WHITE PAPER

BIAS IN AI

What regulation requires and how enterprises can put it in action

Contributors:
Claire Poirson, Blanche de Saint Victor,
Nelia Silva, Roxana Rugina, Émilie Sirvent-Hien,
Christèle Tarnec, Kirsten Thorsén

IMPACT AI

cercle InterL

This White Paper is produced by Impact AI and Cercle InterL with the assistance of FIRSH

Bias in AI: What regulation requires and how enterprises can put it in action

In recent years, artificial intelligence (AI) has revolutionized industries by driving efficiency, fostering innovation, and spurring economic growth. However, this powerful technology also brings significant risks, with one of the most pressing being bias in AI systems. This white paper aims to explore the regulatory framework of the European Union's AI Act, which seeks to address and mitigate bias in AI systems.

A White Paper produced by Impact AI
and Cercle InterL
With the assistance of FIRSH

Contributors to this White Paper: Claire Poirson, Blanche de Saint Victor, Nelia Silva, Roxana Rugina, Emilie Sirvent-Hien, Christèle Tarnec, Kirsten Thorsén.

About Impact AI

The Impact AI association is a think & do tank promoting the adoption of responsible AI since 2018. As a reference player in the responsible use of AI in Europe, the organization provides tools and guidance to more than 80 members, including large companies, IT services companies, consulting firms, AI players, start-ups, NGOs and schools. *For more information: www.impact-ai.fr*

About Cercle InterL

Cercle InterL has been committed for 22 years to promoting diversity and professional equality in the scientific and technological sectors, with the ambition of creating favorable conditions for gender balance and performance. It brings together the networks of 15 industrial and technological companies, whose members mobilize throughout the year in working and reflection groups to promote women's access to positions of responsibility, defend the balance between professional and personal life, and share best practices within the network. *For more information: www.InterL.com*

About FIRSH

FIRSH is a law firm dedicated to creation and innovation. As a mission-driven “entreprise à mission” company, the firm runs a laboratory dedicated to combining law and innovation, with a particular focus on tech and AI issues with an impact on today's and tomorrow's society. FIRSH assisting its client in their AI compliance has intervened here in its capacity of legal expert.

Introduction

In recent years, artificial intelligence (AI) has revolutionized industries by driving efficiency, fostering innovation, and spurring economic growth. However, this powerful technology also brings significant risks, with one of the most pressing being bias in AI systems. Biased data, models, and outcomes can lead to unfair, discriminatory, or harmful consequences, undermining individual rights and eroding societal trust in AI. This white paper aims to explore the regulatory framework of the European Union's AI Act, which seeks, among other things, to address and mitigate bias in AI systems. We will examine the regulation's requirements for enterprises, particularly in identifying, reducing, and managing AI bias. Additionally, we will delve into practical approaches that companies can adopt to comply with the regulation, as well as the challenges they may face from legal, technical and operational perspectives.

Bias and the AI Act

Bias in AI arises when AI systems, often trained on historical data, reinforce or perpetuate existing prejudices or inequalities. These biases can stem from a variety of sources, including unrepresentative datasets, unappropriate models, or the lack of diverse perspectives in the design process. When left unchecked, AI bias can lead to negative societal impacts such as discrimination based on race, gender, or socio-economic status, as well as erosion of public trust in AI technologies leading to allocation or representational harms.

As AI becomes more embedded in decision-making processes—whether in hiring, healthcare, law enforcement, or lending—the urgency of addressing bias grows. The regulatory environment, spearheaded by the European Union, seeks to offer a structured solution/requirements through the AI Act, which defines specific rules and responsibilities for organizations developing and deploying AI systems.

The European Union's AI Act is a pioneering regulatory framework designed to ensure the ethical and responsible use of AI technologies and address and mitigate bias in AI systems. The AI Act mandates that enterprises identify, reduce and manage bias throughout the AI lifecycle. This includes ensuring that training data is representative and free from discriminatory patterns, implementing robust models that can detect and mitigate bias and continuously monitoring AI systems for biased outcomes. By adhering to these requirements, companies can enhance the fairness and transparency of their AI systems, thereby fostering greater trust and acceptance among users.

Potential Benefits of Properly Managed AI Bias:

Properly managing bias in AI systems enhances fairness and equity, ensuring decisions do not disproportionately disadvantage any group. This fosters inclusivity and social justice. Additionally, it improves the accuracy and reliability of AI systems, leading to better decision-making and outcomes, which boosts user trust and acceptance. Organizations that address AI bias proactively gain a competitive advantage by demonstrating their commitment to ethical practices and regulatory compliance, enhancing their reputation and attracting ethically minded employees and customers.

Impact AI & Cercle InterL are leading by example

Several companies at Impact AI and Cercle InterL are setting an industry example. Members from various sectors have adopted a hands-on approach to implementing the AI Act. Through real-world case studies, we highlight how companies are adapting governance models, updating AI development practices, and employing new techniques to minimize bias. Sharing these experiences provides valuable insights into complying with the AI Act and contributing to a more equitable AI landscape.

This white paper reflects the collaborative efforts of Impact AI and Cercle InterL, bringing together companies, startups, academic institutions and industry experts to foster AI innovation while ensuring its ethical use, particularly in addressing bias.

1/ The issue of bias with AI and their difficult definition

Various biases in AI can affect different activities and lead to discrimination, here are some examples:

→1. Gender bias in recruitment:

a. Amazon developed a recruitment tool in 2014 that rejected female applicants due to an unbalanced recruitment history. This program would rate resumes on a score going from 1 to 5. Three years later, Amazon had to drop this program after the discovery of a major flaw: gender discrimination. Indeed, this program was trained on resumes received by the group over a ten-year period, most of which were those of men, reflecting the male predominance in tech. Therefore, the AI program got to the conclusion that men's resumes were better than women's.

b. LinkedIn's job-matching AI was biased against women job

c. More recently, a study found that job-ads generated by GPT-4 were on average 30% more biased than those written by humans.

→2. Gender bias in healthcare: for example, an AI system trained primarily on male data may misdiagnose conditions in women, particularly in areas like heart disease

→3. Racial & socio-economic bias in healthcare: for example, a system predicting medical care need favored white over black patients because health costs were used as a proxy for health needs while less money was spent by black patients who have the same level of need as white patients

→4. Gender bias in Apple's credit card.

In 2019 Apple's credit card in partnership with Goldman Sachs Bank has been investigated after being accused of gender discrimination. Consumers had then reported that the Apple's credit card was making gender dis- crimination on the basis that men had more chances of being granted a higher loan than women. In 2021, The New York Finance Service Regulators finally concluded by saying that there wasn't enough proof of gender discrimination.

→5. Multiple biases in a performance monitoring tool in the food delivery industry: In its ruling of November 17, 2023, the Court of Palermo penalized the food delivery company Foodinho for using a discriminatory algorithm. Foodinho had set up a "score of excellence" system for its "riders", based on criteria of "contributions" and "hours of high demand". This algorithm discriminated against less productive riders, notably because of their disability or age. In addition, some riders may be at a disadvantage because their religion does not always allow them to work during "high demand hours" (e.g. Muslims on Fridays, Jews on Friday evenings and Saturdays, and Christians on Sundays).

→6. Identity bias in facial recognition services for security services or social media, for example, facial recognition systems based on gender classification have led to serious discrimination, such as transgender people misidentification and ban to connect to a social media dedicated to women.

HOW TO DEFINE BIAS, FAIRNESS AND DISCRIMINATION

Bias can be defined as a **deviation from the norm** and in the field of AI four broad families of norms have been identified, leading to four **categories of algorithmic biases**: statistical bias (e.g. taking an average that simplifies a phenomena), methodological bias (e.g. using a device that is not well calibrated or using GT3 that has been trained on data collected in 2020 to explore a topic that has emerged latter), cognitive bias (e.g. making a subjective and irrational decision) and socio-historical bias (e.g. training an LLM on a dataset collected in a single country while using it in a country with different cultural values). Bias has been identified **all along the AI pipeline** (in data, when designing the model or through user's interactions with the AI-based-system[1]) and can be source of allocation or representational harms on people but also of economical, reputational and legal harm for companies as it can lead to unfair decisions.

When using AI-based models to make decisions, these are based on the predictions of statistical and machine learning models that are supposed to replace the subjective human decision-making by an objective one. Indeed, to simplify complexity, humans tend to generalize and there are more than **200 cognitive biases** to which humans can be unconsciously subjected. However, AI-based models reveal themselves to amplify human cognitive biases and to embed in addition statistical biases, which are source of unfairness.

Even though fairness is an incredibly desirable quality in society, it can be difficult to achieve in practice as **there is no universal definition of fairness.** Broadly, **fairness in AI** is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making (Mehrabi et al., 2022). **Prejudice** is defined as damage/harm caused to another person, whether intentionally or unintentionally. It is the deliberate or inadvertent alteration of an asset or right belonging to the person complaining, resulting in a loss of value or opportunity[2] . Unintentional prejudices can conduct to **unintentional bias** and discrimination against the civil society, but companies could also face serious issues: legal, reputation[3] , business[4], employees' motivation and well-being, ESG risk, company value, talent recruitment.

In the field of algorithmic bias management, **generative AI** such as GPT4 presents a particular challenge for companies for several reasons. LLMs, which are used in various fields ranging from virtual assistance to content generation, are increasingly integrated into large-scale applications such as search engines or office suites. An LM (Language Model) is a **statistical model designed to repre- sent natural language**. LLMs are advanced versions of these models, trained on vast data sets and using sophisticated architectures. **Their ability to understand and generate text in a coherent and contextually relevant way is revolutionizing applications, improving the performance of** machine translation, text generation, sentiment analysis and human-machine interaction systems. In LLMs, the huge scale of the pre-training data sets, the adaptation process (aligning with human values, specializing in a particular language or field etc.), the bias mitigation choice and the nature of the prompt (e.g. what kind of roleplay) can cause harms of allocation (unjust distribution of resources) or of representation (reinforcement of stereotypes).

HOW TO DEFINE BIAS, FAIRNESS AND DISCRIMINATION

Whereas harms of allocation produce immediate, easy-to-formalize effects, harms of representation produce long-term effects and are more difficult to formalize. The development pipeline of generative AI is built mainly from **two blocks: a basic model**, developed to encode the language and its rules, **and a second model**, **which is fine-tuned to respond to specific instructions** (e.g. open questions/answers). This second model can then be further tailored to the desired task (e.g. customer relations chatbot) and/or aligned with stated values (e.g. adherence to a company's ethical charter). The biases in LLMs can therefore manifest both in the mo- del itself **(intrinsic bias, which occurs in the model's internal representations)** and in the final decisions it takes **(extrinsic bias, which occurs in final decisions and predictions). Bias in LLMs can be evaluated intrinsically and extrinsically.** Understanding how, where and when biases arise is necessary to draft a data and AI go- vernance process for all companies.

There is no standard definition of fairness, and **over 70 different fairness metrics** have been created. These metrics can be grouped into **three main families**, each corresponding to different notions of fairness that are **not mathematically or morally compatible**. You can focus on equality in terms of **acceptance rate, error rate or calibration** of the model between privileged and unprivileged groups. The choice for the most relevant fairness metric must be made on a case-to-case basis and based on the company/provider's worldview. For example, to select the best candidate for a job without prejudice to an unprivileged social group, one can focus on an equality between the privileged and the unprivileged group in terms of acceptance rate for the job or error rate (equal number of individuals whose request for the job is denied among those who would have made the job) or calibration rate (equal number of individuals who won't make the job properly among those who are given the job).

Fairness in AI is a workflow of identifying bias (the disparate outcomes of two or more groups), performing root cause analysis to determine whether disparities are justified and employing a targeted mitigation strategy if needed. Managing **LLM** bias remains a complex and evolving subject and although there are ways to evaluate (intrinsic and extrinsic methods, existing or new benchmark) and mitigate bias (with or without additional training), the area is not yet fully mature. Within companies, the focus is currently on organization, developing prototypes and experimenting.

To **mitigate bias in AI**, one can use **pre-processing, in-processing or post-processing strategies**. Pre-processing strategies, that focus on changing training data to have fairer data as input of the model and by consequences fairer output as a result, implies having access to training data. In-processing methods, which consist in putting constraints while building the model to improve model fairness, requires access to the model and data scientist expertise. Post-processing methods, applied on the output of the model, means no acting on training data or on the model itself. Vigilance and compromise are sometimes necessary to ensure that bias-mitigation actions do not harm the model's performance.

2/ How does AI Act regulate bias?

AI ACT IN GENERAL: INTRODUCTION AND HISTORY

The AI ACT which is the European Regulation that establishes **harmonized rules** regarding AI in the European Union entered into force on the 1st of August 2024. It will be of general application on the 2nd of August 2026, even though some of its provisions will be progressively of general application before that date (in the course of 2025) or after (2027 at the latest, regarding the classification of the high-risk AI systems and the subsequential obligations.

The AI Act is a historical step in the legal and innovative world. Suggested by the Commission in 2021 and approved by the Parliament and the Council in December 2023, the **AI Act aims at catalyzing a responsible and respectful AI across the EU**.

The approval of this regulation by the different States has been at the center of vivid debates. Deemed to be too regulatory by some of the most "pro-innovation" countries other countries, on the contrary, called for more restraining measures in an attempt to regulate this new market. We finally got to an agreement by making multiple compromises. For Thierry Breton, former European commissioner, this text is the first in the world to recognize "the perfect **balance between innovation and safety**".

The AI Act addresses a large panel of potential risks involved in the use of AI, notably for health, for safety and for the citizen's fundamental rights. The Regulation works around a "risk-based approach", consisting in four categories from the least to the most acceptable risk. Hence, the AI Act implements clear obligations for developers depending on the category of risk in which their system belongs to.

One of the EU's main priorities with the AI Act was to implement **transparency requirements**. The Regulation mentions "transparency" as one of the seven non- binding ethical principles for AI which are intended to help ensure that AI is trustworthy and ethically sound. These guidelines were elaborated in 2019 by a high-level expert group of the Commission AI-HLEG in order to achieve an AI that can be trustworthy[5] , under which, the non-binding principles of "diversity, non-discrimination and equity" have been established. The AI Act defines the concept of "transparency" in its preamble at n°27 as follows : "Transparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights."

Furthermore, the AI which falls into the category of "moderate risk" must comply with specific requirements of transparency and information. For instance, this is the case for generative AI systems such as chatbots. They must make it clear to the users that they are interacting with a machine. In addition, for high-risk AI systems, an appropriate type and degree of transparency shall be ensured.

Another important non-binding principle from the Ethics guidelines for trustworthy AI is the concept of "**accountability**". They define the concept as follows: "Accountability": Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes, plays a key role therein, especially in critical applications. Moreover, adequate and accessible redress should be ensured. They promote the minimization and reporting of negative impact, trade-offs and redress."

How does the AI ACT define the notion of bias, discrimination and equity?

Although the AI Act regulation does not define the concepts of bias, discrimination and equity per se, it still mentions in its preamble the non-binding guidelines to follow in terms of ethics. It refers notably to the fact that AI systems need to be developed and used in a way that **avoids any discriminatory effects and unfair biases, that are forbidden by EU law or even national law**[6].

What are the legal obligations regarding bias management in the AI Act?

The binding rules that are introduced by the AI Act are based on the intensity and the scope of the risks that can be caused by AI systems.

As mentioned below, the AI Act works on a “risk-based approach” articulated around four categories of risks. The first one being “unacceptable risks”, the AI Act forbids the utilization of AI systems that are deemed too dangerous or unethical to be allowed (for instance: social scoring and biometric identification). These types of AI can lead to discriminatory results and exclude certain groups of people.

The second category is “high risks” for AI that can have significant implications for individual rights and safety. High-risk AI systems use techniques that involve training AI models by means of using data developed on the basis of training, validation and test datasets. Their utilization is allowed only if strict obligations of procedure and branding are respected.

The third category is “moderate risks” where some utilizations are allowed only if the obligations of information and transparency are respected (for instance: chatbots).

Finally, the last category is “minimal risk to no risk” their utilization is allowed almost without any restriction.

We will focus on high-risk AI systems as they are the most likely (without considering the unacceptable AI systems) to have biases, cause discriminatory effects and have significant implications on individual’s rights. They are particularly concerned by the legal measures implemented by the AI Act.

High-risk AI systems are subject to controls and governance practices[7]. In particular, they must undergo some investigation for possible biases that could harm the health and safety of individuals, have a negative impact on fundamental rights, or result in discrimination prohibited by Union law, especially when output data influences inputs for future operations.

In addition, these high-risk AI systems are subject to an obligation of transparency. They must include maximum transparency through information sheets and explanatory notices. Indeed, it seems essential that these AI systems are developed and used in such a way as to enable appropriate traceability and explicability[8] . In that respect, on the one hand, data subjects can realize that they are interacting with an AI system, they are informed of the AI's potential, its limits; and on the other hand, they will know their rights (especially when personal data is involved.)

Furthermore, like with the General Data Protection Regulation (or GDPR), certain operators will have to undergo an impact analysis prior to the marketing of these AI systems in the case where personal data is at stake and/or there is a potential risk of infringement of people's fundamental rights, such as discrimination[9]. This will be the case for the operation of a real-time remote biometric identification system in areas accessible to the public[10], as well as for AI systems deployed by public law bodies, private entities providing public services, banking or insurance entities[11].

The AI Act provides for a specific penalty for AI systems with unacceptable risks (higher penalty). The same penalties apply to high and moderate risk systems. Finally, it provides for a sanction, applicable to all levels, specific to the provision of erroneous information to national authorities.

	Unacceptable risks	**High risks**	**Moderate risks**	**Minimal to no risk**
Obligations	Forbidden (art. 5)	Obligation to comply to certification and branding (art.43 to 49) Obligation to undergo controls and Governance practices (art. 10) Obligation of transparency (art.13)	Obligation of information and transparency (art. 50)	No obligations but code of conduite possible

	Unacceptable risks	High risks	Moderate risks	Minimal to no risk
Sanctions (art.99)	35Millons Euros Fine or 7% world turnover + If supply of incorrect information: 7,5 millions Euros Fine or 1% world turnover	15Millions Euros Fine Or 3% world turnover + If supply of incorrect information: 7,5 millions Euros Fine or 1% world turnover	15Millions Euros Fine Or 3% world turnover + If supply of incorrect information: 7,5 millions Euros Fine or 1% world turnover	If supply of incorrect information: 7,5 millions Euros Fine or 1% world turnover

What are the sanctions in the AI Act?

The aim of these legal obligations is to meet the major challenge of respecting the fundamental rights and freedoms of individuals.

Thus, depending on the breach of any of the obligations placed on AI system operators, the European Commission **could impose an administrative fine of up to 35,000,000 Euros** or, if the offender is a company, up to 7% of its total worldwide annual sales in the previous financial year, whichever one is the highest[12] .

National French provisions governing discrimination

The Criminal code: Articles 225-1 to 225-4 of the French Penal Code penalize discrimination committed by private individuals and legal entities. Discrimination committed by public officials is punishable under article 432-7 of the Penal Code.

The Labor code: Articles L1131-1 to L1134-10 of the French Labor Code sets out provisions concerning discrimination in the workplace.

The AI Office

The AI Office is established since May 2024 and exists within the EU Commission. It plays a key role in implementing the AI Act throughout the EU. It enforces the rules for general-purpose AI models and plays the role of a promoter for an innovative ecosystem of a trustworthy AI. Furthermore, the AI Office also collaborates with Member States in order to combine knowledge from diverse expert groups such as the scientific community, industry, open-source system.

AI Act delegates how to implement in standards

Section 40 of the AI Act stipulates that compliance with legal requirements will be presumed by adherence to harmonized standards which will be developed. A request for standardization is prepared by the commission and sent to CEN-CENELEC.

This means that manufacturers who follow these standards will not have to interpret the essential requirements of the AI Act themselves. The limitation is that these documents (ISO/IEC documents dealing with AI standardization for example ISO/IEC TR 24027) show a tendency to provide frameworks and technical tools for assessing and mitigating AI risks, while avoiding defining specific normative thresholds.

They focus on documentation, testing and management processes, often leaving specific ethical decisions to end-users like companies that deploy AI and their stakeholders.

3/ How to implement in practices?

We have seen that bias can emerge from data representing society and from choices and usages from humans around the AI system. It is nearly impossible to avoid bias but the objective for deployers of such technology is to assess the risk, implement trustworthy system and keep the control.

Companies develop a practical approach to comply with the regulation, as well as the challenges they may face from legal, technical, and operational perspectives (define fairness criteria, choose thresholds for example, working with existing teams).

They have not waited for the final text of the regulation to tackle the bias and fairness subject. For 20 years, the Cercle InterL has been committed to gender diversity and professional equity in the scientific and technological sectors, with the ambition to create favorable conditions, gender balance and performance.

The Women& AI Pledge for an accountable and gender fair AI is the collective work of the Cercle InterL companies. This Pledge is a valuable and actionable asset for any company willing to take on this challenge.

It is based on 7 fundamental principles that enable companies to address the risks of discriminatory cognitive bias, during development or while using AI-based solutions or devices and here are some extracts.

AI Ethics Committee and Governance

The company should implement a governance process that ensures an operational action plan and direct access to company Executives for AI and diversity questions.

More specifically, a multidisciplinary AI Ethics Committee, that reports directly to the Executive Management, and where the scope of responsibility extends to the whole company, ensures the detection of gender bias, and rapid corrective action in all AI systems used or produced. The mission of the AI Ethics Committee is explicitly specified, documented, and distributed across the company, specifically the handling of gender bias. The committee has the means to perform their assigned mission efficiently.

Compliance by Design

From the design of an AI product or project to the moment it is delivered into production, the teams involved make sure that the principles of non-discrimination of gender are integrated at each stage. And Involved stakeholders are represented in the process.

Data Selection and Processing

Being aware of gender bias in an AI project starts with data selection. A team of industry experts and developers analyses and classifies data and identifies any imbalance that exists in the data set. If necessary, biased data are corrected or deleted.

The ethics of models

A company that develops models for AI projects should use a tool that can detect gender bias. The results are reviewed by an in-house champion and/or the AI Ethics Committee. A document that guarantees the transparency, traceability, and the explainability of the models used and the results produced, should be kept up to date. Companies that are users of AI-based systems should engage with suppliers who develop AI solutions that do not propagate gender bias in their solutions. A good practice is to perform regular audits on models to identify and mitigate bias.

Evaluation and Monitoring

The company should establish a procedure for reporting and correcting deviations and possible discriminations perpetuated by AI solutions throughout their lifecycles (collecting data, cleaning data, training Machine Learning models, and so on). This procedure has to be compliant with the regulatory context. The AI Committee supervises the detection and correction of deviations and discriminations, handled by AI development team.

AI team diversity

The company should set global or team-specific objectives that will result in the diversification of profiles in AI teams. In line with global policy, gender diversity should be an objective in the recruitment, retention and promotion of employees. AI teams seek to diversify talent by hiring employees possessing hard and soft skills, integrating different profiles in teams, making them more mixed and thus contributing to the elimination of gender discrimination. To increase the proportion of women in AI teams, the company agrees to develop professional training programs. These programs focus on training or re-training women for jobs in AI.

Awareness and Accountability

The company agrees to raise awareness among employees about issues related to gender bias, especially among those who work in the field of AI. For example, the company can raise awareness about gender bias by developing a range of audience-specific content, from simple communications to training adapted to employee roles. The company is encouraged to raise awareness within its wider ecosystem, including schools, universities, and the public.

Recommendations

Finally, we propose some recommendations for companies and public authorities:

For companies a 5-step action plan must be deployed

The first and most important step is to **be aware** of bias in AI. To achieve this, companies must acknowledge the existence of bias and develop their own values around what they consider to be bias. The second step is to create an **inventory** of existing AI systems within the company and assess the level of risk associated with each AI-based service in relation to unfairness and discrimination. Then, once companies have a clear and complete bias and discrimination risk-based mapping of their AI systems, they must **prioritize risks with legal impact for the company**. After that, a **tracking and documentation process** must be implemented for all services with legal risks. Finally, companies must integrate bias management into their Corporate and Social Responsibility (CSR) approach.

Recommendations for public authorities

As for companies, the first action for political bodies is to **raise political awareness** of the challenges of harm and allocation prejudice caused by biased AI-based services. In addition to raising awareness within the various political bodies, it is necessary to **inform and educate** the population, starting with the youngest, who are the citizens of tomorrow, by developing bias management training in schools at all levels. In addition, public authorities should **allocate funds to research** activities on bias management and encourage cross-disciplinary research (computer science and social sciences). In order to **support small and medium-sized enterprises** (SMEs), public authorities must encourage the development of sandboxes for training and testing of models and promote open access to tools and audit framework. Openness is key to innovation in such a dynamic environment and open standardization should be supported, involving civil society and stakeholders. Finally, public authorities should involve stakeholders in consultation and regulatory development.

Conclusion

NAVIGATING BIAS IN AI: A CALL FOR COLLABORATION AND GLOBAL AI GOVERNANCE

The European Union's AI Act sets a critical regulatory foundation to address bias in AI systems, yet it refrains from defining biases and prescribing solutions.

Companies are responsible of designing and implementing strategies to meet compliance requirements while mitigating bias risks.

As AI and regulatory landscapes change, businesses must proactively adapt their practices to lead in ethical AI development.

Many organizations, including members of Impact AI and Cercle InterL, have proactively embraced this challenge, integrating ethical principles, governance frameworks and innovative practices into their operations.

As we have seen earlier, generative AI brings new challenges, increasing risks of unfair outcomes and stereotypes. Tackling these issues requires collaboration between businesses, policymakers, and society.

Companies must ensure that fairness and transparency are embedded in their AI systems to be compliant.

Beyond this, AI Pact represents a proactive step towards responsible AI encouraging companies to set up an action plan before harmonized standards are published.

Public authorities should support this effort through education, partnerships, and the creation of tools and standards for responsible AI use.

Bias is not a problem to eliminate but a dynamic risk to manage. The evolving lands- cape of AI calls for constant vigilance, robust governance and the willingness to adapt.

Shared learning and collaboration across industries can build a fair AI ecosystem, balancing innovation and societal responsibility.

To know more:
oecd.ai/en/catalogue/tools

Annex

In general, one tends to perceive that as a bad thing (and we will get there in this paper), but we should look at some examples that can show us the other side of the bias.

If a Company has 75% of its workforce as a male gender and the Company wants to come close to the parity of gender, then, the algorithm should be biased towards women, so that the Company can achieve a better blend in its workforce.

Different taxonomies and bias definitions

Bias and norms: hal.science

1. Data Bias

- **Historical Bias:** Bias that arises from historical data that reflects past discriminatory practices or societal inequities.
- **Representation Bias:** Bias that occurs when the data used to train AI models does not adequately represent the diversity of the population it aims to serve.
- **Measurement Bias:** Bias that results from the way data is collected or measured, such as using biased survey questions or measurement tools.
- **Aggregation Bias:** Bias that arises when data is aggregated in a way that obscures important differences between subgroups.

2. Algorithmic Bias

- **Prejudice Bias**: Bias that occurs when the algorithm itself is designed in a way that favors certain outcomes or groups over others.
- **Evaluation Bias:** Bias that arises from the way the algorithm's performance is evaluated, such as using biased metrics or benchmarks.
- **Population Bias:** Bias that occurs when the algorithm is trained on data from one population but applied to another, leading to inaccuracies.

3. Interaction Bias

- **Feedback Loop Bias:** Bias that arises when the algorithm's outputs influence the data it receives, creating a self-reinforcing loop that amplifies existing biases.
- **Behavioral Bias:** Bias that occurs when the algorithm's outputs influence human behavior in a way that perpetuates or exacerbates biases.
- **Content Production Bias:** Bias that arises from the way content is generated or curated, such as in recommendation systems or social media platforms.

4. Human Bias

- **Implicit Bias:** Unconscious biases held by individuals involved in the development, deployment, or use of AI systems.
- **Confirmation Bias:** Bias that occurs when individuals selectively interpret or seek out information that confirms their preexisting beliefs or expectations.
- **Attribution Error Bias:** Bias that arises from the way individuals attribute the causes of events or outcomes, often leading to stereotyping or prejudice.

5. Systemic Bias

- **Institutional Bias:** Bias that is embedded in the policies, practices, and structures of organizations or institutions.
- **Cultural Bias:** Bias that reflects the values, norms, and beliefs of a particular culture or society.
- **Structural Bias:** Bias that arises from the way systems and processes are designed and implemented, often leading to disparate impacts on different groups.

Annex

6. Evaluation Bias

- **Benchmark Bias:** Bias that occurs when the criteria or benchmarks used to evaluate AI systems are themselves biased.
- **Validation Bias:** Bias that arises from the way AI systems are validated, such as using biased test data or evaluation methods.
- **Performance Bias:** Bias that occurs when the performance of AI systems is measured in a way that favors certain outcomes or groups over others.

7. Interpretation Bias

- **Explanation Bias:** Bias that arises from the way AI systems' decisions or outputs are explained or interpreted.
- **Communication Bias:** Bias that occurs when the way AI systems' outputs are communicated influences how they are perceived or acted upon.
- **Presentation Bias:** Bias that arises from the way AI systems' outputs are presented, such as using biased visualizations or language.

Understanding these different types of bias is crucial for developing effective strategies to mitigate them and ensure that AI systems are fair, transparent and accountable.

[1] Imagine an AI model that puts specific CVs at the top of its list. HR recruiters tend to interact most with the top results and pay little attention to other CVs. CVs at the top will become more and more popular, not because of the nature of the result but due to the biased interaction and placement of results by these algorithms

[2] dictionnaire-juridique.com/definition/prejudice.php

[3] https://www.finance-investissement.com/nouvelles/developpement-des-affaires/lia-risque-dentacher-la-reputation-de-lindustrie/

[4] In 2022, a survey conducted by DataRobot reported that over 350 U.S and U.K. technologists suffered losses from algorithmic biases. More than half lost revenue, customers, employees or incurred legal fees.

[5] Trustworthy AI" is designed to be fair, ethical and transparent. In contrast to "AI" (simple AI, that is untrustworthy), "Trustworthy AI" protects privacy, reduces biases, and ensures human supervision to maintain security. Understanding the necessity to have a "Trustworthy AI" allows us to build a safeguard against harmful and discriminatory AI systems.

[6] Cons. 27 – AI Act. For example, French labor law punishes any discrimination based on a person's gender.

[7] Art. 10 AI ACT

[8] Art. 13 AI ACT

[9] Cons. 27 AI ACT

[10] Cons. 34 AI ACT

[11] Cons. 96 AI ACT

[12] Art. 99 AI ACT

IMPACT AI

cercle InterL

www.impact-ai.fr
contact@impact-ai.fr

www.interelles.com
sylvie.lechevin@sap.com