Bias in AI: tackling the issues through regulations and standards

Adrian Byrne, CeADAR, University College Dublin
Donghyeok Lee, CeADAR, University College Dublin
Quan Le, CeADAR, University College Dublin

Keywords: AI, bias detection, bias mitigation, discrimination, fairness, regulations, standards, sandboxing, red teaming

Introduction

As companies increase their use of Artificial Intelligence (AI), people are questioning the extent to which human biases have made their way into AI systems (Shedding light on AI bias with real world examples). AI is becoming more involved, not only in the automation of human works, but in decision-making in critical areas that are key to the lives of the users, such as employment, loans, and the justice system. However, when AI bias goes unaddressed, it can impact an organisation's success and hinder people's ability to participate in the economy and society (What is AI bias?). These AI systems, if not responsibly developed, can reconstruct existing biases to even worse levels and give unfair or discriminatory results.

In modern society, where truth and facts are increasingly reduced to quantifiable statistical evidence, the limitations of this truth may be appropriate for most of a given population but can lead to inaccurate or even erroneous outcomes for minority groups or those who fall far from the statistical average. The concept of the Human Starburst, as presented by Jutta Treviranus, clearly visualises this issue, as shown in ***Fig. 1***. Designed as a series of overlapping Gaussian/normal bell-shaped curves representing different quantities of interest, the "outliers" in the tails away from the central tendency have less in common with the majority as well as the other "outliers". This phenomenon is more complicated in AI systems. While AI may generally make accurate decisions for the average person (i.e., the middle 80%), it is likely to make unfavourable decisions for those on the periphery (i.e. the peripheral 20%). This implies that AI systems may reinforce or even exacerbate existing biases. Therefore, in the design and implementation of AI systems, the idea of the 'Human Starburst' is an important consideration rather than relying naively on statistical averages.

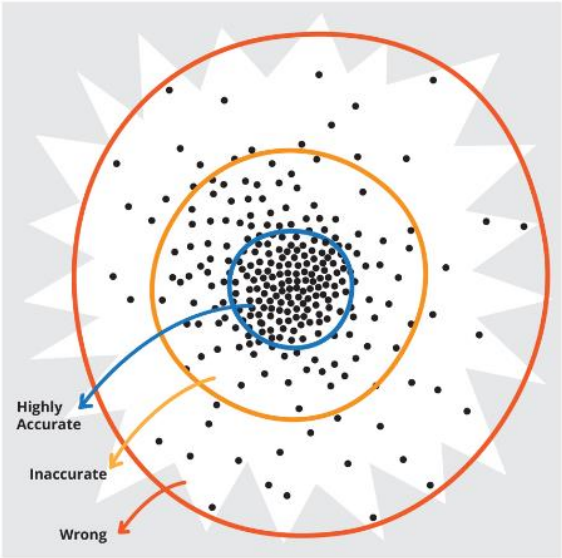

Figure 1 Human Starburst

The increasing uptake of, and dependence on, AI demonstrates the need to make sure that such systems are used fairly and in a manner that is easily understandable. For example, it is crucial to ensure that these decisions do not reflect discriminatory behaviour toward certain groups or populations. To address these concerns, proper regulations and standards have an important role to play. Thus, this paper aims to provide an overview of the role of regulations and standards in the identification of bias in AI and in the formation of public policy as well as industry standards and guidelines.

AI Bias and its Challenges

Bias refers to the systematic errors that occur in decision-making processes, leading to unfair outcomes in human-centred AI applications (Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies). AI biases can be derived from biased training data, skewed algorithms or unintentional decisions made by developers which might have unknown, adverse effects. For example, if the AI system is learning from data that was taken from a prejudiced society that discriminates against non-white people and/or non-males, then the outcomes of that AI system will also discriminate against non-white people and/or non-males. The system is likely to replicate and even amplify these biases in its predictions and decisions as seen in *Fig 2*. The figure shows how the AI lifecycle can potentially lead to inaccurate and/or unfair outcomes in healthcare applications.

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

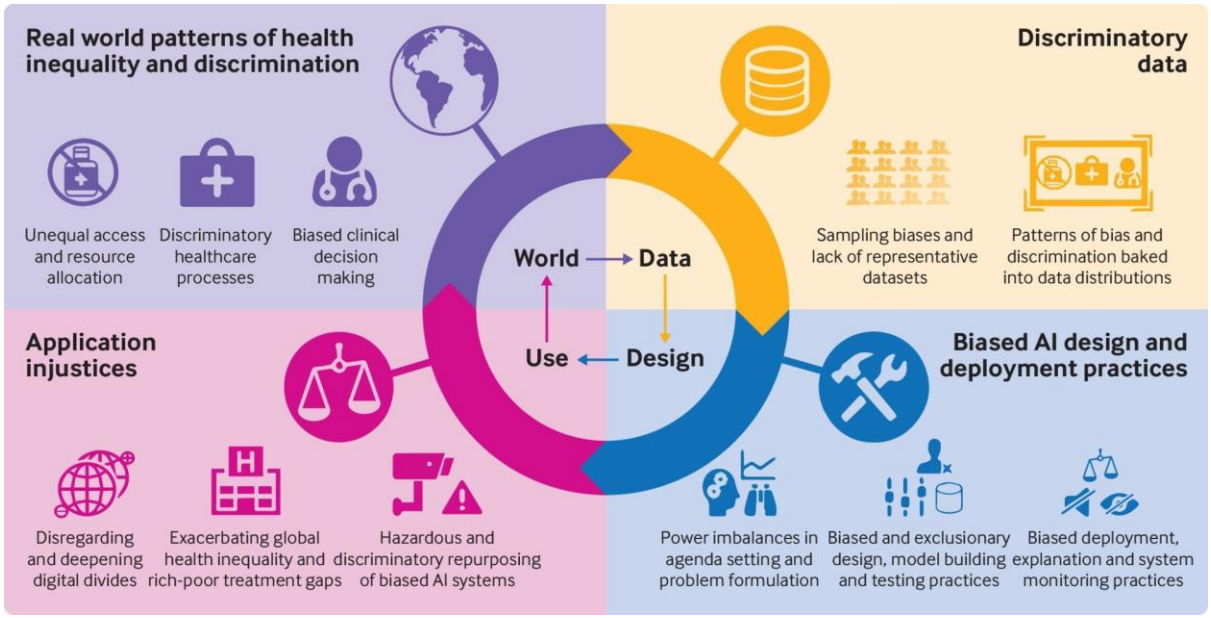

Figure 2 Inequality and discrimination in the design and use of AI in healthcare applications

Bias can occur at various stages in the development and deployment of AI systems. Several types of bias can occur while transitioning from data to algorithms. **Measurement bias** deals with the choice of features used in the model, how they are utilised and how they are measured while **omitted variable bias** occurs when some key variables are excluded from the model. **Representation bias** refers to the data sampling during the data collection process while **aggregation bias** occurs when wrong inferences are made on an individual based on what is observed for the whole group (akin to the Human Starburst figure above). Moreover, there is **sampling bias** since only subgroups are sampled non-randomly, and **linking bias** arises through network features that come with the users' connections, activities and/or interactions.

When transitioning from algorithms to users, several other types of bias can occur. **Algorithmic bias** is the bias that is inherent in the algorithm rather than the input data. **Interaction bias** is the bias that is brought about by the operations of the user interface or the operations that the user deliberately selects. **Popularity bias** refers to the phenomenon that the most popular items get viewed more often and can be influenced by fake reviews and/or social media bots. **Emergent bias** occurs via interaction with real users. This bias arises due to change in population, cultural values or societal knowledge usually sometime after the completion of the AI design. Lastly, there is **evaluation bias** that emerges during the evaluation of the model where inappropriate or disproportionate standards are applied.

There are also biases that arise from users back to the data. **Historical bias** is already existing bias that can be naturally inherent in the data generation process. **Population bias** arises when statistics, demographics, representatives and user characteristics are different in the user

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

population of the platform from the original target population. **Self-selection bias** is one of the subtypes of the selection or sampling bias, in which the subjects themselves choose to be included. **Social bias** happens when others' actions affect our judgement. **Behavioural bias** results from variability or other differences in user behaviour across platforms, contexts or data sets. While **temporal bias** results from variability or differences in populations and behaviours over time. Finally, **content production bias** is caused by structural, lexical, semantic and syntactic variability of the content created by the users.

Bias and Discrimination

The problems of AI bias are severe because they can result in an unfair, distorted representation of results within the AI system, which, in turn, leads to social inequalities and adversely affects AI trustworthiness and fairness. Bias can also manifest as discrimination, which involves human prejudice and stereotyping based on sensitive attributes such as ethnicity, gender, or age. Discrimination can occur both intentionally and unintentionally, and understanding these two types is important for ensuring fairness in AI.

Explainable discrimination (justifiable inequality) relates to differences in treatment and outcomes among different groups that can be explained or justified via some attributes in some cases (e.g., legal discrimination). Whereas, *unexplainable discrimination* (unjustifiable inequality) is considered illegal discrimination. It occurs when individuals or groups are treated unfairly without any legitimate reason, often due to implicit biases embedded in AI systems.

Types of discrimination:

- *Direct discrimination* happens when protected attributes of individuals explicitly result in non-favourable outcomes toward them.
- *Indirect discrimination* is when individuals appear to be treated justly based on seemingly neutral and non-protected attributes. However, protected groups, or individuals, still get to be treated unjustly because of implicit effects connected to their protected attributes.
- *Systemic discrimination* refers to policies, customs, or behaviours that are a part of the culture or structure of an organisation that may perpetuate discrimination against certain subgroups of the population.
- *Statistical discrimination* is a phenomenon where decision-makers use average group statistics to judge an individual belonging to that group.

A well-known case of AI bias arose in relation to hiring practices. Amazon once built an AI system to automate its recruitment process. The algorithm was trained on resumes submitted to the company over ten years, which predominantly came from males. Consequently, the system began favouring male candidates over females, demonstrating a clear bias (What is Algorithmic Bias?). In the same way, AI systems in criminal justice have been deemed to be

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

racist and have caused discrimination in sentencing and policing practices. These examples underscore the importance of adhering to measures that will help with the early identification of bias within AI.

Fairness in AI

When addressing the concerns related to bias and discrimination, the concept of fairness must be considered. Fairness in AI is a multi-dimensional concept that can be approached from various perspectives.

- *Individual fairness:* give similar predictions to similar individuals.
- *Group fairness:* treat different groups equally.
- *Subgroup fairness* intends to obtain the best properties of the group and individual notions of fairness. It picks a group fairness constraint like equalising false positives and asks whether this constraint holds over a large collection of subgroups.
- *Intersectionality* is a concept that different categorisations including ethnicity, gender and age are interlinked and interdependent leading to formations of discrimination. In AI, this leads to the question of how the different sensitive attributes interact with each other to impact on the fairness of the outcomes.

To implement fairness in AI, several technical methods can be applied at different stages of the AI development process. *Pre-processing* techniques try to transform the data, so the underlying discrimination is removed. *In-processing* techniques try to modify and change learning algorithms to remove discrimination during the model training process. In-processing can be used during the training of a model either by incorporating changes into the objective function or imposing a constraint. *Post-processing* means labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase.

One particularly significant fairness metric is the **Conditional Demographic Disparity (CDD)** proposed in 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI', which provides a more nuanced tool for evaluating fairness by considering the context in which decisions are made. This sophisticated fairness metric assesses how AI systems treat different demographic groups by determining whether varying outcomes can be justified based on specific attributes. Unlike simple demographic disparity, CDD considers an attribute (or a set of attributes) R that provides the context in which the disparity is evaluated as seen in ***Fig. 3***. For instance, in a hiring AI system, simple demographic disparity might show that one group is hired more frequently than another. However, when relevant attributes such as qualifications or years of experience (i.e., the attribute(s) R) are considered, CDD allows for a more accurate assessment of whether the disparity is justified or indicative of unwanted bias.

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

This approach is particularly important in a legal context. Legal decisions often need to be justified based on specific, legitimate factors, and, by incorporating CDD into fairness assessments, AI practices can better align with non-discriminatory and fair legal standards. Such fairness concepts and techniques should be incorporated into the development of Al models so that the developed models will be accurate and fair at the same time. Such approaches can assist in mitigating bias and reduce unnecessary discrimination, thereby fostering greater trust in AI and its use being more appreciated by everyone in society.

$$A_R = \frac{Number\ of\ selected\ people\ in\ advantaged\ group\ \textbf{with attributes R}}{Total\ number\ of\ people\ in\ advantaged\ group\ \textbf{with attributes R}}$$

$$D_R = \frac{Number\ of\ selected\ people\ in\ disadvantaged\ group\ \textbf{with attributes R}}{Total\ number\ of\ people\ in\ disadvantaged\ group\ \textbf{with attributes R}}$$

$$\textbf{Conditional demographic (dis)parity} = D_R > A_R$$

Figure 3 Conditional Demographic (Dis)parity

Regulations for AI Bias Detection

EU AI Act

To overcome the problems associated with AI bias, a number of governments and organisations have established certain legal requirements regarding the fairness and trustworthiness of AI. The European Union (EU) is at the forefront of regulating AI with the proposed EU AI Act which establishes a legal framework for the advancement of AI in the EU. The EU AI Act includes specific provisions for bias detection, requiring that high-risk AI systems undergo rigorous testing and validation before their deployment in the EU marketplace. This entails sets of compliance checks that include mandatory assessment of AI systems' biases and the handling of any identified risks.

Within the AI Act, it specifically mentions:

Section 2(f-g): Bias examination and mitigation: Data sets must be examined for potential biases that could negatively impact health, safety, fundamental rights, or lead to prohibited discrimination, especially when the AI system's outputs influence future inputs. Appropriate measures must be implemented to detect, prevent, and mitigate any biases identified.

Section 5(a-f): Processing special categories of personal data: Special categories of personal data can only be processed if bias detection and correction cannot be effectively achieved using other data, including synthetic or anonymised data. These data must be subject to technical limitations on reuse, and protected by state-of-the-art security and privacy measures, such as pseudonymisation. Strict controls, safeguards, and documentation must be in place to secure the data and prevent misuse, ensuring access is limited to authorised persons with

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

confidentiality obligations. The data must not be transmitted or accessed by unauthorised parties. The data must be deleted once the bias is corrected or when the retention period ends, whichever is sooner. Records must include justifications for why processing special categories of data was necessary for bias detection and correction, and why alternative data could not achieve the same objective.

§ 5-301 Bias audit of New York City's Local Law 144

New York City's Local Law 144 of 2021, effective from July 2023, is a unique attempt to audit Automated Employment Decision Tools (AEDTs) deployed within its jurisdiction. Any employer with a business in NYC who are using any AEDTs in the recruitment or promotion process must adhere to an independent third-party bias audit annually. The main purpose of this law is to detect biases based on gender and/or race/ethnicity.

Specifically mentioned within the NYC Local Law 144:

Annual bias audit: Employers or employment agencies must ensure that a bias audit has been conducted within the past year before using or continuing to use an AEDT.

Bias audit content:

a) *Calculation of selection rate:* The AEDT must calculate the selection rate for specific groups (e.g., gender, race/ethnicity, intersectional categories).
b) *Calculation of impact ratio:* The impact ratio for each selected group must be calculated.
c) *Intersectional analysis:* The impact ratio must be separately calculated for combinations of gender, race/ethnicity, and their intersections.
d) *Handling of missing data:* It must be noted if any applicants belong to categories (e.g., gender, race) that are unknown and thus excluded from the calculations.

Role of independent auditors: Independent auditors may exclude categories representing less than 2% of the data from the bias audit. If a category is excluded, the reason must be clearly explained.

Local Law 144 has set an important precedent for other jurisdictions. It is the first law to mandate independent audits of algorithmic systems used in recruitment, and it underscores the growing recognition of the need for regulatory oversight in the deployment of AI-driven decision-making tools. As other regions consider similar legislation, they may look to New York City's experience to refine and strengthen their own approaches to algorithmic accountability.

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

Standards for AI Bias Detection

International standards can play a crucial role in creating AI-based processes that are free from unwanted bias. These standards help create frameworks for best practice that includes guidelines and the promotion of good practices when the developers and organisations are developing and deploying AI systems that are more accountable and transparent. Currently, there is the published ISO/IEC TR 24027:2021 and the forthcoming ISO/IEC 12791.

ISO/IEC TR 24027:2021: Bias in AI Systems and AI-Aided Decision Making

ISO/IEC TR 24027:2021 is the technical report for the best practices for bias detection and mitigation in AI systems. It also makes clear that most AI systems are prone to the various forms of bias that can distort their fairness and effectiveness because they are based on data and algorithms. It offers clear guidance for identifying bias in all stages of an AI system from data collection to deployment.

Key components of ISO/IEC TR 24027:2021 include:

Understanding Bias and Fairness:

- Bias in the standard refers to bias in AI systems that can be introduced as a result of structural deficiencies in system design, arise from human cognitive bias held by stakeholders or be inherent in the datasets used to train models. These include societal bias and statistical bias, which can influence AI systems in various ways.
- The concept of fairness is explored in the context of AI systems, emphasising that fairness is a complex, context-dependent concept that varies across cultures and application domains.

Sources of Unwanted Bias: The report outlines several types of unwanted bias, such as human cognitive bias, data bias, and bias introduced by engineering decisions. For example, data bias can happen if the data is collected with selective sampling or wrong labelling as well as already developed cognitive biases in humans, such as automation bias, group attribution bias, implicit bias, and confirmation bias, which can influence AI systems.

Bias Assessment Techniques: Techniques used to assess bias in AI systems include confusion matrices, equalised odds, equality of opportunity, demographic parity, predictive equality, and other metrics. These methods assist in quantifying the extent of bias and evaluate the fairness of AI systems.

Bias Mitigation Strategies: In this standard, re-weighting techniques, and adversarial methods are promoted to mitigate bias. Continuous monitoring is also important and is an effective way to validate throughout the AI system's lifecycle to ensure that biases are detected and addressed.

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

ISO/IEC 12791: Bias Mitigation Techniques (Forthcoming Standard)

ISO/IEC 12791 is a forthcoming standard expected to provide mitigation techniques that can be applied throughout the AI system life cycle in order to treat unwanted bias. This document describes how to address unwanted bias in AI systems that use machine learning to conduct classification and regression tasks. This document is applicable to all types and sizes of organisation.

IEEE P7003 Standard for Algorithmic Bias Considerations

This document is a work-in-progress report on the '**IEEE P7003: Standard for Algorithmic Bias Considerations**'. This standard is part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and is being developed to help identify and mitigate unintended, unjustified, and inappropriate biases in algorithmic decision-making systems. The primary objective of the standard is to provide a development framework for individuals or organisations that are developing or deploying algorithmic systems to manage issues of bias. The standard addresses various situations where algorithmic bias might occur and provides methodologies for managing these impacts, thereby contributing to improving the fairness of algorithmic decision-making systems, which are increasingly used by businesses, governments, and other organisations.

Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

This is a document by the National Institute of Standards and Technology (NIST) on how to detect and mitigate bias in AI systems. This document discusses three sources of bias that might exist within AI systems, known as systemic bias, statistical/computational bias, and human bias, as well as the possibility of harm that these biases can cause. This document also presents several strategies to address bias, including dataset, testing and evaluation, and human aspects, and provides initial guidelines on how to implement these strategies. The purpose of the document is to shed light on the multifaceted issue of AI bias and foster public trust regarding the challenges of AI by adopting a socio-technical approach. The document aims to improve the reliability of AI technologies as well as further the conversation on the standardisation process with the help of various stakeholders.

The AI industry should adopt such standards for addressing bias. Therefore, organisations can prove they are being 'fairer' and 'more transparent' in their application of AI. Furthermore, it helps make AI systems more compatible for different countries and different industries thereby uniting the world in the development of new technology.

Best Practice (Sandboxing and Red Teaming)

With sandboxing (which is written into the EU AI Act), decision-making by AI systems is subjected to testing in a deliberate manner to help expose any risks in the system such as

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

determining the existence of unwanted bias. This approach helps regulators oversee the AI systems and ensure their output is fair and transparent as required.

The 2024 Generative AI Red Teaming Transparency Report showcases a recent example of red teaming, whereby specialists deliberately attempt to identify and exploit weaknesses, limitations, and vulnerabilities in AI systems, with the goal of making these models stronger. In a regulatory sandbox, red teaming enables the representation of real-life situations where AI can facilitate discriminatory and unfair results. This method is vital for detecting systemic faults that may not be normally revealed in traditional testing approaches. During the DEF CON GRT (Generative AI Red Teaming) challenge, various AI models were subjected to tests designed to find biased or harmful outputs. The event highlighted how public red teaming, a form of structured public feedback, can reveal contextual biases that might be overlooked in normal testing environments.

Combining the sandbox environment with red teaming provides the regulator with an insight into how AI systems operate under designated environments, thereby enhancing the EU AI Act regulation. By incorporating these methods into the regulatory sandbox, regulators can better understand the existing limitations and challenges of using AI systems, as well as making sound decisions regarding its deployment. It is imperative to avoid the use of AI systems that may perpetuate or even exacerbate biases in order to protect public trust in such technologies.

Conclusion

As AI technologies continue to evolve and play an increasingly significant role in society, the problem of bias detection becomes even more important. The potential for AI systems to perpetuate or exacerbate existing biases underscores the necessity for robust regulatory frameworks and international standards. These tools are necessary for ensuring that AI systems are developed and deployed in a manner that is fair, transparent, and accountable.

This paper has discussed various types of bias and discrimination which can exist in AI systems and introduced essential ideas and methodologies. Additionally, one of the more sophisticated fairness metrics, i.e., Conditional Demographic Disparity (CDD) was introduced as it helps gauge how fairly an AI system works in practice. In addition, this paper outlined some existing regulations including the EU AI Act and the New York City's Local Law 144, as well as international standards such as ISO/IEC TR 24027:2021, IEEE P7003, and others that are widely used in the development of a fair and transparent AI system. These regulations and standards help in directing AI developers and organisations towards designing more accountable and transparent AI systems and thus make AI systems more reliable as well as improving the fairness of the society where the AI system is deployed.

Moving forward, the continuous development of regulations and standards will be necessary to ensure that AI systems remain fair and trustworthy. Bias detection and mitigation strategies must be effectively integrated into the AI development process from the outset. By prioritising

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie

these ethical considerations, AI can evolve into a technology that benefits everyone, fostering a more equitable and just society.

UCD **Geary Institute** for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland. Phone +353 1 716 4615
Website: https://www.ucd.ie/geary Email: geary@ucd.ie