive outcome) is the same for all groups. In other words, predictive parity is achieved when individuals from different groups, who receive the same predicted outcome, have an equal likelihood of actually experiencing that outcome.

PREDICTIVE POLICING: the use of analytical techniques to identify promising targets

PROTECTED ATTRIBUTES: qualities, traits or characteristics that, by law, cannot be discriminated against.

STATISTICAL PARITY DIFFERENCE (SPD): the difference in the probability of a positive outcome for each observed demographic group. Mathematically, $SPD= P(Y=1\boxtimes G=A)-P(Y|G=B)$.

TRANSPARENCY: the degree to which the operations, processes and decisions of an AI system are open and understandable to stakeholders.

TREATMENT EQUALITY (COST RATIO): a classifier achieves treatment equality if the ratio of false negatives to false positives (FNs/FPs) is the same across observed demographic categories.

TRUE NEGATIVE (TN): a case when the predicted and actual outcomes are both in the negative class.

TRUE POSITIVE (TP): a case when the predicted and actual outcomes are both in the positive class.

TRUE POSITIVE RATE (TPR): the fraction of positive cases correctly predicted to be in the positive class out of all actual positive cases, $TP/TP+FN$. TPR is often referred to as sensitivity or recall; it represents the probability of the truly positive subject to be identified as such, $P(d = 1|Y = 1)$.

TRUE NEGATIVE RATE (TNR): the fraction of negative cases correctly predicted to be in the negative class out of all actual negative cases, $TN/ FP+TN$. TNR represents the probability of a subject from the negative class to be assigned to the negative class, $P(d = 0|Y = 0)$.

WELL-CALIBRATION: if satisfied, this metric means that for any predicted probability score *s*, individuals from all observed demographic groups should not only have an equal likelihood of actually being in the positive class but that this likelihood should precisely match *s*.

Foreword

Artificial Intelligence (AI) offers remarkable opportunities for enhancing the efficiency and accuracy of law enforcement investigations, border security, criminal justice procedures and asylum processes. For example, AI can assist in analysing vast datasets, identifying patterns in criminal behaviour, predicting threats and streamlining case management. However, alongside these benefits lies the ethical challenge of AI bias, which can lead to discrimination and the erosion of of trust, thereby resulting in inaccurate policing, such as the misidentification of individuals.

This report, based on extensive research conducted by Europol's Innovation Lab, addresses the pressing issue of AI bias in law enforcement. It clarifies the concept of AI bias, explores its potential harms – including its impact on decision-making in critical areas such as criminal investigations – and explains strategies for detecting bias. Based on these findings, as well as the requirements outlined in the AI Act, the report provides targeted recommendations for preventing and mitigating bias in AI systems.

As the central law enforcement agency in the European Union, Europol plays a crucial role in supporting Member States.
This includes ensuring that the integration of AI across law enforcement agencies is not only compliant with EU legislation, such as the AI Act, but also transparent, efficient and aligned with ethical standards. By investing resources in promoting a better understanding of how to mitigate AI bias, Europol aims to enable Member States to integrate AI responsibly, thereby ensuring both fairness and public safety while respecting fundamental rights.

European Union legislation, such as the AI Act, stipulates the need for safe, transparent, and unbiased AI use. By respecting ethical standards, and being proactive in understanding, anticipating, preventing and mitigating AI bias, law enforcement agencies can be responsible users of AI. This report aims to be a guide for law enforcement in terms of using AI as a force for good, with the objective of promoting fairness and public safety.

Catherine De Bolle
Executive Director of Europol

Executive summary

This report examines the critical issue of AI bias in law enforcement, focusing on its implications for operational effectiveness, public trust and fairness. While law enforcement applications of AI technologies – such as predictive policing, automated pattern identification and advanced data analysis – offer significant benefits, they also carry inherent risks. These risks arise from biases embedded in the design, development and deployment of AI systems, which can perpetuate discrimination, reinforce societal inequalities and compromise the integrity of law enforcement activities. The report highlights the necessity of addressing these challenges to ensure responsible and fair use of AI in law enforcement.

AI bias can emerge at any stage of the system lifecycle. In the design phase, historical and representation biases often reflect unequal societal patterns, leading to skewed outcomes. During the deployment phase, misuse bias and over-reliance on AI outputs (automation bias) can lead to wrongful actions, such as unnecessary surveillance or wrongful arrests. These issues are especially prominent in predictive policing, where biased data and feedback loops can reinforce stereotypes and disproportionately target marginalised communities.

The report emphasises the complexity of defining and measuring fairness in AI systems, as notions of fairness often depend on context and specific use cases. Fairness metrics, both statistical and causal, are valuable for identifying and quantifying biases. Furthermore, AI bias mitigation methods provide mechanisms to reduce bias at different stages of AI system development. However, these efforts often require trade-offs, particularly between fairness and AI model accuracy or privacy, highlighting the nuanced decision-making involved in deploying AI responsibly.

The report calls for the consistent application of legal safeguards, such as those outlined in the AI Act, for mitigating bias and respecting fundamental rights. These measures include risk management systems, transparency requirements, and human oversight protocols. By complying with these regulations and promoting understanding of bias detection and mitigation strategies, law enforcement agencies can responsibly integrate AI technologies into their operations.

By understanding the nature and impacts of AI bias and implementing the recommendations proposed, law enforcement agencies can more safely take advantage of AI technologies. This report serves as a foundational step in guiding agencies towards achieving these goals.

Key recommendations for law enforcement

- **Documentation and transparency**
  Maintain detailed documentation of all AI lifecycle stages, from problem definition to development and deployment, including decisions made based on AI outputs. As humans are involved in and heavily influence an AI system throughout its lifecycle, clear documentation of each step of the AI lifecycle processes is necessary. This ensures traceability, accountability, and aids in identifying where biases may occur, particularly focusing on confirmation bias.

- **Holistic Evaluation Framework**
  To effectively evaluate and implement AI models, develop a comprehensive socio-technical framework that not only assesses technical accuracy but also thoroughly considers historical, social and demographic contexts. To achieve this successfully, active engagement of a diverse group of stakeholders with varied expertise, such as engineers, data scientists, lawyers, sociologists, etc., is recommended. Their involvement ensures a broader range of perspectives and helps identify potential pitfalls, making AI systems more equitable. This integrated approach prioritises not just technical precision but also fairness.

- **Regular Training and awareness**
  Conduct ongoing training for all law enforcement personnel involved with AI tools to deepen their understanding of AI technologies, bias implications, and the importance and meaning of fairness metrics. Such training should emphasise the value of human evaluation in reviewing AI-generated outputs, the responsible interpretation of those outputs and the potential for confirmation bias.

- **Rigorous pre-deployment testing**
  Rigorous performance, impact assessments and bias testing should be conducted for all available datasets before deployment. When working with pre-trained models, it may not always be possible to access or inspect the original training data. In such cases, rely on available documentation or metadata and test the model's outputs for indicators of bias. If you are fine-tuning a pre-trained model, carefully evaluate and test the fine-tuning dataset for biases before and during the fine-tuning process to ensure that no new biases are introduced and that existing biases are not amplified.

- **Case-by-case analysis and technical training**
  Determining what is fair can be difficult and often depends on the context in which it is evaluated. With various conceptions of fairness presenting different trade-offs based on the situation, a well-informed, case-by-case analysis is crucial for the responsible use of AI by law enforcement agencies. As a result, fairness metrics are hard to generalise and can cause problems if applied inappropriately. It is essential to have a comprehensive understanding of their implications. Therefore, regular training on

understanding different AI-biases and their relation with fairness metrics and bias mitigation methods for staff is advised.

- **Continuous bias assessment and mitigation**
  Implement regular testing and re-evaluation of AI models throughout their lifecycle to detect and mitigate biases. This includes analysing decision-making processes and the backgrounds of those influencing AI development. Multi-stakeholder engagement should be encouraged to ensure assessments are comprehensive and balanced. These include AI practitioners and analysts, legal and compliance teams, data scientists and engineers, AI and ML researchers, social scientists, ethics and fairness specialists, data protection authorities, AI oversight agencies, policy experts, civil society and advocacy groups.

- **Human-in-the-loop and human evaluation**
  Human evaluation uses human judgment to evaluate qualitative aspects of an AI system that sometimes might be difficult to capture using predefined metrics. This form of evaluation should be an integral part of AI system assessment. Fairness testing and post-processing bias mitigation techniques should be applied on both AI system output and the final decisions made by human experts who rely on those outputs. Furthermore, it is essential to clarify whether the final prediction of the AI system and human-in-the-loop is based on causality or correlation as this is of crucial importance for legal domain applications. Ultimately, it will be up to human investigators to determine how to act on the information and suggestions generated by AI, which in turn brings up questions regarding the suitability of certain actions.

- **Trade-offs between fairness and quality of the models**
  Balancing fairness and quality in AI models present significant challenges, especially in high-stakes law enforcement applications. Mitigating bias often involves excluding sensitive attributes or applying fairness constraints, which can reduce the model's predictive accuracy. Law enforcement agencies (LEAs) must carefully evaluate the context and objectives of each AI application, aligning fairness measures with operational goals to ensure both ethical and effective outcomes.

- **Contextual and statistical consistency**
  AI models are considered properly evaluated for a given purpose only when both contextual and statistical consistency are met. This means that the decision-making context or environment and statistical properties of data samples on which decisions are to be made remain consistent between evaluation and usage. It is important to be especially careful when implementing off-label AI models, i.e. models applied to a task for which they were not designed.

- **Standardisation of procedures**
  Standardise fairness metrics and mitigation strategies across the organisation to ensure consistent practices in bias assessment.

This includes adopting common frameworks for bias testing and mitigation in order to make results comparable and reliable across different units and deployments. Convert proposed actions and standards into documentation templates that guide users through workflows. These templates should help in ensuring consistent application of bias mitigation strategies, fostering transparency and accountability in AI deployments.

Introduction

This report focuses on artificial intelligence (AI) bias detection and mitigation practices that are useful for LEAs' operational work. The aim is to:

- highlight which AI biases are relevant for LEAs;
- compile a comprehensive set of bias metrics for assessment purposes and bias mitigation algorithms that are available (or potentially available) for use in LEAs;
- assist LEAs in AI bias prevention and mitigation through a set of recommendations;
- and to highlight potential limitations specific to LEAs, such as legislative constraints.

Background

With growing availability of data and computational power, AI systems are becoming increasingly accurate and, consequently, increasingly relevant and present in the public sector. Consequently, AI has emerged as a crucial enabling technology in public services, with its usage consistently rising[1].

In law enforcement, AI is particularly valuable as there are numerous applications for different use cases. Law enforcement applications benefitting from AI span several domains, including data analytics, biometric identification, natural language processing (NLP) and computer vision[2].

However, inherent characteristics of AI systems designed for use in LEAs can produce biased outcomes and lead to discrimination, particularly concerning age, ethnicity, sex or disabilities. For example, in the Netherlands, the use of the System Risk Indication (SyRI) profiling, an AI-based system identifying social benefits fraud, was banned in 2020[3]. SyRI used data from multiple databases, which included personal and sensitive information, to generate a risk-prediction score for individuals, indicating their potential likelihood of committing a crime or offense. The algorithm's risk model incorporated a number of unidentified risk indicators, such as those associated with education, taxes, health insurance and residence status. Based on these indicators, further investigations could be triggered[4]. This system was judged to be detrimental to human rights and considered as unlawful by a court of law in 2020[5], which acknowledged that the extensive data processing and use of risk profiles were problematic, and that 'links were created based on bias, such as a lower socio-economic status or an immigration background'. Other examples of predictive policing bias are listed in Potential harm from AI bias in LE.

EU law enforcement and criminal justice authorities may increasingly utilise AI systems to profile areas, forecast crime occurrences and evaluate the likelihood of future offences. These predictive measures and risk evaluations, targeting individuals, groups and specific locations, have the potential to shape or lead to excessive policing and biased criminal justice outcomes. This could manifest in heightened surveillance, stop-and-search procedures, fines, unwarranted interrogations and other policing actions. Consequently, such practices may result in wrongful arrests, detentions and prosecutions.

Under the EU Law Enforcement Directive (LED)[6], it is required that these systems be piloted and implemented with the necessary safeguards and subjected to impact assessments to ensure compliance with the principles of data protection. This includes adhering to the principle of lawfulness, fairness and transparency, ensuring that personal data processing is conducted in a lawful and fair manner and is transparent to individuals.

The EU has actively addressed AI and bias, especially in sensitive areas like law enforcement, through initiatives such as the High-Level Expert Group on Artificial Intelligence (AI HLEG) and its 'Ethics Guidelines on Trustworthy AI' (2019). Key legislation includes the AI Act (Regulation (EU) 2024/1689)[7], effective August 2024, alongside the Digital Services Act (DSA)[8] and the General Data Protection Regulation (GDPR)[9]. These laws target high-risk domains, including law enforcement, migration, and border control and mandate measures to detect, prevent, and mitigate AI bias (Article 10(5)). High-risk systems must also address feedback loops (Article 15(4)) and document steps taken to correct bias, ensuring transparency and accountability (Article 10(2)(g)).

Scope & purpose

While there is a surge in the usage of AI in other domains, LEAs are facing many challenges in their operational work. These include the rapidly increasing volumes of data LEAs need to process, but also the need to be able to keep pace with a steadily evolving criminal landscape that sees criminals actively exploiting emerging technologies. It being a very sensitive topic with regards to fundamental rights, the usage of AI by LEAs' has to be well regulated. However, despite LEAs' AI systems being classified as high-risk[10] and public concern regarding LEAs' usage of AI that results in numerous publications from human rights advocates, there is a lack of a law enforcement-centric perspective on the problem.

This report aims to provide this perspective in order to assist LEAs in using AI more safely by analysing existing AI biases and identifying potentially harmful ones. It reviews current frameworks and procedures for detecting, preventing and mitigating bias, and proposes measures to help LEAs build public trust in their use of AI tools.

Important notice: Although the measures recommended in this report may assist LEAs in ensuring compliance with their legal obligations under the applicable legal framework, this report is not intended to provide any comprehensive guidance to LEAs regarding such obligations nor specifically regarding their compliance with the AI Act.

Non-AI bias and law enforcement

Bias, in the sense of human cognitive bias, occurs when humans are processing and interpreting information[11], showing favouritism towards some things, people, or groups over others[12]. Human cognitive bias can manifest itself openly as explicit or conscious bias. In contrast, implicit or unconscious bias often operates beyond our conscious awareness. Even individuals who strive to be fair may unintentionally apply these biases. As defined by NIST, 'implicit bias is an unconscious belief, attitude, feeling, association, or stereotype that can affect the way in which humans process information, make decisions, and take actions'[13]. The term bias is often used to refer to unfair disparities based on demographic characteristics, such as race, age, gender, ethnicity, etc. in a decision-making process that are objectionable for societal reasons[14]. As defined in 'Encyclopedia of Critical Psychology', this report refers to a minority or minority group as 'a subgroup of the population with unique social, religious, ethnic, racial, and/or other characteristics that differ from those of a majority group. The term usually refers to any group that is subjected to oppression and discrimination by those in more powerful social positions, whether or not the group is a numerical minority'[15].

Discrimination or less favourable treatments of individuals based on protected characteristics occurs even without usage of AI systems and originates in the unconscious human cognitive bias. As stated in a recent report by the EU Fundamental Rights Agency (FRA) on bias in algorithms[16], around 14 % of the general population experienced a police stop during the past year and 27 % experienced one in the past five years. This information is obtained from the FRA's 2019 Fundamental Rights Survey, which is based on around 35 000 interviews across the EU, North Macedonia and the United Kingdom. Police stops more often concern men, young people, people from ethnic minorities, people from different religious groups, i.e. Muslims, and members of the LGBTIQ+ community. For example, out of the people who consider themselves to be part of an ethnic minority, 22 % in the EU-27 were stopped by the police in the 12 months before the survey, as opposed to 13 % of people who do not consider themselves to be part of an ethnic minority. Furthermore, it is shown that judges' decisions are influenced by cognitive and societal biases, e.g. their own personal characteristics, material irrelevant to the case, stereotypes about social groups, gender, race, etc.[17].

There are several ways to define fairness, and thus many different ways to measure and reduce unfair bias. According to White House Office of Science and Technology Policy publication on implicit bias[18], two methods are used to assess implicit bias.

The Implicit Association Test (IAT) is commonly used to measure implicit bias in individuals. The IAT measures the strength of associations between concepts (e.g. skin colour, age, or sexual orientation) and evaluations (e.g. good or bad) or characteristics (e.g. athletic, smart or clumsy). The IAT is based on the observation that people place two words in the same category more quickly if the words are already associated in the brain.

For example, the rate at which a person can link the words 'black' or 'white' with 'good' or 'bad' indicates their implicit bias. In this way, the IAT measures attitudes and beliefs that people may be unwilling or unable to report.

The second method of measuring implicit bias uses randomised experiments on populations of people. In these studies, each participant is asked to evaluate an item, which might be a résumé, a photograph or a job performance description. One characteristic of that item is varied randomly.

For instance, in one type of experiment all evaluators see the same résumé, which has been randomly assigned a woman's or a man's name. If the evaluators who have seen the résumé with the man's name are more likely to hire the candidate, but they believe they have no a priori preference for a man or woman, then this is evidence that, on average, this group of evaluators is expressing implicit bias. Some stereotypes are fictional, whereas some are real generalities about a demographic group, but either way, can lead to flawed assessments of individuals.

For example, when evaluators are asked to estimate heights of subjects standing in a doorway, the evaluators will typically underestimate the heights of the women and overestimate the heights of the men[19]. In this case, the bias is based on a true generalisation – men are, on average, taller than women are – but applying the bias that is derived from the generalisation to assessments of individuals leads to erroneous estimates about them.

Several well-studied implicit biases are particularly relevant to law enforcement because they link social groups with traits related to crime and violence. General trait and behavioural stereotypes are linking certain socio-economical groups with aggression and certain behaviour of members of those groups is qualified as more aggressive than the identical behaviour of members of non-minority groups[20]. Additionally, specific race-crime stereotypes link faces of certain races with crime and with weapons[21].

As highlighted in the White House Office of Science and Technology Policy publication on implicit bias [22], the incidence of implicit bias has not changed over several decades, demonstrating the persistence of such bias across time and generations. There are several intervention strategies used to mitigate implicit racial bias. Namely, increasing exposure to individuals who challenge stereotypes associated with their group and actively rejecting stereotypical associations while affirming counter-stereotypical ones[23 24]. Additionally, openly discussing implicit bias within organisations or communities has also been shown to lessen its impact on behaviours[25 26].

In conclusion, implicit or unconscious bias is common and difficult to mitigate, as it is deeply ingrained in us, while the factors that may have influenced our thinking throughout our lives are difficult to identify.

AI & law enforcement: use cases and legislation

There are many definitions of AI but for the purpose of this document, we will use the definition created by the High-Level Expert Group on Artificial Intelligence (AI HLEG)[27], i.e. 'As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimisation), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)' (Figure 1). As defined in the AI Act, Article 3(1), 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments[28].

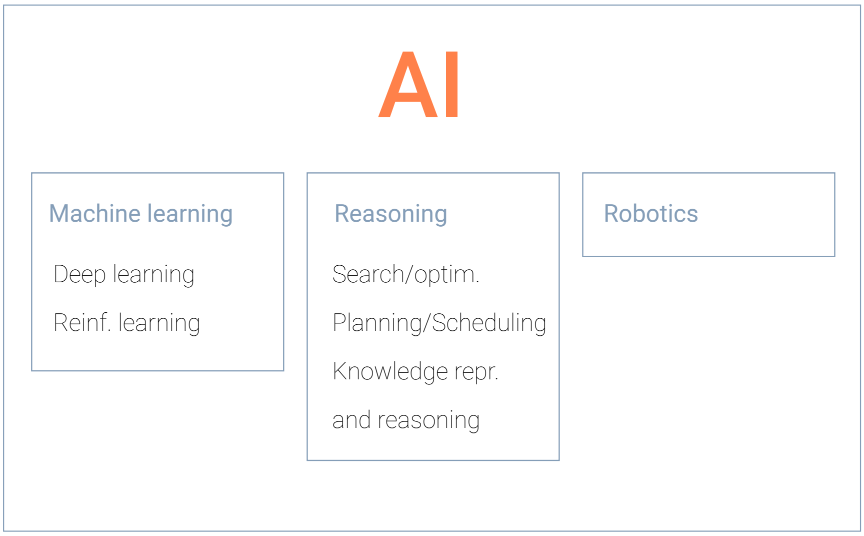


Figure 1 AI's sub-disciplines and their relationship. Source: AI HLEG. (2018). A definition of AI: Main capabilities and scientific disciplines.

One of the definitions of machine learning (ML) is that it refers more specifically to the 'field of study that gives computers the ability to learn without being explicitly programmed'[29]. In this report, the term AI systems refers to ML-based AI systems.

AI use cases

AI offers valuable tools for enhancing various aspects of law enforcement operations. For instance, predictive policing can help officers assess potential criminal activity and automated surveillance – such as using cameras in front of a suspect's location – can monitor who enters and exits, providing crucial information on suspect movements. Additionally, AI assists in analysing large datasets from confiscated devices, police reports and cold cases, allowing officers to uncover case-relevant information faster and more accurately.

AI also facilitates more efficient reporting and documentation, streamlining processes like automated report writing using speech-to-text technology, which helps officers document witness statements or perform wiretapping quickly[30]. In operations, AI technologies like facial recognition can support investigations by identifying suspects or witnesses more accurately.

By addressing repetitive desk tasks, AI enables officers to focus more on fieldwork, improving their connection with the community and enhancing job satisfaction. This improvement can also lead to a more efficient and effective police force, addressing both operational needs and community engagement.

Moreover, AI can expand the accessibility of police services. For example, crime reporting can be made more inclusive by offering multiple methods, such as in-person, phone and online options.

Legislation

Various legal frameworks govern the development and use of AI by LEAs, particularly in AI-driven crime analytics. At the EU level, these include the right to privacy and personal data protection under the Charter of Fundamental Rights of the EU[31], the General Data Protection Regulation (GDPR), the AI Act and the Law Enforcement Directive LED[32], which addresses the protection of natural persons concerning the processing of personal data for law enforcement purposes, including the prevention, investigation, detection, and prosecution of criminal offences or the execution of criminal penalties. Furthermore, these frameworks consider the principles of fairness, accountability, and transparency in AI systems, along with requirements for data security, accuracy, and non-discrimination.

Additionally, the principle of purpose limitation and data minimisation necessitates that data collected for one purpose is not used for incompatible purposes and that only the minimum amount of data necessary is processed. AI systems must also comply with the principle of accuracy, ensuring that the data used is accurate and up to date.

Automated processing and profiling must adhere to the principle of safeguards against automated individual decision-making, ensuring that such practices do not lead to significant adverse effects on individuals without appropriate safeguards. Moreover, the LED mandates that a data protection impact assessment must be conducted when processing is likely to result in high risks to individuals' rights and freedoms.

The LED specifically addresses the handling of personal data for the purposes of preventing, investigating, detecting, or prosecuting criminal activities, as well as enforcing criminal penalties, requiring that these operations be carried out with respect for these principles to protect individual rights and uphold legal standards

The application of AI in law enforcement is contentious, particularly when used in areas such as predictive policing and facial recognition, which enable automatic identification or authentication of individuals, and in criminal proceedings to assess the risk of recidivism[33]. The AI Act addresses the deployment of AI in law enforcement through several measures. It generally prohibits the use of real-time remote biometric identification systems in publicly accessible spaces for law enforcement purposes, with certain exceptions allowed under strict conditions and only when such use is strictly necessary (Article 5(1)(h) and Article 5(2) of the AI Act). These exceptions include for example targeted searches for specific victims of abduction, trafficking in human beings or sexual exploitation, as well as searches for missing persons, the prevention of a specific, substantial and imminent threat to life or physical safety of natural persons or a genuine and present or genuine and foreseeable threat of a terrorist attack. Such uses are permitted only under necessary and proportionate safeguards and conditions, including appropriate judicial oversight and the development of a fundamental rights impact assessment, among others. Additionally, the AI Act designates other AI applications in law enforcement as high-risk, which also entails compliance to very strict requirements, recognising the significant threats the use of AI in law enforcement may pose to fundamental rights (Annex III to the AI Act)[34].

Therefore, AI software applications intended for law enforcement must meet specific requirements before they can be marketed or utilised within the EU, which include, but are not limited to:

- **Article 9** (Risk management system) of the AI Act requires a continuous risk management system for high-risk AI systems, focusing on identifying, analysing, evaluating and mitigating risks to health, safety or fundamental rights throughout the system's lifecycle. It emphasises the need for thorough testing and appropriate risk management measures, particularly considering the impact on vulnerable groups, including minors;

- **Article 10** (Data and data governance) of the AI Act mandates that high-risk AI systems using model training techniques must utilise training, validation, and testing datasets that meet specified quality criteria. They should have appropriate statistical properties and account for the specific geographical, contextual, behavioural or functional factors relevant to the setting in which the AI system will be used. These datasets must adhere to data governance and management practices suitable for the intended purpose, addressing design choices, data collection origins, preparation processes, and bias assessment. Special measures must be taken to detect, prevent, and mitigate biases that are likely to have negative impact on fundamental rights or lead to discrimination, and the use of special categories of personal data is permitted only when necessary for bias detection and correction, with strict safeguards in place;
- **Article 11** (Documentation) of the AI Act requires that detailed technical documentation for high-risk AI systems be prepared before the system is marketed or deployed, kept up to date, and include sufficient information to demonstrate compliance with regulatory requirements;
- **Article 13** (Transparency and provision of information to deployers) of the AI Act mandates that high-risk AI systems must be designed for sufficient transparency, allowing deployers to interpret and use the system's outputs appropriately, and must be accompanied by clear, complete, and accessible instructions for use, detailing, among other things, the provider's information, system capabilities, limitations and human oversight measures;
- **Article 14** (Human oversight) of the AI Act specifies that high-risk AI systems must be designed with appropriate human oversight tools, enabling natural persons to effectively monitor and intervene to prevent or minimise risks to health, safety or fundamental rights, ensuring the oversight measures are proportionate to the system's risks, autonomy and context of use. The oversight measures must enable the humans to properly understand the limitations of the high-risk system and remain aware of automation bias;
- **Article 15** (Accuracy, robustness and cybersecurity) of the AI Act mandates that high-risk AI systems must be designed and developed to maintain appropriate levels of accuracy, robustness and cybersecurity throughout their lifecycle, with these levels and relevant metrics declared in the accompanying instructions. Development of high-risk AI systems based on learning models must include measures to eliminate or reduce as far as possible the risk of biased outputs influencing inputs within feedback loops. Additionally, systems must be resilient to errors, faults, and unauthorised tampering, including measures to prevent, detect and mitigate AI-specific vulnerabilities such as data poisoning and adversarial attacks.

The above list includes requirements under Section II of Chapter III of the AI Act on high-risk AI systems. The responsibility to comply with these requirements rests with the provider of the AI system. The same applies to the additional requirements for providers set out in Section III of the same Chapter III (e.g. registration in the EU database, conformity assessment, CE marking) and Chapter IV on transparency obligations and Chapter V on general-purpose AI models. LEAs will be responsible for such compliance as long as it is the provider and not just the deployer of the high-risk AI system and provided that it cannot be considered as a provider, for instance, after having put its name and logo on the high-risk AI system already on the market, or for having made a substantial modification to it or having changed the purpose of an AI system with the result of making it become a high-risk AI system (Article 25 AI Act). Yet, a responsible LEA will need to ensure that the provider has made available comprehensive Instructions for Use and has complied with these requirements before using the product and that the provider has also in place a quality management system (QMS) as defined in Article 17 of the AI Act.

If a LEA is just a deployer of a high-risk AI system, it still has to comply with obligations on deployers such as those under Articles 26 and 27 of the AI Act. In particular, to ensure that the input data is relevant and sufficiently representative in view of the intended purpose of the AI system, to assign human oversight, to monitor risks, report incidents, maintain logs, perform and document data protection impact assessment (DPIA), and Fundamental rights impact assessments under Article 27 where the categories of people likely to be affected by its use and the risks of harm it can cause should be identified. If a risk is identified, the provider of the AI system and relevant authorities must be informed immediately.

These measures are not exhaustive and yet they are essential to prevent AI systems in law enforcement from reinforcing biases or discriminating against certain populations, and to ensure transparency, fairness and safety. By implementing a robust internal compliance mechanism that ensures that these requirements and all applicable legal obligations are fulfilled, crucial advance towards the responsible and ethical deployment of automated law enforcement technologies, safeguarding individual rights and promoting public trust will be made.

Understanding AI bias and AI bias types

Similar to many definitions of AI, there are many definitions of AI bias[35]. Perhaps the most appropriate definition to use in the context of fundamental human rights is that bias is a 'differential treatment based on protected characteristics, such as discrimination and bias-motivated crimes'[36]. The consequences of bias and inherent unfairness have gathered significant attention, especially as AI becomes more prevalent in sensitive fields such as healthcare, hiring practices, law enforcement and criminal justice.

The assumption that data-driven decision-making increases fairness is true only to an extent. For example, while it is shown that algorithms could help reduce racial disparities in the criminal justice system[37], data-driven systems are not always fair. When considering fairness in AI system design and deployment, it is crucial to remember that these technologies, despite appearing neutral, are created by humans who bring their own contextual limitations and biases to the process[38] and these are difficult to mitigate as shown in the previous chapters. Human and societal bias, whether explicit or implicit, can be embedded in AI bias through the dataset used for training. For example, if people of a certain gender, race and age appear more in social media or academic environments, this can be reflected in datasets created via those sources and the ML algorithms trained on those datasets will possibly perform better on that type of population. Furthermore, decisions made during the AI system development phase can introduce additional biases, which can be reinforced in the deployment phase if AI system contains feedback loops. Additionally, humans perform certain actions based on AI system outputs. These actions are affected by both AI and human biases and can cause discrimination.

Although the term ‘bias’ often has a negative connotation; bias is an inherent part of every ML model. That is, every ML model is inherently biased by design as it is designed to spot patterns in the training data. This can be illustrated with an example where an ML model developed to differentiate between cars and bicycles must be biased towards identifying vehicles with four wheels and an enclosed structure as cars. This type of bias does not lead to discrimination or errors in decision making. Furthermore, in ML and statistics, the bias of an estimator is defined as the difference between this estimator’s expected value and the true value of the parameter being estimated and it is always present[39].

In this report, we focus on unwanted AI biases that are harmful and the remainder of this chapter will enlist and clarify different AI bias types as defined and used in literature. The classification of different AI bias types can be done based on its occurrence in the AI lifecycle or based on its source. Both classifications are useful. The former facilitates referencing bias types within AI bias management frameworks and guides actions to be undertaken during different phases of the AI life cycle. The latter helps in selecting appropriate fairness metrics and mitigation algorithms for bias detection.

AI bias types based on the AI lifecycle phase

AI bias can be introduced in all stages in the AI lifecycle[40]. A scheme of the AI lifecycle representing three phases is shown in Figure 2[41]. AI bias can be introduced in any of the three phases and, below, we list the types of biases according to the AI lifecycle phases in which they appear.

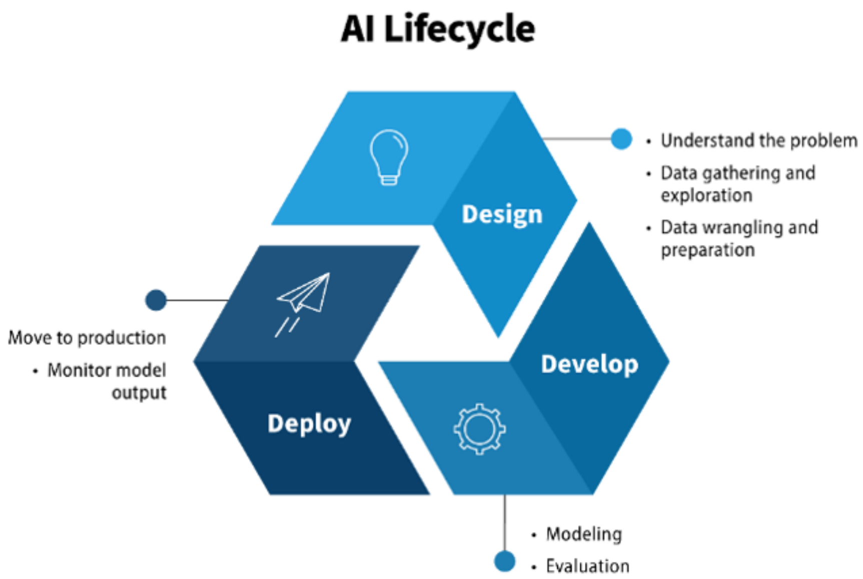


Figure 2 AI lifecycle phases: Design, Development and Deployment. Source: https://coe.gsa.gov/coe/AI guide-for-government/understanding-managing-AI lifecycle/

Design phase – problem formulation phase

PROBLEM FORMULATION BIAS

The first stage of the AI cycle is the problem formulation stage. During this stage, the objective of the AI system and how this objective will be achieved are defined. At the very first step of the AI lifecycle, a target variable that the AI system will predict is identified. The target variable might not accurately capture the complexity of the system's predictive goals, often being too simplistic or imprecise for effective analysis.

For example, take an AI system is designed to optimise the deployment of traffic patrols by predicting areas where traffic violations are most likely to occur. The problem is defined narrowly as 'minimising the number of traffic violations recorded', leading the AI to focus on areas with high levels of traffic enforcement infrastructure (e.g. cameras, sensors) rather than on areas with high accident rates or safety concerns. This could result in under-prioritising locations where accidents frequently occur but lack prior data due to insufficient monitoring.

Design phase – data acquisition

Before an ML model can be trained, data needs to be collected. During this step, a target population is defined, and its features and labels are defined and measured. In this phase, several biases occur.

HISTORICAL BIAS

Bias in AI systems often originates from the data itself, which can carry forward intricate societal and historical patterns influenced by human biases, both explicit and implicit. This creates a misalignment between the current state of the world and the ideal values or goals we aim to embed in a model. Even if a system accurately reflects the world and the data is impeccably measured and sampled, it can still perpetuate harm, such as reinforcing stereotypes related to race, ethnicity or gender. This demonstrates that a seemingly objective model can unintentionally uphold and propagate existing inequalities.

SAMPLING OR REPRESENTATION BIAS

Sampling or representation bias arises when collected data is unbalanced and does not accurately represent the population for which the AI system is supposed to be used. The algorithms might not perform well on such data. For example, if a face recognition algorithm is trained on mostly white male subjects, it might not work well on women of other ethnicities.

LABELLING BIAS

Labelling bias arises when the individuals responsible for tagging data allow their subjective views to influence the labelling process. For instance, imagine a law enforcement agency is creating a training dataset for an AI system intended to identify 'suspicious behaviour' from security camera footage. Human annotators watch hundreds of video clips and assign labels like 'suspicious' or 'not suspicious' to individuals' actions. If some annotators hold subconscious biases or stereotypes – such as perceiving individuals of certain ethnic backgrounds as inherently more suspicious – they may consistently label clips with those individuals as 'suspicious' at a higher rate than identical behaviours performed by individuals of other backgrounds. This pattern of skewed labelling, influenced by the annotators' subjective views rather than objective criteria, represents a clear case of labelling bias.

Development phase

During the development phase, several actions are conducted. Firstly, the collected data is prepared so that it can be used to train an AI model. The data undergoes normalisation and cleaning processes (see Glossary), and measurable features are identified to serve as proxies for underlying concepts that cannot be directly observed. After that, the algorithms to be employed are chosen and the AI model is trained. During this phase, several biases can arise.

MEASUREMENT OR PROXY BIAS

There are several sources of measurement bias in AI systems. One cause is the inaccurate measurement of specific features due to faulty or inconsistent measuring instruments. Additionally, measurement precision may vary across different groups within the population being studied. Another issue arises when the chosen features do not adequately represent the complex concept the AI system is designed to address. Moreover, bias can occur if the system uses non-sensitive attributes that are highly correlated to sensitive attributes like race, gender or socio-economic status, leading to biased results. For example, using a zip code as a variable can inadvertently serve as a proxy for socio-economic and ethnic demographics. It is crucial to distinguish between correlation and causation; AI algorithms identify correlations between variables and predicted outcomes but a strong correlation does not necessarily indicate a causal relationship. For example, strong correlation between a zip code and higher rate of criminal behaviour does not establish a causal link; the zip code itself does not cause criminal behaviour, it merely reflects broader systemic issues.

AGGREGATION BIAS

Aggregation bias arises during model construction when a single, uniform model is applied to data that contains distinct subgroups that require individualised consideration. As a result, the model is optimised for the predominant group and may perform poorly for other subgroups. For instance, predictive policing tools may exhibit aggregation bias if they are developed using crime data from various cities or regions without accounting for the unique social, economic and cultural contexts of each location.

LEARNING BIAS

Algorithmic bias arises when the choice of algorithms or the design of the learning process itself introduces bias into the AI model or amplifies undesirable biases in the training data. For example, a Naïve Bayes ML model assumes features to be conditionally independent and can amplify small differences in data leading to existing bias amplification.

For example, imagine an AI system is analysing social media posts for hate speech and finds the words (features) ‘threat’ and ‘violence’ in a post. The Naïve Bayes model treats these words as independent signals, multiplying their impact and concluding that the post is very likely to be hate speech. In reality, these words often appear together in similar contexts and they do not double the chance the post is hate speech related.

EVALUATION BIAS

In the final stages of model development, after selecting an algorithmic approach and training it, the model’s performance is typically assessed using predetermined evaluation metrics and test data – yet this process itself can introduce evaluation bias. Evaluation bias occurs when the chosen evaluation metrics, the test data or the interpretation of the results fail to accurately reflect how the AI system will perform in real-world conditions. Some metrics might not be suitable for the way how the AI system will be used in practice. However, the same set of evaluation metrics are used so that it can be easy to compare the performance of various ML models. Success of a given ML model when using a certain type of metrics could be misleading in considering the model to be successful for any task. For example, if a model scores high in accuracy, it can still have high false positive rate.

Additionally, evaluation bias can be caused by a non-representative benchmark test set. Similarly, as with the sampling/representation bias in the design phase, a benchmark set that is not representative for data on which the AI system would actually be used can lead to misleading evaluation of its performance.

Deployment phase

During deployment phase, an AI system may encounter unforeseen situations when facing real world scenarios. This can result in input data to differ compared to the data used to train and evaluate the model. More generally, when an AI system enters a societal context that surrounds decision-making systems, it becomes so-called 'sociotechnical system'[42] that is more complex than the technical system it was in the design phase. This may result in the following biases:

DEPLOYMENT BIAS

Deployment bias occurs when a model struggles to apply its learned patterns to new, unseen data. This often happens because the model was overly tailored or overfitted, to the specific characteristics of the training data, or due to shifts in the relationship between the target variable and real-world data during deployment. Consequently, the model's predictions become less reliable and accurate in practical applications.

MISUSE BIAS

Misuse bias occurs when an AI system designed for one purpose is used for another, a practice known as 'off-label deployment'. For example, an AI system initially developed to monitor traffic patterns and optimise traffic flow might be repurposed by law enforcement to identify and predict criminal activity based on vehicle movement. This can result in wrongful suspicion and surveillance, as the system was not intended or validated for detecting criminal behaviour.

AUTOMATION BIAS

AI systems often generate outputs that human decision-makers must interpret. Even if these systems demonstrate strong performance in controlled environments, they can produce unforeseen results when deployed[43]. This can lead to over-reliance on the AI, causing individuals to overlook their own expertise and alter their accurate decisions to align with the AI's recommendations.

AI bias types based on its source

In their publication on bias mitigation[44], NIST[45] (National Institute of Standards and Technology) classifies biases according to their sources (see Figure 3). This leads to a different categorisation of biases than the one proposed in this report. This is a more detailed categorisation as some categories listed in the previous paragraph are a subset of categories shown in Figure 3. Furthermore, when biases in different AI lifecycle phases come from same sources, usually similar fairness metrics and similar mitigation algorithms would be used to mitigate them. Hence, categorisation based on source, in combination with the categorisation based on the AI lifecycle phase in which bias occurs, could help enhance the bias detection and mitigation process.

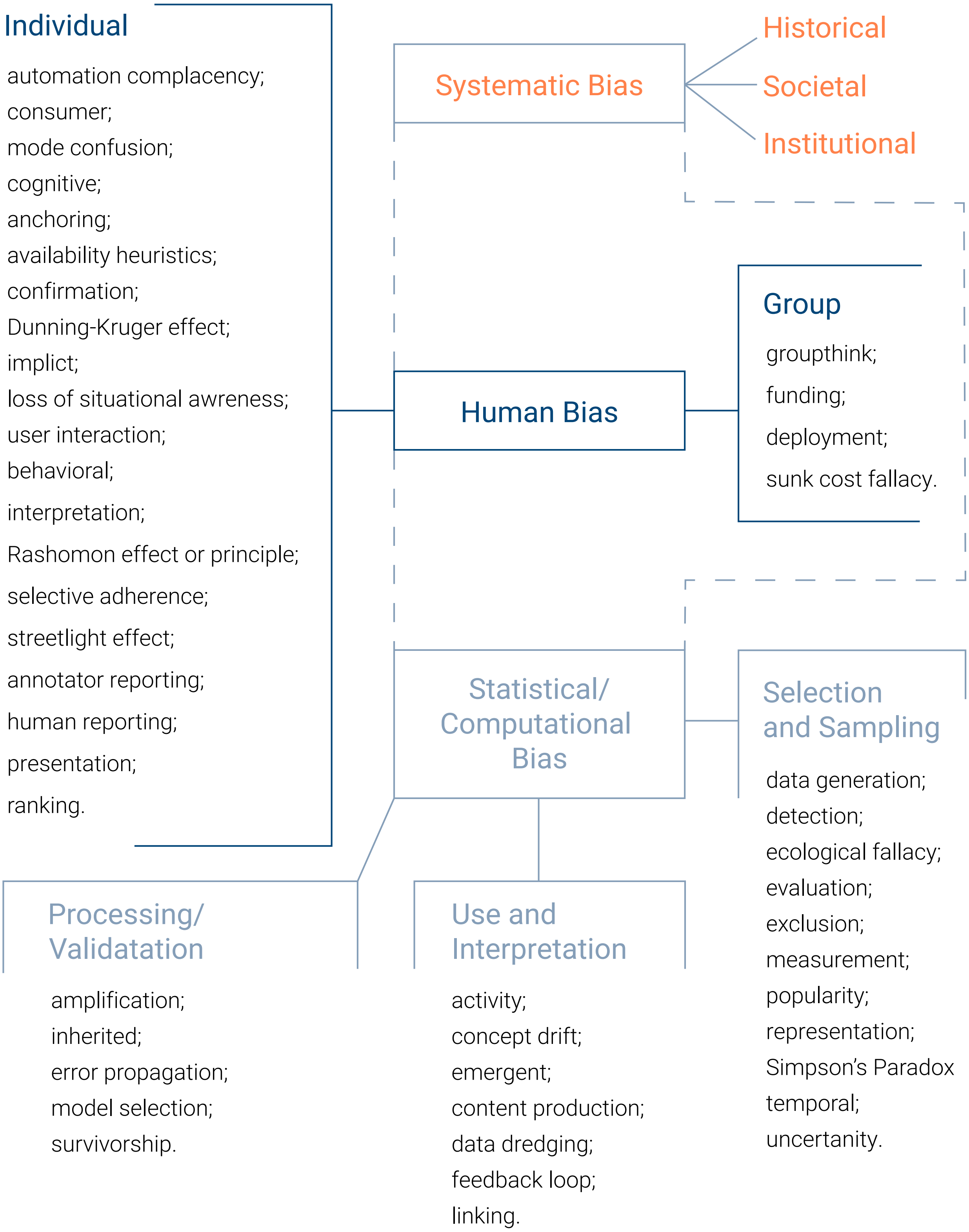


Figure 3 Categories of AI Bias. Source: NIST, 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence', 2022

SYSTEMIC BIAS

Systemic biases emerge from established procedures and norms that inadvertently favour certain social groups over others. These biases can manifest without any deliberate prejudice or discriminatory intent, often simply by adhering to prevailing rules or standards[46]. Such biases are inherent in the datasets utilised for AI development, as well as the institutional practices, norms, and procedures throughout the AI lifecycle. Systemic bias corresponds to historical bias in the previous categorisation scheme.

STATISTICAL AND COMPUTATIONAL BIAS

Statistical and computational biases stem from errors that result when the data sample is not representative of the population. These biases can occur in the absence of prejudice, partiality, or discriminatory intent[47]. In AI systems, these biases can occur in different AI lifecycles, and correspond to Sampling or representation bias of the Design phase, Learning bias, Evaluation bias and Aggregation bias of the Development phase, and the Deployment bias of the Deployment phase listed in the previous chapter.

HUMAN BIAS

Human biases are frequently subconscious and influence how individuals or groups process information, such as AI-generated outputs, to make decisions or fill gaps in knowledge[48]. These biases permeate decision-making at all levels – institutional, group, and individual – throughout the AI development and deployment stages. Although human biases are not exclusive to AI, they are deeply ingrained in human cognition. Simply becoming aware of these unconscious biases is not sufficient to ensure they are adequately controlled or mitigated.

LLM bias challenges

Large Language Models (LLMs) have significantly advanced text processing and generation, as demonstrated by tools like ChatGPT and various chatbot applications. While they exhibit impressive capabilities, they are also susceptible to absorbing and amplifying existing societal biases. These biases can stem from various sources, including the unfiltered data they are trained on, the design of the models and the objectives set during their development.

Training data for LLMs often comes from vast, unvetted internet sources, leading to the inclusion of stereotypes, misrepresentations, and biased language. Such biases can disproportionately harm marginalised groups. For instance, labelling bias arises when annotators' social and linguistic norms differ from those of the originators of that data, causing misinterpretations. A common scenario is misidentifying benign language from specific ethnic groups as offensive due to unfamiliarity with those groups' communication styles[49] [50] [51].

Moreover, LLMs trained on biased datasets can perpetuate these biases when processing new data, as seen in sentiment analysis tools that yield different results based on demographic indicators like gender or language spoken[52 53 54]. Additionally, LLMs predominantly focus on English and Indo-European languages, neglecting many global languages and perpetuating a cycle of underrepresentation. This focus not only limits the models' applicability but also reinforces linguistic inequities.

To mitigate these issues, it is essential to adopt comprehensive data curation practices, develop robust bias detection[55] and correction mechanisms[56,57,58], and expand research to include a broader range of languages and cultural contexts. This approach will help ensure that LLMs are fairer and representative in their applications. Despite these efforts, completely eliminating bias in LLMs is a formidable task, as it requires not only technical solutions but also a deep understanding of the socio-cultural contexts from which the data originates.

Potential harm from AI bias in LE

The question of harm is a very important one as it is also reflected in the fact that the AI Act is based on the potential of AI technology to cause harm. As stated in Recital 5 of the AI Act, '… AI may generate risks and cause harm to public interests and fundamental rights that are protected by Union law. Such harm might be material or immaterial, including physical, psychological, societal or economic harm'[59].

AI-related harm can be examined from various angles. From a technological standpoint, particularly in the field of ML, the focus often falls on the risks and negative impacts associated with these systems, especially due to their opaque nature, commonly known as 'black-box' algorithms. A black-box system is a system whose internal contents or implementation are not known or accessible, focusing solely on input-output behaviour[a].

The perceptions of citizens regarding AI risks are crucial for developing a comprehensive understanding of AI as a socio-technical system. Legally, the focus shifts to accountability and liability for damages that can be clearly attributed to AI. This prompts critical inquiries into the adequacy of current legal structures to manage AI-induced risks, particularly in light of the AI Act and/or any other applicable AI regulations.

Ethically, the evaluation of risks and harms involves researching the moral foundations of either employing or refraining from using various AI applications. This entails assessing whether the benefits of using AI to address challenges or solve problems outweigh the potential moral costs. Both fundamental and applied research

[a] IEEE Standard Glossary of Software Engineering Terminology.

are actively being conducted in these domains to thoroughly assess and comprehend the full spectrum of AI's impacts and the risks it poses[60].

The concept of fairness [b] is difficult to define due to its multifaceted nature. There exists a plethora of methods to define and quantify fairness, underscoring the complexity of the challenges faced in this domain[61]. Each scenario in law enforcement presents unique problems, necessitating distinct trade-offs and meticulous case-by-case analyses. The utilisation of AI systems in such contexts amplifies the need for careful consideration; however, the main element in employing such technologies is not the mere reception of data from these systems. Rather, it is the subsequent human actions and decisions, influenced by the output of AI, that are crucial. Suppose there is an AI system designed to analyse surveillance footage for the detection of unattended baggage in public areas, a common security concern in places like airports and train stations. The AI system can quickly sift through hours of video to identify objects that have been left unattended for a suspicious duration. The crucial aspect in this scenario is the human intervention that follows the AI system's notification. Once the AI system identifies potential unattended baggage, the responsibility shifts to security personnel to assess the situation. They must determine whether the identified object poses a real threat, which might involve physically checking the item, reviewing additional camera feeds to trace the item's owner, or even evacuating the area if the situation escalates. This example illustrates the essential role of human judgment in interpreting AI outputs, where the technology aids in alerting and performing an initial identification but human decision-making is vital for appropriate response and resolution.

When unwanted bias occurs in law enforcement, it can lead to discrimination. This happens when persons or groups in similar or the same situation are treated or affected differently based on certain protected grounds/attributes[62]. This relates to attributes such as race, gender, age, or sexual orientation. These attributes are called protected attributes or characteristics. However, not all forms of bias relate to protected characteristics. Additionally, bias can be related to a non-protected attribute that is highly correlated to the protected one (Measurement or proxy bias). Finally, even if the algorithm is biased with respect to protected attribute, this still can be justified with the intended usage of the algorithm. For instance, if data indicate that certain protected groups – such as specific ethnic communities or LGBTQ+ individuals – are disproportionately victims of certain types of crimes, law enforcement agencies could use this information to tailor victim support services specifically for these groups. This might involve training officers in culturally competent responses, or creating partnerships with community organisations that provide support and advocacy for these groups.

[b] The term 'fairness' as used in this report does not refer to, nor should it be interpreted in accordance with, the concept of a 'fair trial' as set forth in Article 47 of the Charter of Fundamental Rights of the European Union.

Predictive policing

In LEAs, AI bias could potentially be particularly dangerous in predictive policing systems. Predictive policing refers to the use of analytical techniques by law enforcement to make statistical predictions about potential criminal activity. It is a policing strategy that uses algorithmic surveillance to predict future crimes, criminals and victims to intervene before crimes occur. A distinction can be made between two types of predictive policing: predictive mapping and predictive identification[63].

The most commonly used type of predictive policing is predictive mapping or place-based predictive policing. This refers to advanced geospatial analyses to predict when and/or where a crime may take place at an aggregate level of analysis. An aggregate level of analysis involves examining crime data collectively across specific geographical areas or time periods to identify patterns and trends. This approach focuses on predicting crime hotspots or times of increased risk, rather than targeting specific individuals or incidents. If now place-based predictive policing reacting to higher incidence rate of crime in some neighbourhoods ends up targeting groups associated with a protected characteristic more than others, this is indirect discrimination. For instance, the use of predictive policing algorithms may result in unnecessarily increased police presence in areas mainly inhabited by certain ethnic minorities, whereby the area itself becomes a proxy for ethnic origin[64].

Predictive identification is the analysis at the individual or group level and personal data is processed; this can focus on predicting potential offenders, offenders' identities, criminal behaviour, and potential victims of crime. Recently adopted AI Act limits the use of AI systems for predictive policing[65]. In Europe, human verification, with two humans in the loop, of any AI matches is a standard practice, and that minimises the chances of wrongful arrests[66],[67]. As an illustration, a few cases of predictive policing software potentially harmful and used in Europe will be mentioned here.

For instance, a Dutch Sensing project, which ran between January 2019 and October 2020, aimed to counter crimes like shoplifting in the south-eastern city of Roermond. The Sensing Project used remote sensors in and around the city to detect the make, colour and route of cars carrying people suspected of what police call 'mobile banditry'. Sensing project identified vehicles with Eastern European licence plates in an attempt to single out Roma as suspected pickpockets and shoplifters. The model itself was specifically biased against non-Dutch nationals. Roermond police took the clearly biased step of excluding Dutch nationals from the definition of 'mobile banditry' and narrowing the focus of the Sensing Project[68]. More specifically, the target profile was biased towards designating higher risk scores for individuals with an Eastern European nationality and/or Roma ethnicity, resulting in this group being more likely to be subjected to measures, such as storage of their data in police databases[69].

In Germany and Switzerland a predictive mapping software Precobs is used[70]. PRECOBS uses algorithms and knowledge about crimes committed in the past to predict the commitment of so-called 'near repeat' crimes. By using offence data from the recent past, predictions are made for police authorities for a defined 'district' and used for operational measures and crime prevention. Predictive policing is designed to make predictions for specific offences (e.g. burglaries, vehicle offences, robberies, arsons). The forecasting software was developed by the Institute for Pattern-based Forecasting Technology IfmPt, which was taken over by Logobject Deutschland GmbH in 2021. German police forces of Karlsruhe and Stuttgart decided to stop using PRECOBS software because there was insufficient crime data to make reliable predictions[71].

Above-mentioned systems use various mathematical tools for prediction purposes, that use different data to be trained. One of the main challenges in predictive policing is utilisation of biased data. The data used to train automated algorithms is historical data, from police databases, and might not be representative for the present time. The data collected is twofold, from reported crimes and from the crimes observed or detected by the police themselves. Data distribution is affected by reluctance of certain socio-economic groups to report crimes, by increased presence of police in certain areas, by the type of crime making it more or less 'observable' or likely to be reported, etc.[72]. Furthermore, learning bias can be present where the ML model itself can amplify already existing bias. Finally, sampling bias occurs because more crime is observed where more patrols are located or when the police more closely follow a specific person. With presence of these biases, and without mitigation techniques, after certain amount of time, due to a feedback loop mechanism, where, in this case, biased AI system outputs are fed back as inputs, majority of police forces are assigned to one area where historically there were more crimes than in other ones[73]. This leads to the creation of datasets that appear to reflect higher crime rates, but which really reflect greater police attention. Additionally, ML models themselves can form feedback loops if they continue learning after deployment. Batch learning models that are periodically trained using data accumulated in batches, or online learning models that are continuously trained as the new data keep arriving, are typical examples. Such algorithm was used in the Dutch childcare benefit fraud scandal[74]. As seen in this case, bias by proxy is especially harmful and when vulnerable categories are involved, such as for example single mothers with medical conditions, it could very severely affect their lives [75]. However, complications arise when a model must account for legitimate differences in offending across demographics. For instance, men commit crime at significantly higher rates than women and are more likely to commit violent offences[76]. Age is also strongly correlated with offending; it is well known that offending tends to peak in the teenage years and then decay over time[77].

As highlighted in Alikhademi et al. work on predictive policing[78], some scholars proposed different approaches to predictive policing that did not involve predicting crime or assessing individuals, but instead involve removing motivations for crime. This approach acknowledges that certain individuals face challenges that increase their inclination to engage in violent acts and it recommends generating a public health model for identifying these people and their needs. For example, it is possible to use AI to assist police officers in understanding the context in which they work[79]. Additionally, Asaro[80] proposes an ethical framework for adopting an 'AI Ethics of Care' approach that promotes care for all stakeholders rather than models of threat. This approach requires training and guidance to educate users of AI systems on the complex socio-technological frame in which they operate.

The concerns related to predictive policing usage are valid and may lead potentially to serious infringements upon fundamental human rights. In response, the EU with the AI Act (Article 5 (1)(d))[81], prohibits the use of an AI system for predictive policing based solely on the profiling of a natural person or on assessing their personality traits and characteristics. Exception to this ban are AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity. Moreover, while the current regulatory framework appears to exclude certain area-based or event-based predictive applications, challenges may arise in the implementation process, especially as applications and models overlap.

Fairness metrics

Fairness metrics constitute the foundation of AI bias detection and mitigation methods. It is important to understand them in order to be able to use them correctly. In this chapter, we will list and explain different fairness metrics through law enforcement related examples.

AI systems influencing human lives should be viewed as dynamic entities, where their predictions and actions reshape their operational context, continually feeding new information back into the system. A crucial aim of ethical AI is to identify and mitigate disparities produced by these systems that are harmful or unjustified. This requires a comprehensive approach to evaluating and refining algorithms to ensure they do not perpetuate inequities, aiming for AI technologies that are both efficient and ethically responsible.

The European Commission has emphasised the necessity of preventing harm from AI systems, including biases, by issuing guidelines that stress transparency, accountability, explainability, and fairness[82]. These guidelines stress the importance of removing biases from data before it is used for training, stating the need for inclusion and diversity throughout the AI system's lifecycle. To achieve this objective, tools that help detect and understand biases in data and models, and manage trade-offs between bias mitigation and decision quality, are essential.

A thorough understanding of AI fairness requires exploring the broader fairness literature, which includes varied definitions and techniques. Definitions of fairness are task-dependent, and numerous types of biases can exist within data, leading to confusion due to differing terminologies and metrics. Studies like those by Verma and Rubin provide clarity[83]. Ongoing research continues to expand the list of fairness metrics used in both scientific literature and AI fairness tools.

In ML, fairness is often measured by predicted outcomes, actual outcomes, or similarity measures[84]. Researchers, including Bellamy et al., define fairness in terms of protected attributes such as race or sex, which can divide populations into privileged and disadvantaged groups, historically favouring the privileged[85]. Various definitions of fairness in ML aim to prevent AI systems from reinforcing these historical inequities. Several fairness metrics, referring to both individual and group fairness can be defined. In this chapter, we will follow the separation on statistical, similarity-based and causal metrics as defined in the work of Verma and Rubin[86].

Statistical metrics

Statistical measures of fairness are derived from the values in the classifier confusion matrix or analysis of predicted probabilities, and many of them are parity measures. A parity measure is an evaluation criterion used in the context of fairness and bias in AI and ML. It ensures that evaluation metrics, such as accuracy, error rates, or other performance indicators, are independent of protected characteristics like race, gender, age, or other attributes that could lead to discrimination. Essentially, it requires that the performance of a model be consistent across different groups defined by these protected characteristics, aiming to prevent biased outcomes and promote fairness.

To illustrate and compare these fairness metrics, we will use a concrete example of an AI system deployed to detect and issue fines for vehicles exceeding speed limits using traffic cameras and radar sensors. We will consider two types of vehicles, cars and motorcycles, each representing one demographic category.

Definitions based on predicted outcome

These definitions focus on a predicted outcome for various demographic distributions of subjects. This type of metrics can be used to estimate how AI system behaves, without knowing the ground truth.

> Demographic parity, also known as Statistical parity, Group fairness, Classification parity, Proportional parity, requires that the probability of a positive outcome is the same across all observed demographic groups, irrespective of the ground truth[87]. Following the traffic

example, Demographic parity requires that the probability of issuing a speeding ticket is the same across different vehicle types, regardless of actual behaviour.

Metrics derived from this one that appear in different fairness toolkits are:

- **Statistical Parity Difference (SPD)**[88]– This is the difference in the probability of a positive outcome for each observed demographic group. If the statistical parity difference is zero, and two demographic groups are observed, it indicates that the model treats both groups equally in terms of the positive outcome rate. However, a non-zero value suggests that there is a disparity in how the groups are treated, indicating potential bias.
  In the traffic example, if SPD equals zero, it means that the probability of a motorcycle getting a fine is the same as a car getting a fine. However, if SPD = -15 %, it means that the higher percentage of one of the vehicle types are being issued a ticket.

- **Disparate Impact**[89]– This is the ratio in the probability of positive outcome between the minority and non-minority demographic groups. A value of 1 implies both groups are treated equally, while a value smaller than 1 implies higher benefit for the privileged group. In the traffic example, if we assume motorcycles are a minority group, and the Disparate impact is equal to 2.5, it means that the motorcycles are 2.5 times more likely to get a speeding ticket.

Demographic Parity is appropriate when the goal is to ensure that all groups have equal chances of receiving positive outcomes (in this case, being issued a ticket), regardless of differences in actual behaviour. In practice, enforcing strict demographic parity without considering actual speeding behaviour may not be desirable, as it could lead to unfair enforcement – either over-penalising or under-penalising certain groups.

Conditional statistical parity allows differences in **probability of a positive outcome** across groups if they can be justified by legitimate, non-discriminatory factors (legitimate risk factors)[90]. For example, suppose that in the traffic example, the likelihood of speeding varies with the time of day, and the AI system uses time of day as a legitimate factor. Within each value of the legitimate factor(s) being conditioned upon, the positive prediction rates must be equal across groups. If we find that the probabilities of a positive outcomes are still different for cars and motorcycles after conditioning on time of day, Conditional Statistical Parity is not achieved.

This metric allows for flexibility in enforcement, acknowledging that different groups may have different behaviours due to legitimate reasons, as long as these differences are justified and not discriminatory.

Metric name	Example of a fair AI system	Purpose	Use it when
Demographic Parity	The system fines 10 % of both cars and motorcycles, regardless of who is actually speeding.	Ensure equal rates of positive outcomes across groups.	Ensuring to prevent disparities in opportunities or penalties between group
Disparate Impact	Motorcycles are fined at 1.5 times the rate of cars; this ratio should be close to 1 to avoid unfair impact.		
Conditional Statistical Parity	Among vehicles driving at night, the system fines the same percentage (X) of cars and motorcycles; and among vehicles driving during the day, the same percentage (Y) of cars and motorcycles is fined.	Allow justified differences based on legitimate factors.	Differences in outcomes are acceptable if they are based on appropriate criteria.

Table 1 Fairness Definitions Based on Predicted Outcome

Definitions based on predicted and actual outcome

The definitions in this section consider not only the outcomes predicted by the AI system but also compare these predictions to the actual outcomes, often referred to as the 'ground truth'. This comparison is performed using confusion matrix elements: number of false positives, false negatives, true positives, true negatives (see Glossary for definitions).

Predictive parity also knowns as **Outcome test**. It focuses on achieving equal positive predictive value (PPV) across groups. PPV is the proportion of true positive outcomes among all instances that the model predicts as positive. In our traffic example, PPV is the probability that vehicles predicted to be speeding are actually speeding.

Predictive parity is achieved if when the AI system flags a vehicle for speeding, the likelihood that the vehicle was actually speeding is the same for both cars and motorcycles. It ensures that among all vehicles predicted to be speeding, the proportion that actually are speeding is the same across groups.

Predictive Parity is suitable when the goal is to ensure that enforcement actions (e.g. issuing fines) are equally reliable across vehicle types. It helps maintain fairness in the traffic violation process and ensures that neither group is disproportionately subjected to incorrect fines.

Equal opportunity (also known as False negative error rate balance, True Positive Rate balance)[91]. An AI system satisfies this definition if individuals from different demographic groups have an equal chance of receiving a positive outcome, given that they belong to the positive class (i.e. they qualify for the positive outcome).

Specifically, it requires that the true positive rate (the proportion of actual positive cases correctly identified by the AI model), TPR, is the same across all observed demographic groups. In our traffic example, TPR is the probability that speeding vehicles are correctly identified. If both cars and motorcycles have equal TPR, it means that the vehicles that are actually speeding are equally likely to be detected, irrespective of their type.

Predictive equality (also known as False positive error rate balance). A classifier satisfies this definition if the false positive rates (FPRs) are equal across different demographic groups. This means that non-speeding vehicles are equally unlikely to be incorrectly fined, whether they are cars or motorcycles. Predictive Equality is important when the aim is to minimise incorrect fines (false positives) equally across vehicle types, preventing unjust penalties.

Equalised odds[92] (also known as Conditional procedure accuracy equality[93], Disparate mistreatment[94]). This definition combines the previous two: a classifier satisfies the definition if it has equal true positive rates and equal false positive rates across different demographic groups.

In our traffic example, this ensures that the detection rates and error rates are balanced between cars and motorcycles. Equalised Odds is appropriate when it's necessary to enforce traffic laws fairly, ensuring that both the chances of catching violators and the risk of penalising innocent drivers are equal across vehicle types.

Conditional use accuracy equality[95] (also known as Predictive Value Parity). This fairness metric ensures that the accuracy of a model's predictions is equal across different demographic groups, conditional on the predicted outcome. This means that for individuals predicted to have a positive outcome, the accuracy of those predictions should be the same for all groups. Similarly, for individuals predicted to have a negative outcome, the accuracy of those predictions should also be consistent across all groups.

In our traffic example, this means that both fines and non-fines are equally reliable across vehicle types. It aims to ensure fairness in both detection violations and confirming compliance, preventing any demographic group from experiencing disproportionately high rates of incorrect predictions. It ensures the reliability of the system's decisions is consistent, depending on the outcome (fine or no fine). It is used to prevent both, unfair penalisation of one group due to higher false positives (incorrect fines), and under-enforcement for one group due to higher false negatives (missed violations).

Overall accuracy equality[96] (also known as Accuracy Parity and Equal accuracy) aims to ensure that the model's overall accuracy – the proportion of all correct predictions, true positives and true negatives together, out of total number of predictions – is the same across different demographic or subgroup populations. In other words, it values the model's accuracy in both predicting when something should happen and when it should not happen.

If we use the traffic example, the Overall accuracy equality ensures that the AI system is not generally biased in accuracy towards any of the two groups, motorcycles or cars. It prevents one group of vehicles experiencing a higher rate of overall errors made by the AI system.

Treatment equality[97] (Cost ratio) – it focuses on the balance of errors made by a classifier rather than its overall accuracy. According to this definition, a classifier achieves treatment equality if the ratio of false negatives to false positives is the same across observed demographic categories. This concept can be extended to account for scenarios where false positives are considered less desirable or more costly than false negatives by a specified cost ratio.

Following the traffic example, the Treatment Equality is relevant when it's important to balance the risk of missing violators (false negatives) and incorrectly penalising compliant drivers (false positives) equally across vehicle types.

Metric name	Example of a fair AI system	Purpose	Use it when
Predictive Parity	Of all vehicles fined, 80 % are actually speeding for both cars and motorcycles.	Ensure equal reliability (accuracy) of positive predictions.	It is important that positive decisions have the same validity across groups.
Equal Opportunity	The system correctly fines 80 % of speeding cars and 80 % of speeding motorcycles.	Ensure equal true positive rates across groups.	Focusing on fairness in detecting actual positives, such as catching violators equally across groups.
Predictive Equality	The system wrongly fines 10 % of non-speeding cars and 10 % of non-speeding motorcycles.	Ensure equal false positive rates across groups.	Aiming to prevent one group from being unfairly subjected to more incorrect penalties
Equalised Odds	The system correctly fines 80 % of speeders and wrongly fines 10 % of non-speeders equally for cars and motorcycles.	Balance true positive and false positive rates across groups.	It is important to ensure fairness among groups in both detecting actual positives and avoiding incorrect positives.
Conditional Use Accuracy Equality	When the system fines or does not fine a vehicle, the chance that it is correct is 80 % for both cars and motorcycles.	Ensure equal accuracy of predictions conditioned on the decision made.	The goal is to have equal trust in the system's decisions across different groups.
Overall Accuracy Equality	The system is 80 % accurate overall for both cars and motorcycles.	Ensure equal overall performance across groups.	Ensuring that no group is generally disadvantaged by a less accurate system.
Treatment Equality	The ratio of wrongly fined to wrongly not fined vehicles is the same for cars and motorcycles.	Balance the ratio of different errors across groups.	Wanting to equalise the burden of errors between groups.

Table 2 Fairness Metrics Based on Predicted and Actual Outcome

Definitions based on predicted probabilities and actual outcome

Fairness definitions based on predicted probabilities and actual outcomes – such as Calibration, Well-Calibration, Balance for Positive Class, and Balance for Negative Class – focus on ensuring that the AI system's probability estimates accurately reflect true outcomes across different groups. In difference to metrics Definitions based on predicted and actual outcome, these metrics focus on the accuracy and reliability of the probability estimates themselves, rather than just the final binary predictions, and are probability-based metrics. They are important when decisions are influenced by predicted probabilities, such as setting thresholds for accepting or rejecting an outcome. Examples of such decision-making AI system would be face recognition systems for automated passport control or AI systems in cargo and baggage screening.

Calibration[98] (also known as Test-fairness[99], Matching conditional frequencies[100]) refers to an idea that the AI system's predicted probabilities accurately reflect the true likelihood of an event occurring. It deems a classifier fair if individuals from all observed demographic groups have the same likelihood of a positive classification for any given predicted probability.

For example, if the AI system predicts a 70 % chance of speeding for vehicles, then approximately 70 % of those vehicles should actually be speeding, for both cars and motorcycles.

Well-calibration[101] extends the concept of calibration by ensuring the calibration holds even when conditioned on additional factors. It requires that within any subgroup defined by certain attributes (e.g. time of day, weather conditions), the predicted probabilities are accurate for each group. For example, well-calibration is satisfied if during daytime when the system predicts a 70 % chance of speeding, about 70 % of vehicles are actually speeding; and if, during night-time, the when the system predicts a 90 % chance of speeding, then about 90 % of vehicles are actually speeding. This concept can be extended to, for example, different subtypes of motorcycles and cars.

Balance for positive class[102] means that the average predicted probability score for individuals in the positive class is equal across all observed demographic groups.

In our traffic example, balance for positive class requires that among all the vehicles that are actually speeding (the positive class), the average predicted score assigned by the AI system is the same across groups. This means that speeding cars and speeding motorcycles should, on average, receive the same risk score from the system.

Balance for negative class[103]. This definition is a reversed version of the previous one, stating that individuals in the negative class from all observed demographic groups, should have equal average predicted probability scores. This means that non-speeding cars and motorcycles should, on average, receive the same risk score from the system.

Metric name	Example of a fair AI system	Purpose	Use it when
Calibration	For both cars and motorcycles, if the system predicts a 70% chance of speeding, 70% actually are speeding.	Ensure predicted probabilities reflect actual outcomes and can be trusted for all groups.	Decisions are influenced by the predicted risk levels, and it's important that these are reliable
Well-calibration	In all conditions (e.g. day or night), when the system predicts a 70% chance of speeding, 70% actually are speeding for both cars and motorcycles.	Ensure calibration holds across all subgroups.	Preventing hidden biases that might only appear under certain conditions.
Balance for positive class	Among actual speeders, the average predicted risk score is the same for cars and motorcycles.	Ensure equal risk assessments among violators.	Avoiding bias in assessing the severity of cases within the positive class (speeders).
Balance for negative class	Among non-speeders, the average predicted risk score is the same for cars and motorcycles.	Ensure fair treatment of non-violators.	Ensuring that non-violators are not unfairly suspected due to higher risk scores.

Table 3 Fairness Metrics Based on Predicted Probabilities and Actual Outcome

Similarity-based metrics

Statistical metrics are observational, that is, they depend only on the joint statistical distribution of classifier, protected attribute, features and outcomes[104]. They focus on the protected attribute, while largely ignoring other circumstances influencing the AI system's lifecycle. To illustrate drawbacks of such approach, imagine a scenario in which an AI system is used by a police department to determine which individuals should be flagged for further investigation based on their risk of engaging in criminal activity. Suppose the AI system assigns a 'risk score' equally across two demographic groups – say, Group A and Group B. From a statistical standpoint, this might suggest fairness, as the same proportion of individuals from each group receives a high-risk score. Suppose for Group A, the risk score is assigned based on a broad range of factors, including recent criminal activity, but for Group B, the score is predominantly based on one specific factor, such as living in a certain neighbourhood known for higher crime rates. Even though the outcome (the proportion of high risk scores) is statistically balanced between the groups, the criteria and context for assigning these scores are substantially different. In this case, although statistical parity might indicate that the classifier is fair, the underlying methodology reveals a disparity in how the assessments are made – Group B individuals are being judged by a narrower and potentially biased criterion.

The following definitions aim to tackle these issues by avoiding the marginalisation of insensitive attributes of the classified subject.

Causal discrimination is satisfied if a classifier produces the same classification for any two subjects with the exact same attributes other than the currently observed protected one[105].

Fairness through awareness defines an AI system as fair if it produces similar outcomes for individuals who are similar based on certain metrics. This is a task-specific metric, as the notion of similarity is defined according to the specific task. For example, a distance metric might define the distance between two individuals as 0 if all attributes except gender are identical, and 1 if any other attributes differ. Similarly, outcome metrics could be set to 0 if the classifier gives the same prediction and 1 if it gives different predictions. Essentially, this approach reduces the problem to defining causal discrimination.

On the other hand, **fairness through unawareness**, also known as **Anti-classification**[106] defines fairness as not using protected attributes in the decision-making process in order to avoid any un-intentional consequences[107].

Causal reasoning metrics

Often, when statistical inferences are performed, it does not mean that causality between different variables is established, but rather correlation. However, depending on the context, correlation alone may be insufficient. Both statistical and similarity-based metrics are not measuring causal relationships between the attributes in the complex dataset points system and they can miss measuring Measurement or proxy bias. Causal graphs are used to represent causal relationships between outcome and attributes. Based on these graphs, a set of structural equations are derived to formalise dependencies and set conditions for unfair AI systems[108] [109] [110]. To illustrate how statistical metrics detect non-existent discrimination by ignoring relationships between attributes, we will use claimed gender-based discrimination in college admission that Pearl describes[111]. Bickel had shown that the reason for a lower college-wide admission rate for women than for men was not discrimination, but the simple fact that women applied in more competitive departments (see Figure 4)[112]. Furthermore, women even had higher acceptance rate then men when data was adjusted for department choice.

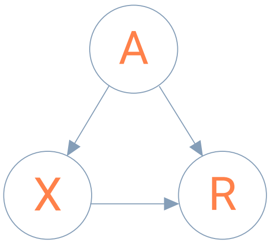


Figure 4 The admission decision R does not only directly depend on gender A, but also on department choice X, which in turn is also affected by gender A[122]. [Source: Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B., 'Avoiding Discrimination Through Causal Reasoning', Advances in Neural Information Processing Systems, 2017.]

To explain the logic of causal reasoning, we will modify an example given in literature[113] by using attributes from a social benefits fraud case (Figure 5).

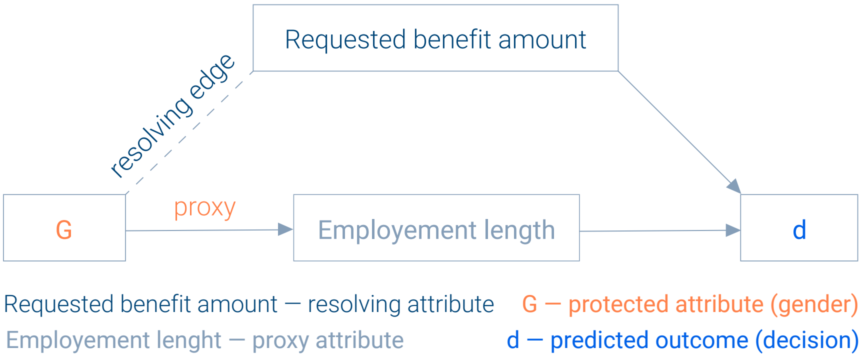

Requested benefit amount
resolving edge
proxy
G
Employement length
d
Requested benefit amount — resolving attribute
G — protected attribute (gender)
Employement lenght — proxy attribute
d — predicted outcome (decision)

Figure 5 Causal graph example representing a social benefits fraud case: G-protected attribute (gender), Employment length – proxy attribute (one can derive applicant's gender from the length of employment), Requested benefit amount - resolving attribute (influenced by the protected attribute in a non-discriminatory way), d – predicted outcome (decision)[114].

Counterfactual fairness[115]. The underlying principle of this measure is that it ensures that an algorithm's decisions remain fair by considering hypothetical scenarios. A decision is considered counterfactually fair if it would remain unchanged in a hypothetical world where a sensitive attribute (such as race, gender, or age) is different, while all other attributes remain the same[116]. It can be expressed in a form of a counterfactual statement of how the world would have to be different for a desirable outcome to occur. For example, 'you were labelled as suspicious for re-offence because your frequency of committing a crime was higher than X, had it been lower, you would not have been labelled by the algorithm'.

More formally, a causal graph achieves counterfactual fairness if the predicted outcome d is not influenced by any descendant of the protected attribute G. For example, in the scenario depicted in Figure 5, d depends on factors such as requested benefit amount and employment length. Since employment length is directly influenced by G, this causal model is not counterfactually fair. This fairness measure is similar to the Causal discrimination metric, but here expressed in the form of a graph.

No unresolved discrimination: is a fairness principle in ML that ensures all identifiable biases and disparities in decision-making processes are addressed and resolved. A causal graph displays no unresolved discrimination if there is no path from the protected attribute G to the predicted outcome d unless it passes through a resolving variable. In our example, Figure 5, the path from G to d via requested benefit amount is considered non-discriminatory because the requested benefit amount acts as a resolving variable.

Conversely, the path through employment length is discriminatory. Therefore, this graph shows unresolved discrimination.

No proxy discrimination: A causal graph is free from proxy discrimination if there is no proxy variable or no path from the protected attribute G to the predicted outcome d. In Figure 5, there is an indirect path from G to d that goes through the proxy variable employment length, indicating the presence of proxy discrimination in the graph.

Higher-level metrics classification

At a high level, fairness definitions can be viewed from two perspectives[117]: individual fairness and group fairness[118]. Metrics defined for individual fairness focus on similar outcomes for similar individuals, while metrics used for group fairness focus on treating different groups in similar ways and is typically identified with protected or sensitive attributes such as gender, race, etc.[119]. In simple words, the goal of individual fairness is similar individuals to be treated similarly, while group fairness seeks for some statistical measure to be equal across different demographic groups[120].

Additionally, these above-mentioned fairness metrics based on moral notions can be connected and grouped under mathematically formalised non-discrimination criteria. These criteria aim to define absence of discrimination in terms of statistical expressions involving random variables describing a classification or decision-making scenario[121]. These criteria are: Independence, Separation, Sufficiency, and Causation.

1. **Independence:** This criterion necessitates that the sensitive attribute is statistically independent of the model's predicted score. In essence, the model's predictions must not be influenced by sensitive characteristics. A model adheres to independence if its predictions are unaffected by these attributes;
2. **Separation:** This criterion is fulfilled when the model's predictions are independent of the sensitive attributes, provided the true label is known. Essentially, this means that within each subgroup of the original dataset, such as those defined by the sensitive attributes, error rates are calculated separately, ensuring the model's performance is consistent across these subgroups;
3. **Sufficiency:** This criterion is met if, given the model's predictions, the sensitive attributes and the true outcomes are independent of each other. This concept is closely related to calibration, ensuring that for any given predicted outcome, the likelihood of that outcome being correct is the same across different groups defined by the sensitive attributes;
4. **Causation**, is the only statistical measure that does not refer to group fairness and corresponds to counterfactual fairness.

As shown in the Table 4, all statistical metrics are group metrics and correspond to one of the three criteria, Independence, Separation and Sufficiency. While all similarity-based and causal reasoning metrics are individual and correspond to the criteria of Causation.

Demographic parity	Independence	Statistical	Group
Conditional Demographic parity	Independence	Statistical	Group
Equal opportunity	Separation	Statistical	Group
Predictive equality	Separation	Statistical	Group
Equalised odds	Separation	Statistical	Group
Balance for positive class	Separation	Statistical	Group
Balance for negative class	Separation	Statistical	Group
Calibration	Sufficiency	Statistical	Group
Conditional use accuracy equality	Sufficiency	Statistical	Group
Predictive parity	Sufficiency	Statistical	Group
Causal discrimination	Causation	Similarity-based	Individual
Fairness through awareness	Causation	Similarity-based	Individual
Fairness through unawareness	Causation	Similarity-based	Individual
Counterfactual fairness	Causation	Causal reasoning	Individual

Table 4 Fairness metrics relation with individual and group fairness and higher definition metrics

Methods to archieve model fairness

The techniques to achieve model fairness, or to mitigate bias, fall into three different categories[123](see Figure 6):

- Pre-processing: Applied to training data before training an ML model, adjusting the training data's feature space to eliminate any correlation with the protected attribute;
- In-training: Applied during the ML model training, imposing fairness constraints into the optimisation process that builds the classifier from the training data;
- Post-processing: Applied on outputs of the trained ML model, modifying the outputs of the trained classifier to ensure they are not correlated with the protected attribute.

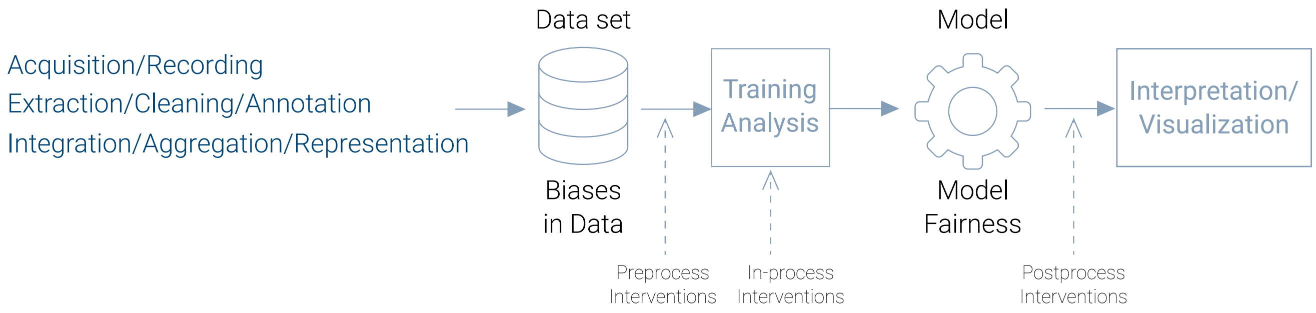


Figure 6. Illustration of bias and fairness in data analytics pipeline (source: Shahbazi, N., Lin, Y., Asudeh, A., Jagadish, H.V., 'Representation Bias in Data: A Survey on Identification and Resolution Techniques', Association for Computing Machinery, 2021.)

In the remaining of this chapter, we will give a short description of bias mitigation techniques in each of the three groups, with references for exact application of the mitigation methods. These mitigation methods are incorporated in the numerous AI bias mitigation toolkits, such as, Fairlearn[124], What-if tool[125], AI Fairness 360[126], etc.

Pre-processing methods

Pre-processing bias mitigation methods involve modifying the data before model training to remove biases related to protected attributes ensuring independence, and an unbiased dataset as the foundation for a fair and accurate ML model. The pre-processing bias mitigation methods focus solely on creating a fair and unbiased dataset, without considering or being affected by how the adjusted data will be used in later stages or applications. The main objective is to ensure fairness in the data itself, leaving the specific use of the transformed data to be determined independently. Commonly used pre-processing methods are listed in Table 5.

Pre-processing methods	Method description
Data reweighting[127]	Assign different weights to data points during the training process to ensure that the model treats various groups more equitably. By adjusting these weights, the algorithm can be guided to give more importance to underrepresented or minority groups, thus reducing the bias in the model's predictions.
Data sampling methods[128]	Oversampling or under sampling data representing different demographic groups, using, e.g. random, systematic, stratified or cluster sampling, to ensure the dataset has the same or similar distribution of different demographic groups as in reality.
Modifying feature representations[129]	Developing an intermediate representation that preserves all crucial information while eliminating any indication of the sensitive attribute, by making distribution of both protected and unprotected groups similar[130] [131] or by using fair dimensionality reduction[132].
Synthetic data generation[133]	Enhance group fairness when the original data is limited by generating artificial data that mimics real data characteristics using generative adversarial network (GAN) or statistical transformations.
Causal methods[134] [135] [136]	In order to mitigate proxy bias, causal methods identify and understand the causal relationships within the data and ensure that the outcome is conditionally independent of the sensitive attributes.

Table 5 Pre-processing de-biasing methods

In-training methods

In-training methods integrate the constraint directly into the optimisation process that builds the classifier from the training data. These methods can be divided in several categories presented in Table 6.

Pre-processing methods	Method description
Regularisation of the objective function[137] [138] [139]	Adding regularisation terms to the objective function that penalises the dependency of the prediction on the sensitive attributes.
Adversarial learning[140] [141] [142] [143]	This technique is used to reduce bias in ML models by using an adversarial setup during training. In this approach, a model (the predictor) is trained to perform its primary task, such as classification or regression, while simultaneously an adversary is trained to predict sensitive attributes (e.g. race, gender) from the same data. The goal is to ensure that the predictor becomes proficient at its primary task without encoding information about the sensitive attributes.
Bandit methods[144]	In cases when it is difficult to define what is fair, so-called bandit methods based on bandit theory, a statistical learning model aiming to make a choice between several actions based on the reward they generate.

Table 6 In-training de-biasing methods

Post-processing methods

Post-processing refers to the process of taking results of the trained classifier and manipulate them in order to achieve independence, i.e. fairness among different groups. Usually, these approaches set different classifier outputs thresholds for different groups in order to achieve equalised odds. Commonly used terms in post-processing bias mitigation algorithms are Single Threshold and Group Threshold. Single Threshold is a uniform threshold for all data points, determined solely by the specified cost ratio, regardless of any protected attributes. Group Thresholds are different decision thresholds for different demographic groups, defined by different protected attributes, and based on the specified cost ratio.

Pre-processing methods	Method description
Output Adjustments[145] [146] [147]	Changing thresholding of classifiers outputs to achieve equalised odds or to maximise accuracy while minimising demographic parity.

Trade-offs between fairness, privacy and quality of the models

LEAs can implement a variety of technical strategies to mitigate the risk of bias in AI systems, although these measures often result in decreased accuracy. For instance, applying an anti-classification fairness approach to an AI model used for predictive policing would necessitate excluding any protected characteristics and their proxies, such as postcodes, from the model.

This exclusion aims to prevent discriminatory outcomes but can also diminish the model's accuracy, as the postcode might have served as an indicator for legitimate risk factors that improve predictive performance.

However, the trade-off between accuracy and fairness is not always inevitable. In some cases, improving both fairness and accuracy is possible by collecting more data, especially if the model's discriminatory outcomes are due to insufficient data on minority populations. To collect more relevant data, it may be necessary to gather information on protected characteristics. Additionally, the collection of personal data for the purpose of detection and mitigation of AI bias in high-risk AI systems is allowed under conditions listed in Article 10(5) of the AI Act[148]. This scenario introduces as another trade-off; between privacy and fairness.

Additionally, explaining the logic behind an AI system can have unintended consequences. Specifically, it can reveal too much about how the model operates, leading individuals to intentionally change their behaviour. This adjusted behaviour could manipulate the system to produce incorrect or misleading results. For example, if people understand how their actions influence the AI's decisions, they might trick the system to receive more favourable outcomes, thus compromising the accuracy and integrity of the AI model. For instance, intentionally misspelling words to avoid detection can cause text analysis algorithms to fail to recognise malicious intentions.

Regarding fairness metrics trade-offs, we know that in the most constrained cases, when rates of positive outcomes differ across groups, it is impossible to achieve calibration while also satisfying Equalised Odds[149] [150]. The Impossibility Theorem Kleinberg et al. (2016)[151] states that no more than one of the three fairness metrics of Demographic Parity, Predictive Parity and Equalised Odds can hold at the same time for a well calibrated classifier and a sensitive attribute capable of introducing machine bias. Additionally, it has been shown that imperfect predictors cannot simultaneously satisfy equal odds and calibration unless the groups have identical base rates, i.e. rates of positive outcomes[152] [153].

Conclusion and recommendations

While applying mathematical methods embedded in fairness toolkits[154] [155] [156] is necessary, it is not sufficient to address all sources of bias in AI systems used in law enforcement. Human, institutional and societal factors also play significant roles in contributing to bias. Understanding AI as a socio-technical system is essential to overcoming these challenges. A socio-technical approach considers the values and behaviours derived from datasets, human interactions with AI systems, and the intricate organisational elements involved in their design, development, deployment and maintenance[157]. This comprehensive approach allows for the inclusion of human, societal, and systemic influences, resulting in a broader and deeper understanding of the benefits and challenges AI systems bring to operational work.

Human decisions play a crucial role in ensuring the fairness of an AI system, starting from the design phase – where the decision on whether AI is needed in the first place is made – to choices about data, target labels and algorithms, and finally to how humans and the organisation interpret the AI system's outcomes and decide on subsequent actions. During this process, many obstacles can arise, such as inherent human biases, the fact that those responsible for AI oversight or end users may not be AI experts, and the tendency for decision-makers to trust AI outputs too much, potentially leading to confirmation bias. Therefore, documenting potential sources of AI bias throughout the AI lifecycle is essential. This documentation enhances model transparency and explainability and helps address issues of bias and fairness.

Impact assessment is particularly important for AI systems used in the law enforcement domain, where the potential for harm is high. This assessment must be conducted throughout all phases of the AI lifecycle, taking into account the context in which the AI system is implemented. It should also include a thorough evaluation of data protection issues, such as the collection, storage, and use of personal data, ensuring compliance with relevant privacy laws and regulations. All potential risks, harms, and data protection concerns that the AI system might cause should be carefully evaluated and documented to ensure transparency, accountability, and the safeguarding of individual rights. More specifically, fundamental rights impact assessment for high-risk AI systems is mandated in Article 27 of the AI Act[158], and it includes describing how and when the system will be used, the categories of people likely to be affected by its use and the risks, and human oversight measures. As advised by BSA software Alliance guide[159], impact on people, context and purpose of the system, the degree of human oversight and the type of training data should be considered by all the stakeholders. In order to avoid confirmation bias, different stakeholders must be included in this analysis. Concretely, personnel included in the oversight of an AI bias management framework implementation should include operational staff, algorithms development teams consisting of IT experts, engineers, data scientists and legal experts.

In addition, ensuring that the team developing an AI project is highly diverse and interdisciplinary is crucial. This diversity should encompass a range of expertise and professional experiences, as well as varied personal backgrounds. Such a composition enables the team to more effectively identify and address potential biases and discriminatory outcomes in the AI system. A heterogeneous team brings multiple perspectives to the table, enhancing the ability to recognise subtle biases that a more homogeneous group might overlook. This diversity is essential for building more fair and equitable AI systems, as it promotes a broader understanding of different societal impacts and challenges.

Recognising that bias is intrinsically linked to the context in which an AI system is deployed is critical. As discussed, and evidenced by the many fairness metrics available, there is no one-size-fits-all definition of fairness. Selecting appropriate fairness metrics requires careful consideration of the specific context. Developing methods for identifying suitable metrics and balancing trade-offs within real-world contexts, such as predictive mapping, is vital. Additionally, it is essential to evaluate gaps in current fairness metrics and processes based on context.

Sometimes it can be difficult for fairness metrics to capture the complexity of the real world. This is why human evaluation is crucial. 'Human evaluation' refers to the process of having people (e.g., domain experts, annotators, or end-users) directly assess, review, or rate the performance, outputs, or behaviour of an AI system. Unlike automated metrics – which rely purely on computational checks – human evaluation uses human judgment to evaluate qualitative aspects that might be difficult to capture using predefined metrics. This can include assessing the relevance, coherence, fairness or interpretability of a model's predictions or generated content, as well as detecting subtle biases, cultural sensitivities or ethical issues that are not easily measurable by automated tests.

Numerous trade-offs must be navigated, as detailed in previous chapters. Developing a systematic approach to managing these trade-offs is essential. Moreover, it is crucial to investigate methods for incorporating contextual information into the machine learning pipeline and to understand how humans interpret and act upon the AI model's outputs. By doing so, we can ensure that the AI system's decisions are more aligned with real-world scenarios and better support human decision-making processes.

A block diagram, showing all phases of AI lifecycle and incorporating AI bias detection and mitigation techniques described in Fairness metrics and Methods to achieve model fairness is shown in Figure 7. A prominent role is given to the importance of having a human-in-the-loop, and the actions taken by humans once they receive the outputs and decisions made by AI systems. Post-processing bias mitigation techniques should be applied on both AI system output and the final decisions made by human experts. Finally, it is essential to clarify whether the final prediction of the AI system and human-in-the-loop is based on causality or correlation as this is of crucial importance for legal domain applications.

Design & Development

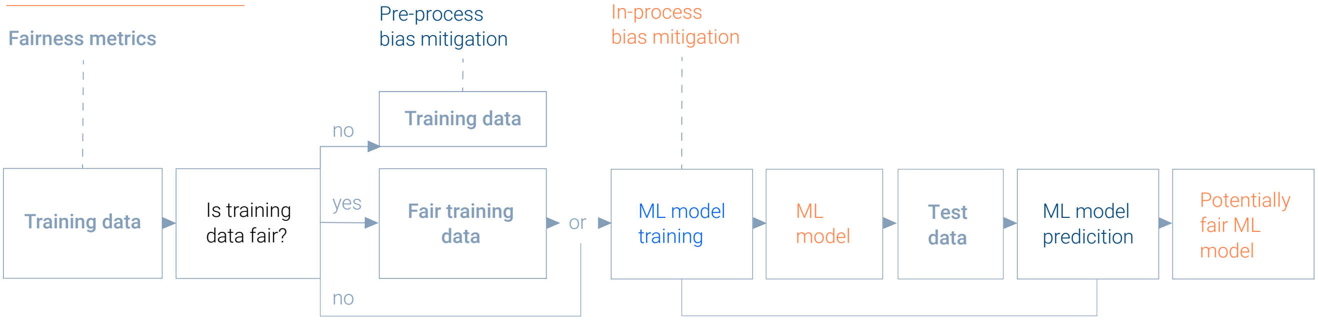
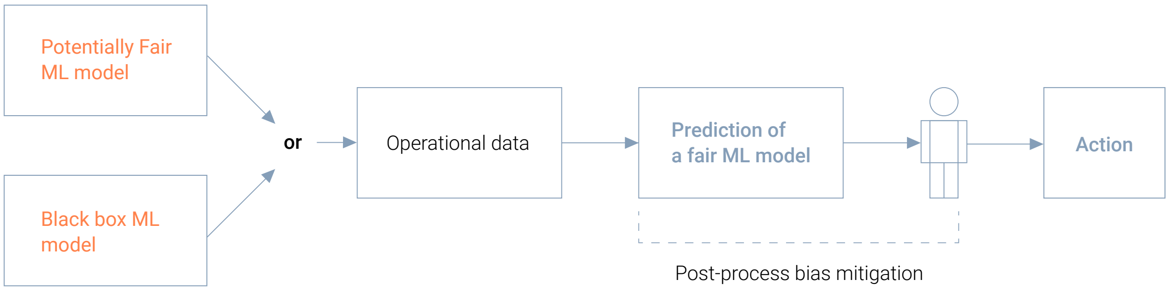


Figure 7 Block diagram of AI bias detection and mitigation process with a human in the loop.

Deployment


Potentially Fair ML model
Black box ML model
or
Operational data
Prediction of a fair ML model
Action
Post-process bias mitigation

We recommend AI bias mitigation frameworks to consist of a set of actions, ensuring best practices for identifying and mitigating risk of AI bias. This set of guidelines should span over all three phases of the AI lifecycle and should focus on three important pillars:

- documentation of every decision, action and process;
- inclusion of diverse stakeholders in order to encapsulate a complex socio-technological perspective of AI bias;
- repeated testing, including impact assessment and human evaluation, throughout the AI lifecycle, and testing of the final human decisions after receiving AI output for bias.

It is recommended to assess and document training, test and evaluation datasets used in the AI system development phase with respect to their origin, collection, motivation behind creation, funding, possible conflict of interest, sensitivity, data protection, composition and technical characteristics.

In summary, addressing AI bias in law enforcement requires a multifaceted approach that integrates technical, human, and contextual factors. By fostering diversity, documenting biases, performing thorough impact assessments, and selecting appropriate fairness metrics, we can develop AI systems that are fairer, more transparent, and better suited to support just and equitable law enforcement practices.

ENDNOTES

[1] European Parliament, Artificial Intelligence and public services, 2021, [accessed 31/05/2024], https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/662936/IPOL_BRI(2021)662936_EN.pdf

[2] EUROPOL, AI and policing, September, 2024, [accessed 15/11/2024], https://www.europol.europa.eu/publication-events/main-reports/ai-and-policing

[3] Case number / cause list number: C/09/550982 / HA ZA 18-388, ECLI:NL:RBDHA:2020:1878, Rechtbank Den Haag, C-09-550982-HA ZA 18-388 (English), [accessed 31/05/2024], https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:1878&showbutton=true&keyword=C%252f09%252f550982%2B%252f%2BHA%2BZA%2B18-388&idx=2

[4] A. Rachovitsa, Johann N.., 'The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case', Human Rights Law Review, 2022

[5] Case number / cause list number: C/09/550982 / HA ZA 18-388, ECLI:NL:RBDHA:2020:1878, Rechtbank Den Haag, C-09-550982-HA ZA 18-388 (English), [accessed 06/09/2024], https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:1878&showbutton=true&keyword=C%252f09%252f550982%2B%252f%2BHA%2BZA%2B18-388&idx=2

[6] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.

[7] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 13 June 2024, Official Journal of the European Union, [accessed 29/08/2024], https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[8] European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Official Journal of the European Union, [accessed 29/08/2024], https://eur-lex.europa.eu/eli/reg/2022/2065/oj

[9] European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council 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 (General Data Protection Regulation). Official Journal of the European Union, [accessed 29/08/2024], http://data.europa.eu/eli/reg/2016/679/oj

[10] European Parliament, EU AI Act: first regulation on artificial intelligence, 2023, [accessed 31/05/2024], https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[11] ISO, 'Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making', 2021, ISO/IEC TR 24027:2021, [accessed 31/05/2024], https://www.iso.org/standard/77607.html

[12] ISO, 'Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence', 2020, ISO-24028:2020, [accessed 31/05/2024], https://www.iso.org/standard/77608.html

[13] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P., 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence', NIST Special Publication 1270, 2022, [accessed 31/05/2024],

[14] Barocas S, Hardt M., Narayanan A., 'FAIRNESS AND MACHINE LEARNING Limitations and Opportunities', 2023, [accessed 31/05/2024], https://fairmlbook.org/

[15] Perkins, K., Wiley, S., 'Minorities. In: Teo, T. (eds) Encyclopedia of Critical Psychology', Springer, New York, NY, 2014, [accessed 31/05/2024], https://link.springer.com/referenceworkentry/10.1007/978-1-4614-5583-7_188

[16] FRA, 'Bias in algorithms – AI and discrimination', 2023.

[17] Australian Law Reform Commission, 'Judicial Impartiality – Cognitive and Societal Biases in Judicial Decision-Making', 2021, [accessed 31/05/2024], https://www.alrc.gov.au/publication/cognitive-social-biases-ji6/

[18] Handelsman J., Sakraney N., 'Implicit Bias', White House Office of Science and Technology Policy, 2012. [accessed 31/05/2024], https://fellowshipapp.aaas.org/library/doclib/2024/9/Implicit-Bias-OSTP-Paper.pdf

[19] Biernat M., Manis, M., Nelson T., ‘Stereotypes and standards of judgment’, Journal of Personality and Social Psychology 66:5-20., 1991.

[20] Sagar HA, Schofield JW., ‘Racial and behavioral cues in black and white children’s perceptions of ambiguously aggressive acts.’, J Pers Soc Psychol., 1980, [accessed 31/05/2024], https://pubmed.ncbi.nlm.nih.gov/7431207/

[21] Sagar HA, Schofield JW., ‘Racial and behavioral cues in black and white children’s perceptions of ambiguously aggressive acts.’, J Pers Soc Psychol., 1980, [accessed 31/05/2024], https://pubmed.ncbi.nlm.nih.gov/7431207/ Handelsman J., Sakraney N., ‘Implicit Bias’, White House Office of Science and Technology Policy, 2012 [accessed 31/05/2024],

[22] Handelsman J., Sakraney N., ‘Implicit Bias’, White House Office of Science and Technology Policy, 2012 [accessed 31/05/2024],

[23] Dasgupta N., Greenwald A. G., ‘On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals’, Journal of Personality and Social Psychology 81: 800-814, 2001.

[24] Dasgupta N., Asgari S., ‘Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping’, Journal of Experimental Social Psychology 40:642-658, 2004.

[25] Zawadzki M. J., Danube C. L., Shields S. A., ‘How to talk about gender inequity in the workplace: Using WAGES as an experiential learning tool to reduce reactance and promote self-efficacy.’, Sex Roles 67: 605-616, 2012.

[26] Carnes M., Devine P. G., Manwell L. B., Byars-Winston A., Fine E., Ford C. E., Sheridan J., ‘The effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial.’ Academic Medicine 90:221-230, 2015.

[27] AI HLEG, ‘A definition of AI: Main capabilities and scientific disciplines.’, 2018, [accessed 31/05/2024], https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

[28] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 13 June 2024, Official Journal of the European Union, [accessed 29/08/2024], https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[29] Samuel A. L., ‘Some Studies in Machine Learning Using the Game of Checkers,’ IBM Journal of Research and Development, vol. 3, no. 3, pp. 210–229, 1959.

[30] EUROPOL, AI and policing, 2024, https://www.europol.europa.eu/publication-events/main-reports/ai-and-policing

[31] European Parliament, the Council of Ministers and the European Commission, ‘Charter of Fundamental Rights of the EU’, 2020, [accessed 31/05/2024], https://commission.europa.eu/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en

[32] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.

[33] Fuster G. G., ‘Artificial Intelligence and Law Enforcement. Impact on Fundamental Rights’, 2020, [accessed 31/05/2024], https://www.europarl.europa.eu/thinktank/en/document/IPOLSTU(2020)656295

[34] Kafteranis D., Sachoulidou A., Turksen U., ‘Artificial Intelligence in Law Enforcement Settings’, 2023, [accessed 31/05/2024], https://eucrim.eu/articles/artificial-intelligence-in-law-enforcement-settings/

[35] FRA, ‘Bias in algorithms – AI and discrimination’, 2023.

[36] FRA, ‘Bias in algorithms – AI and discrimination’, 2023.

[37] Kleinberg J., Lakkaraju H., Leskovec J., Ludwig J., Mullainathan S., ‘Human Decisions and Machine Predictions’, The Quarterly Journal of Economics, Volume 133, Issue 1, February 2018, Pages 237–293, [accessed 31/05/2024], https://doi.org/10.1093/qje/qjx032

[38] Leslie D., ‘Understanding artificial intelligence ethics and safety’, Alan Turing Institute, 2020.

[39] Edenberg E., Wood A., ‘Disambiguating Algorithmic Bias: From Neutrality to Justice’, AIES, 2023, [accessed 31/05/20-24], https://dl.acm.org/doi/abs/10.1145/3600211.3604695

40 Suresh H., Guttag J. V., 'A Framework for Understanding Unintended Consequences of Machine Learning', 2020, [accessed 31/05/2024], https://arxiv.org/pdf/1901.10002.pdf

41 de Silva D., Alahakoon D., 'An artificial intelligence life cycle: From conception to production', Patterns, 2022, [accessed 31/05/2024], https://doi.org/10.1016/j.patter.2022.100489

42 Selbst A. D., Vertesi J., 'Fairness and abstraction in sociotechnical systems', In Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019.

43 Suresh H., Guttag J. V., 'A Framework for Understanding Unintended Consequences of Machine Learning', MIT, 2020.

44 Schwartz R., Vasslev A., Greene K. K., Burt A., Hall P., 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence', 2022, [accessed 31/05/2024], https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence

45 https://www.nist.gov/

46 Schwartz R., Vasslev A., Greene K. K., Burt A., Hall P., 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence', 2022, [accessed 31/05/2024], https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

47 Schwartz R., Vasslev A., Greene K. K., Burt A., Hall P., 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence', 2022, [accessed 31/05/2024], https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

48 Schwartz R., Vasslev A., Greene K. K., Burt A., Hall P., 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence', 2022, [accessed 31/05/2024], https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

49 Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1668–1678, [accessed 18/11/2024], https://homes.cs.washington.edu/~nasmith/papers/sap+card+gabriel+choi+smith.acl19.pdf

50 Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. Proceedings of the Third Workshop on Abusive Language Online, 25–35. [accessed 18/11/2024], https://doi.org/10.48550/arXiv.1905.12516

51 Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of 'Bias' in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454–5476. [accessed 18/11/2024], https://doi.org/10.48550/arXiv.2005.14050

52 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

53 Kiritchenko, S., & Mohammad, S. M. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM 2018), 43–53. https://arxiv.org/abs/1805.04508

54 Nozza, D. (2021). Exposing the Limits of Zero-Shot Cross-Lingual Hate Speech Detection. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), 907–914. https://aclanthology.org/2021.acl-short.114.pdf

55 Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring Stereotypical Bias in Pre-trained Language Models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021), 5356–5371. https://doi.org/10.48550/arXiv.2004.09456

56 Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., ... & Chang, K.-W. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), 1630–1640. https://doi.org/10.48550/arXiv.1906.08976

57 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

58 Li, Y., Feng, S., Qian, T., Chen, Y., & Yu, M. (2020). Towards Debiasing Sentiment Classification via Gender-Neutral Word Embedding. Proceedings of the 28th International

Conference on Computational Linguistics (COLING 2020), 4113–4122. https://dl.acm.org/doi/10.1145/3336191.3371779

59 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Recital 5, 13 June 2024, Official Journal of the European Union, [accessed 29/08/2024], https://eur-lex.europa.eu/eli/reg/2024/1689/oj

60 European Parliament, 'Regulatory divergences in the draft AI act, differences in public and private sector obligations', May 2022.

61 Verma S., Rubin J., 'Fairness Definitions Explained', ACM/IEEE International Workshop on Software Fairness, 2018.

62 FRA, 'Handbook on European non-discrimination law', (direct discrimination p. 43, indirect discrimination p. 53), 2018, [accessed 31/05/2024], https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-handbook-non-discrimination-law-2018_en.pdf

63 Van Brakel R., 'Pre-emptive big data surveillance and its (dis)empowering consequences: the case of predictive policing', In Exploring the boundaries of big data, 117-141, Amsterdam University Press, 2016.

64 FRA, 'Bias in algorithms – AI and discrimination', 2023.

65 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 5(d), 13 June 2024, Official Journal of the European Union, [accessed 29/08/2024], https://eur-lex.europa.eu/eli/reg/2024/1689/oj

66 TELEFI, Summary Report of the project 'Towards the European Level Exchange of Facial Images', 2021, [accessed 31/05/2024], https://www.telefi-project.eu/sites/default/files/TELEFI_SummaryReport.pdf

67 ENFSI, 'Best Practice Manual for Facial Image Comparison', 2018, [accessed 31/05/2024], https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf

68 Fair Trials, Automating injustice: the use of artificial intelligence & automated Decision-making systems in criminal justice in Europe, 2021, https://www.fairtrials.org/app/uploads/2021/11/Automating_Injustice.pdf

69 Amnesty International, We sense trouble: Automated Discrimination and Mass Surveillance in Predictive Policing in the Netherlands, 2020

70 Egbert, S. and S. Krasmann. 2019. Predictive policing: Not yet, but soon preemptive? Policing and Society, doi:10.1080/10439463.2019.1611821.

71 Brakel, R. V. (2021). Rethinking predictive policing: Towards a holistic framework of democratic algorithmic surveillance. In Algorithmic societies: Power, knowledge and technology in the age of algorithms (pp. 104-118)

72 FRA, 'Bias in algorithms – AI and discrimination', 2023.

73 FRA, 'Bias in algorithms – AI and discrimination', 2023.

74 Heikkilä M., 'Dutch scandal serves as a warning for Europe over risks of using algorithms', 2022, https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

75 Wired, 'This algorithm could ruin your life', 2023, [accessed 31/05/2024], https://www.wired.com/story/welfare-algorithms-discrimination/#:~:text=More%20than%2020%2C000%20families%20were,in%20response%20in%20January%202021

76 Schwartz J., Steffensmeier D, Zhong Hm, Ackerman J., 'Trends in the Gender Gap in Violence: Reevaluating NCVS and Other Evidence', Criminology, 2009.

77 FarringtonD. P., 'Age and Crime', Crime and Justice, 1986.

78 Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., & Gilbert, J.E., 'A review of predictive policing from the perspective of fairness'. Artificial Intelligence and Law, 2021.

79 Nissan E, 'Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement', Ai & Society, 2021.

80 Asaro P. M., 'Ai ethics in predictive policing: From models of threat to an ethics of care', IEEE Technology and Society Magazine, 2019.

81 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 5(d), 13 June 2024, Official Journal of the European Union, [accessed 29/08/2024], https://eur-lex.europa.eu/eli/reg/2024/1689/oj

82 AI HLEG, 'Ethics guidelines for trustworthy AI', 2019, [accessed 31/05/2024], https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

83 Verma S., Rubin J., 'Fairness Definitions Explained', In Proceedings of the International Workshop on Software Fairness, 2018, [accessed 31/05/2024], https://doi.org/10.1145/3194770.3194776

84 Verma S., Rubin J., 'Fairness Definitions Explained', In Proceedings of the International Workshop on Software Fairness, 2018, [accessed 31/05/2024], https://doi.org/10.1145/3194770.3194776

85 Bellamy, R.K., Dey, K., Hind, M., Homan, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al., 'AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias', arXiv preprint, 2018, [accessed 31/05/2024], https://arxiv.org/abs/1810.01943.

86 Verma S., Rubin J., 'Fairness Definitions Explained', In Proceedings of the International Workshop on Software Fairness, 2018, [accessed 31/05/2024], https://doi.org/10.1145/3194770.3194776

87 Kusner, M.J., Loftus, J., Russell, C., Silva, R., 'Counterfactual Fairness', Advances in Neural Information Processing Systems, 2017.

88 MathWorks, 'Explore Fairness Metrics for Credit Scoring Model', 2024, [accessed 31/05/2024], https://www.mathworks.com/help/risk/explore-fairness-metrics-for-credit-scoring-model.html

89 MathWorks, 'Explore Fairness Metrics for Credit Scoring Model', 2024, [accessed 31/05/2024], https://www.mathworks.com/help/risk/explore-fairness-metrics-for-credit-scoring-model.html

90 Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A., 'Algorithmic Decision Making and the Cost of Fairness', Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.

91 Hardt, M., Price, E., Srebro, N., 'Equality of Opportunity in Supervised Learning', Advances in Neural Information Processing Systems, 2016.

92 Hardt, M., Price, E., Srebro, N., 'Equality of Opportunity in Supervised Learning', Advances in Neural Information Processing Systems, 2016.

93 Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A., 'Fairness in Criminal Justice Risk Assessments: The State of the Art', 2017.

94 Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P., 'Fairness Beyond Disparate Treatment Disparate Impact: Learning Classification Without Disparate Mistreatment', Proceedings of the 26th International World Wide Web Conference (WWW '17), 2017.

95 Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A., 'Fairness in Criminal Justice Risk Assessments: The State of the Art', 2017.

96 Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A., 'Fairness in Criminal Justice Risk Assessments: The State of the Art', 2017.

97 Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A., 'Fairness in Criminal Justice Risk Assessments: The State of the Art', 2017.

98 Chouldechova, A., 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments', Big Data, 2017, [accessed 31/05/2024], https://www.liebertpub.com/doi/abs/10.1089/big.2016.0047?journalCode=big

99 Chouldechova, A., 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments', Big Data, 2017, [accessed 31/05/2024], https://www.liebertpub.com/doi/abs/10.1089/big.2016.0047?journalCode=big

100 Hardt, M., Price, E., Srebro, N., 'Equality of Opportunity in Supervised Learning', Advances in Neural Information Processing Systems, 2016.

101 Kleinberg, J.M., Mullainathan, S., Raghavan, M., 'Inherent Trade-Offs in the Fair Determination of Risk Scores', ITCS, 2017.

102 Kleinberg, J.M., Mullainathan, S., Raghavan, M., 'Inherent Trade-Offs in the Fair Determination of Risk Scores', ITCS, 2017.

103 Kleinberg, J.M., Mullainathan, S., Raghavan, M., 'Inherent Trade-Offs in the Fair Determination of Risk Scores', ITCS, 2017.

104 Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B., 'Avoiding Discrimination Through Causal Reasoning', Advances in Neural Information Processing Systems, 2017.

105 Galhotra, S., Brun, Y., Meliou, A., 'Fairness Testing: Testing Software for Discrimination', Proceedings of the 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE '17), 2017.

106 Corbett-Davies, S., Goel, S., 'The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning', arXiv preprint, 2018, [accessed 31/05/2024], https://arxiv.org/abs/1808.00023

107 Kusner, M.J., Loftus, J., Russell, C., Silva, R., 'Counterfactual Fairness', Advances in Neural Information Processing Systems, 2017.

108 Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B., 'Avoiding Discrimination Through Causal Reasoning', Advances in Neural Information Processing Systems, 2017.

109 Kusner, M.J., Loftus, J.R., Russell, C., Silva, R., 'Counterfactual Fairness', Advances in Neural Information Processing Systems, 2017

110 Nabi, R., Shpitser, I., 'Fair Inference on Outcomes', AAAI, 2018.

111 Pearl, J., 'Causality', Cambridge University Press, 2009.

112 Bickel, P.J., Hammel, E.A., O'Connell, J.W., 'Sex Bias in Graduate Admissions: Data from Berkeley', Science, 1975.

113 Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B., 'Avoiding Discrimination Through Causal Reasoning', Advances in Neural Information Processing Systems, 2017.

114 Verma, S., Rubin, J., 'Fairness Definitions Explained', ACM/IEEE International Workshop on Software Fairness, 2018.

115 Verma, S., Rubin, J., 'Fairness Definitions Explained', ACM/IEEE International Workshop on Software Fairness, 2018.

116 Kusner, M.J., Loftus, J., Russell, C., Silva, R., 'Counterfactual Fairness', Advances in Neural Information Processing Systems, 2017.

117 Kusner, M.J., Loftus, J., Russell, C., Silva, R., 'Counterfactual Fairness', Advances in Neural Information Processing Systems, 2017.

118 Barocas S, Hardt M., Narayanan A., 'FAIRNESS AND MACHINE LEARNING Limitations and Opportunities', 2023, [accessed 31/05/2024], https://fairmlbook.org/

119 Shahbazi, N., et al., 'Representation Bias in Data: A Survey on Identification and Resolution Techniques', Association for Computing Machinery, 2021.

120 Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., Gilbert, J.E., 'A Review of Predictive Policing from the Perspective of Fairness', 2021

121 https://aif360.res.ibm.com/resources#guidance

122 Barocas S, Hardt M., Narayanan A., 'FAIRNESS AND MACHINE LEARNING Limitations and Opportunities', 2023, [accessed 31/05/2024], https://fairmlbook.org/

123 Barocas S, Hardt M., Narayanan A., 'Fairness and machine learning Limitations and Opportunities', 2023, [accessed 31/05/2024], https://fairmlbook.org/

124 Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K., 'Fairlearn: Assessing and Improving Fairness of AI Systems', arXiv preprint, 2023, [accessed 31/05/2024], https://arxiv.org/abs/2303.16626, https://fairlearn.org, https://github.com/fairlearn/fairlearn

125 Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J., 'The What-If Tool: Interactive Probing of Machine Learning Models', arXiv preprint, 2019, [accessed 31/05/2024], https://arxiv.org/abs/1907.04135., https://pair-code.github.io/what-if-tool/

126 Bellamy, et al., AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, 2018, https://arxiv.org/abs/1810.01943

127 Calders, T., Kamiran, F., Pechenizkiy, M., 'Building Classifiers with Independency Constraints', 2009 IEEE International Conference on Data Mining Workshops, 2009.

128 Kamiran, F., Calders, T., 'Data Preprocessing Techniques for Classification Without Discrimination', Knowledge and Information Systems, 2011.

129 Zemel, R.S., Wu, L.Y., Swersky, K., Pitassi, T., Dwork, C., 'Learning Fair Representations', ICML, 2013.

130 Zemel, R.S., Wu, L.Y., Swersky, K., Pitassi, T., Dwork, C., 'Learning Fair Representations', ICML, 2013.

131 Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K., 'Optimized Pre-processing for Discrimination Prevention', Advances in Neural Information Processing Systems, 2017, pp. 3992-4001, [accessed 31/05/2024], https://papers.nips.cc/paper/2017/file/9a49a25d845a483fae4be7e34136 8e36-Paper.pdf.

132 Samadi, S., Tantipongpipat, U., Morgenstern, J., Singh, M., Vempala, S., 'The Price of Fair PCA: One Extra Dimension', Advances in Neural Information Processing Systems, 2018, pp. 10976-10987, [accessed 31/05/2024], https://proceedings.neurips.cc/paper/2018/file/cc4af25fa9d2d5c953496579b75f6f6c-Paper.pdf.

133 Yuan, L., Zhang, X., Wu, X., 'FairGAN: Fairness-aware Generative Adversarial Networks', 2018 IEEE International Conference on Big Data (Big Data), 2018.

134 Galhotra, S., Brun, Y., Meliou, A., 'Fairness Testing: Testing Software for Discrimination', Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, 2017.

135 Glymour, B., Herington, J., 'Measuring the Biases That Matter: The Ethical and Causal Foundations for Measures of Fairness in Algorithms', Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019.

136 Salimi, B., Rodriguez, L., Howe, B., Suciu, D., 'Interventional Fairness: Causal Database Repair for Algorithmic Fairness', Proceedings of the 2019 International Conference on Management of Data, 2019.

137 Bechavod, Y., Ligett, K., 'Penalizing Unfairness in Binary Classification', arXiv preprint arXiv:1707.00044, 2017.

138 Woodworth, B., Gunasekar, S., Ohannessian, M.I., Srebro, N., 'Learning Non-discriminatory Predictors', Conference on Learning Theory, PMLR, 2017.

139 Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H., 'A Reductions Approach to Fair Classification', International Conference on Machine Learning, PMLR, 2018.

140 Wadsworth, C., Vera, F., Piech, C., 'Achieving Fairness Through Adversarial Learning: An Application to Recidivism Prediction', arXiv preprint arXiv:1807.00199, 2018, [accessed 31/05/2024], https://doi.org/10.48550/arXiv.1807.00199

141 Han, X., Baldwin, T., Cohn, T., 'Towards Equal Opportunity Fairness Through Adversarial Learning', arXiv preprint arXiv:2203.06317, 2022, [accessed 31/05/2024], https://arxiv.org/abs/2203.06317

142 Han, X., Baldwin, T., Cohn, T., 'Towards Equal Opportunity Fairness Through Adversarial Learning', arXiv preprint arXiv:2203.06317, 2022, [accessed 31/05/2024], https://arxiv.org/abs/2203.06317

143 Zhang, B.H., Lemoine, B., Mitchell, M., 'Mitigating Unwanted Biases with Adversarial Learning', Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018.

144 Metevier, B., Giguere, S., Brockman, S., Kobren, A., Brun, Y., Brunskill, E., Thomas, P., 'Offline Contextual Bandits with High Probability Fairness Guarantees', 33rd Annual Conference on Neural Information Processing Systems (NeurIPS), Advances in Neural Information Processing Systems 32, 2019.

145 Barocas S, Hardt M., Narayanan A., 'Fairness and Machine Learning Limitations and Opportunities', 2023, [accessed 31/05/2024], https://fairmlbook.org/

146 Hardt, M., Price, E., Srebro, N., 'Equality of Opportunity in Supervised Learning', Advances in Neural Information Processing Systems, 2016, [accessed 31/05/2024], https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

147 Kamiran, F., Mansha, S., Karim, A., Zhang, X., 'Exploiting Reject Option in Classification for Social Discrimination Control', Information Sciences, 2018, [accessed 31/05/2024], https://dl.acm.org/doi/10.1016/j.ins.2017.09.064.

148 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 10(5), 13 June 2024, Official Journal of the European Union, [accessed 05/09/2024], https://eur-lex.europa.eu/eli/reg/2024/1689/oj

149 Chouldechova, A., 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments', Big Data, 2017, [accessed 31/05/2024], https://www.liebertpub.com/doi/abs/10.1089/big.2016.0047?journalCode=big

[150] Chouldechova, A., ‘Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments’, Big Data, 2017, [accessed 31/05/2024], https://www.liebertpub.com/doi/abs/10.1089/big.2016.0047?journalCode=big

[151] Kleinberg, J., Mullainathan, S., Raghavan, M., ‘Inherent Trade-Offs in the Fair Determination of Risk Scores’, arXiv preprint, 2016, [accessed 31/05/2024], https://arxiv.org/abs/1609.05807

[152] Kleinberg, J., Mullainathan, S., Raghavan, M., ‘Inherent Trade-Offs in the Fair Determination of Risk Scores’, arXiv preprint, 2016, [accessed 31/05/2024], https://arxiv.org/abs/1609.05807

[153] Chouldechova, A., ‘Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments’, Big Data, 2017, [accessed 31/05/2024], https://www.liebertpub.com/doi/abs/10.1089/big.2016.0047?journalCode=big

[154] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K., ‘Fairlearn: Assessing and Improving Fairness of AI Systems’, arXiv preprint, 2023, [accessed 31/05/2024], https://arxiv.org/abs/2303.16626, https://fairlearn.org, https://github.com/fairlearn/fairlearn

[155] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J., ‘The What-If Tool: Interactive Probing of Machine Learning Models’, arXiv preprint, 2019, [accessed 31/05/2024], https://arxiv.org/abs/1907.04135., https://pair-code.github.io/what-if-tool/

[156] Bellamy, R.K., Dey, K., Hind, M., Hemanth, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y., ‘‘AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic ’Bias’, arXiv preprint, 2018, [accessed 31/05/2024], https://arxiv.org/abs/1810.01943.

[157] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P., ‘Towards a Standard for Identifying and Managing Bias in Artificial Intelligence’, NIST Special Publication 1270, 2022, [accessed 31/05/2024], https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

[158] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 27, 13 June 2024, Official Journal of the European Union, [accessed 29/08/2024], https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[159] BSA, Confronting Bias: BSA’s framework to Build Trust in AI, 2024, [accessed 31/05/2024], https://www.bsa.org/reports/confronting-bias-bsas-framework-to-build-trust-in-ai

About the Europol Innovation Lab

Technology has a major impact on the nature of crime. Criminals quickly integrate new technologies into their modus operandi, or build brand-new business models around them. At the same time, emerging technologies create opportunities for law enforcement to counter these new criminal threats. Thanks to technological innovation, law enforcement authorities can now access an increased number of suitable tools to fight crime. When exploring these new tools, respect for fundamental rights must remain a key consideration.

In October 2019, the Ministers of the Justice and Home Affairs Council called for the creation of an Innovation Lab within Europol, which would develop a centralised capability for strategic foresight on disruptive technologies to inform EU policing strategies.

Strategic foresight and scenario methods offer a way to understand and prepare for the potential impact of new technologies on law enforcement. The Europol Innovation Lab's Observatory function monitors technological developments that are relevant for law enforcement and reports on the risks, threats and opportunities of these emerging technologies.

www.europol.europa.eu