APEC
Asia-Pacific
Economic Cooperation

Advancing Free Trade
for Asia-Pacific **Prosperity**

Best Practices to Detect and Avoid Harmful Biases in Artificial Intelligence Systems

APEC Digital Economy Steering Group

September 2023

Asia-Pacific
Economic Cooperation

Best Practices to Detect and Avoid Harmful Biases in Artificial Intelligence Systems

APEC Digital Economy Steering Group

September 2023

APEC Project: DESG 05 2021A

Produced by

National Artificial Intelligence Research Center (CENIA Chile)
&
Foresight Consultancy
Avenida Vicuña Mackenna 4860, Macul
Chile 7820436
Tel: (+56) 9 77680087
email: rodrigo.duran@cenia.cl
Website: www.cenia.cl

Project Overseers
Future and Social Adoption of Technology office (FAST) of Ministry of Economy of Chile
Ministry of Science, Technology, Knowledge and Innovation of Chile

For
Asia-Pacific Economic Cooperation Secretariat
35 Heng Mui Keng Terrace
Singapore 119616
Tel: (65) 68919 600
Fax: (65) 68919 690
Email: info@apec.org
Website: www.apec.org

APEC#223-CT-01.16

Executive summary

Artificial Intelligence (AI) is a widely used general-purpose technology that permeates various aspects of everyday life, encompassing diverse fields of human activity. As AI systems operate within complex socio-technical contexts, they are susceptible to replicating and amplifying existing biases shaped by human influence. Consequently, economies around the world have adopted diverse strategies, both institutional and technical, to address biases arising from the widespread use of AI systems.

The recommendations presented in this report are part of the APEC project DESG 05 2021A, titled "Comparative Study on Best Practices to Detect and Avoid Harmful Biases in Artificial Intelligence Systems." These recommendations synthesize the main findings derived from both on-desk research and the workshop held in May 2023 with experts from various APEC economies. Some significant findings from the study include:

1) APEC economies have developed different strategies to detect, avoid, and mitigate harmful biases in AI systems, with the most successful ones establishing robust institutional frameworks and actively implementing measures to address biases.
2) In terms of institutional strategies, common measures to reinforce trust in AI ecosystems include the creation of domestic policies, the adoption of ethical frameworks and principles, and the establishment of guidelines and pilot projects.
3) The experts highlighted various biases that can emerge at different stages of the AI system lifecycle, particularly those biases influenced by human intervention. As a result, there is a strong focus on addressing biases in the problem formulation, data collection, and feedback stages of AI systems.

Consequently, based on successful case studies, recommendations have been established in two fields: institutional and technical. Some institutional best practices include the creation and adoption of domestic frameworks, the incorporation of participatory instances in the early phases of AI system development, adopting a multistakeholder approach, and integrating progressive regulatory mechanisms, among others. On the other hand, some technical best practices encompass problem-oriented AI solutions, establishing diverse and multidisciplinary groups in earlier stages, and opening datasets for inspection and auditing errors during the pre-processing and labeling stage, among others.

Furthermore, the experts have identified several dilemmas that need to be addressed in the future development of AI systems, particularly in sensitive fields crucial to the successful deployment of fair AI systems. These key challenges encompass topics such as the protection of privacy and data availability necessary for training algorithms, the application of general global principles to local contexts, and the interaction between public and private regulatory frameworks, with regulatory experimentation as a mechanism to foster convergence.

These findings will be included along with the detailed on-desk research and interviews process in the document titled "Comparative Study on Best Practices to Detect and Avoid Harmful Biases in Artificial Intelligence Systems." This project is an initiative of the APEC Secretariat and the Digital Economy Steering Group (DESG) aiming to foster technological innovation in APEC economies while promoting fair, equitable, and diverse societies, thereby contributing to sustainable development. Detecting, avoiding, and mitigating harmful biases in AI systems is crucial for achieving trustworthy AI ecosystems and creating new opportunities for all APEC economies.

1. Summary of Key Findings in the Research Report and Online Workshop

Artificial Intelligence (AI) has become pervasive in everyday life, encompassing various fields of human activity. Being a general-purpose technology, AI operates within complex socio-technical systems, where social values, customs, and knowledge interact with technological systems, establishing a symbiotic relationship. Consequently, AI systems are prone to replicating and exacerbating existing biases influenced by humans.

Different economies worldwide have adopted diverse strategies to tackle biases arising from the emergence of AI systems, resulting in distinct developmental paths. These strategies blend institutional and technical policies tailored to each economy’s specific context. Successful cases have emerged when local ecosystems align the interests of the public, private, academic, and civil society sectors.

Institutional Perspective

In terms of institutions, most APEC economies have published official documents to promote the adoption of AI. These efforts have involved building institutional capacities through domestic strategies, national policies, roadmaps, and other initiatives. Among the 21 APEC economy members, 19 have implemented various degrees of these initiatives. Only Brunei Darussalam and Papua New Guinea have not reported any deployment of such initiatives, to the best of our knowledge.

The strategies of the 19 economies often encompass ethical considerations related to the fairness and trustworthiness of AI systems, as well as the societal impact of AI adoption. There is widespread awareness among APEC economies about the need to develop AI with a human-centered vision, leveraging its potential while addressing associated risks. However, although these strategies cover a broad range of topics, they generally lack specific mechanisms to address harmful biases in AI systems. Nonetheless, they emphasize the importance of avoiding or mitigating unintentional consequences.

Furthermore, the majority of APEC economies endorse ethical principles to ensure a fair deployment of AI. In some instances, economies indirectly subscribe to these principles by participating in international forums like the OECD and adhering to common principles, even though they are not obligatory. In other cases, economies have developed their own principles through participatory processes, enabling them to address specific areas of interest such as privacy protection or the inclusion of marginalized populations.

Economies that develop their own principles tend to perform better, as they create innovative approaches to manage the risks associated with AI adoption. In such cases, these principles serve as a framework that allows for the adoption of various initiatives, providing clarity on permissible matters, prohibitions, and priority areas concerning ethical concerns. However, similar effects can be achieved if economies promote internal discussions on principles, even if they have been adopted from external sources such as international forums. Additionally, participatory processes serve as effective means to raise awareness about ethical principles and foster engagement in their adoption.

A few APEC economy members have officially outlined approaches to identify, manage, and mitigate harmful biases in AI systems. Notable examples include:

- The United States, where the National Institute of Standards and Technology (NIST) published a document in 2019 to promote technical standards and related tools in AI, followed by the release of the Blueprint for an AI Bill of Rights in 2022.
- China, which addresses Trustworthy Artificial Intelligence in a White Paper released in 2021, creating a Trustworthy AI Framework.
- Australia, through the Australian Human Rights Commission, which published a Technical Paper in 2020 addressing algorithmic biases in Artificial Intelligence.
- Korea, where the National Information Society Agency (NIA) launched a practical guide in 2020 that applies OECD recommendations for trustworthy AI in public organizations.
- Canada, which is currently discussing regulations for the development and application of AI related to Personal Data Protection and Consumer Privacy in 2022.

However, even among economies that have reached a similar level of development in the elaboration of a trustful AI ecosystem, there is no universally agreed-upon or standardized approach to identifying and mitigating harmful biases in AI systems. Instead, the classifications used to categorize biases and establish mitigation approaches are tailored to the specific priorities of each economy, rather than following a uniform global standard. For example, China's White Paper emphasizes that fairness and diversity depend on training datasets to avoid mistrust caused by data bias, focusing on the quality of training data. On the other hand, the Australian Human Rights Commission focuses on various mitigation strategies throughout the AI system lifecycle. It outlines five general approaches to address biases, which means considering different stages of the AI lifecycle, not just data training.

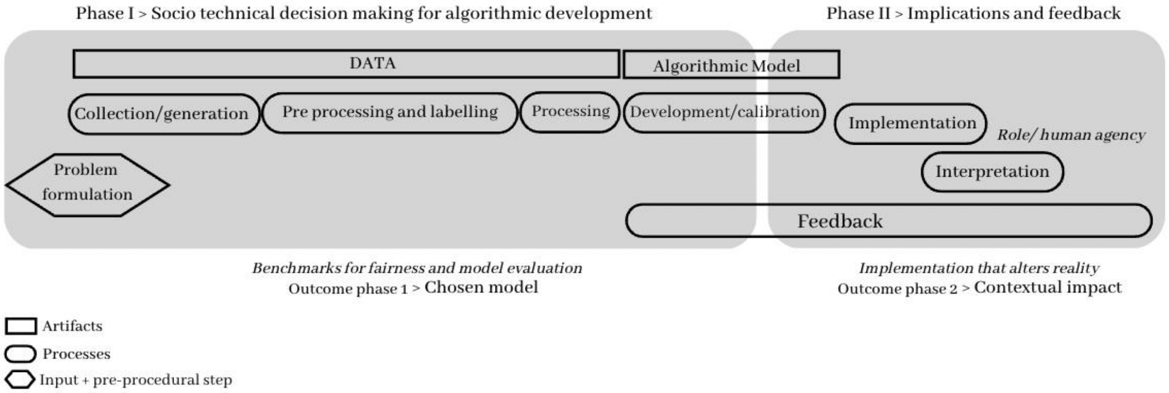
As a result of the diversity in conceptual approaches and frameworks, economies showcase different application cases of AI systems, accumulating lessons from a variety of experiences. Some examples of these applications include Australia, which has implemented pilots in various industries such as the Commonwealth Bank of Australia and Microsoft Chatbots; Canada, which is exploring an Indigenous Protocol for Artificial Intelligence and testing AI solutions in its healthcare system; Singapore, which has compiled practical cases on identity authentication and educational admissions selection, among others; and the United States, which is deploying AI solutions across a wide range of areas, including medical devices.

Furthermore, collaboration between the public and private sectors appears to be crucial in successfully fostering safe and trustworthy AI systems. Interestingly, most initiatives led by official authorities in APEC economies are in the form of guidelines and recommendations, making them voluntary rather than mandatory. Due to the voluntary nature of these frameworks, the alliance between different actors in the AI ecosystem becomes a key requirement for implementing the principles and generating the necessary learnings to strengthen the reliability and trustworthiness of deployed AI systems. Therefore, the avoidance and mitigation of harmful biases in AI systems are more likely to occur in favorable environments where governments, academia, institutes, and industry come together to pursue a common goal of fairer, more inclusive, and equitable development of AI solutions.

Technical Perspective

In terms of technical measures, harmful biases can manifest at any point in the Artificial Intelligence pipeline or lifecycle, including earlier stages such as problem formulation for a specific AI solution. During these stages, AI systems are particularly susceptible to assimilating and reproducing cognitive biases as a result of direct human intervention. Figure 1 illustrates the stages encompassing the AI systems pipeline, as outlined in the first phase research report based on the systematic bibliographic review.

Figure 1. The pipeline of Artificial Intelligence systems


Phase I > Socio technical decision making for algorithmic development
Phase II > Implications and feedback
DATA
Algorithmic Model
Collection/generation
Pre processing and labelling
Processing
Development/calibration
Implementation
Role/ human agency
Problem formulation
Interpretation
Feedback
Benchmarks for fairness and model evaluation
Outcome phase 1 > Chosen model
Implementation that alters reality
Outcome phase 2 > Contextual impact
Artifacts
Processes
Input + pre-procedural step

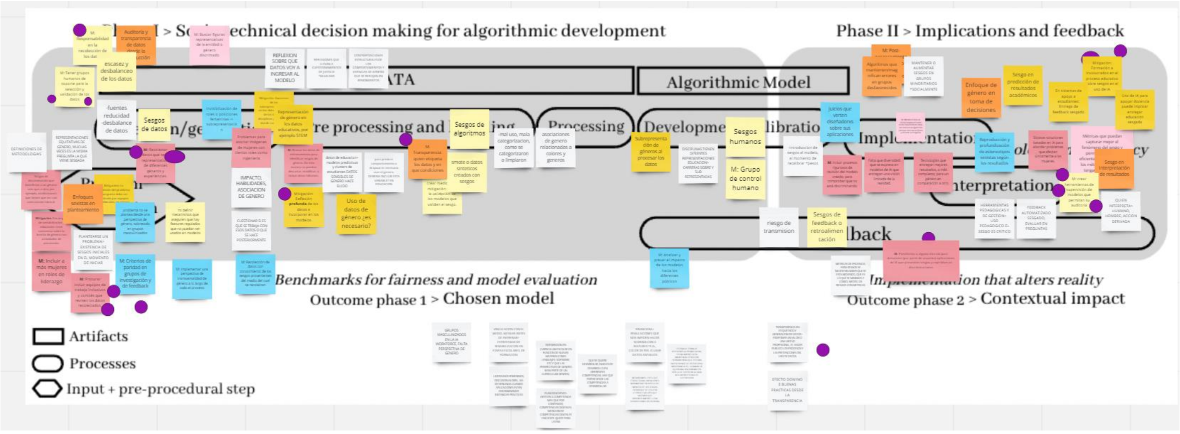
Experts from APEC economies were invited to participate in a two-day workshop with the objective of identifying prominent biases across various stages and topics, including gender bias, the public and financial sectors, recruitment, and the healthcare sector. Each day, approximately 20 participants attended, representing academia, civil society, and the public and private sectors from economies such as Canada; Chile; Mexico; the Philippines; Russia; and Singapore. To facilitate discussions, participants were assigned to breakout rooms focusing on specific topics. Within these rooms, a moderator guided the interactive discussions using an interactive platform. The participants worked together to identify biases, discuss their impact on the AI system pipeline, recommend mitigation strategies, and propose solutions to address future challenges in AI development. The findings were then analyzed and processed, allowing for the synthesis of key topics and conclusions from the two sessions. Finally, the research team reviewed and verified the results, distilling the main insights and recommendations derived from the participatory methodology.

The main conclusions for each topic were as follows:

- Gender bias: Harmful biases related to gender in AI systems encompass the scarcity and imbalance of data, gender associations with colors and roles, under-representation of genders in data processing, and human biases. According to the workshop participants, mitigation strategies should prioritize transparent data collection, diverse problem-solving teams, and bias auditing of models. Education is crucial in addressing the issue early on and raising awareness of how historical gender biases can be perpetuated in AI systems throughout their lifecycle. Feedback mechanisms can help prevent discrimination, and compensation mechanisms should be established once the system is deployed.

Figure 2 provides a summary of the workshop discussions. Each participant used a color to indicate the areas in the AI pipeline where biases were more likely to occur. The figure illustrates that gender biases can manifest at various stages of the AI lifecycle. Particularly, biases were observed to be prominent in the early stages, such as problem formulation, and in the later stages, including feedback and interpretation. These findings highlight the significance of factors such as educational background and awareness in relation to gender biases.

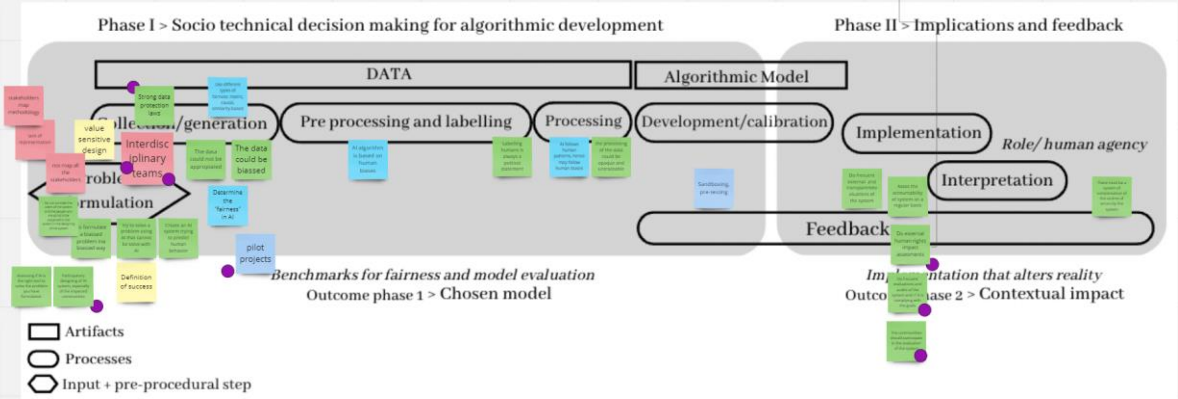
Figure 2. Biases and mitigation strategies identified throughout the AI pipeline for gender


Phase I > Socio-technical decision making for algorithmic development
Phase II > Implications and feedback
DATA
Algorithmic Model
Problem formulation/generation
Pre processing and
Processing
Development
Calibration
Implementation
Interpretation
Feedback
Benchmarks for fairness and model evaluation
Outcome phase 1 > Chosen model
Implementation that alters reality
Outcome phase 2 > Contextual impact
Artifacts
Processes
Input + pre-procedural step
Sesgos de datos
Sesgos de algoritmos
Sesgos humanos
M: Grupo de control humano
Uso de datos de género ¿es necesario?
Enfoque de género en toma de decisiones
Sesgos en predicción de resultados académicos
riesgo de transmisión
Sesgos de feedback o retroalimentación

- Public and financial sector: Harmful biases in the public and financial sectors were prominently identified by the workshop participants in the earlier stages of the AI system development process. The problem formulation for an AI solution can be affected by the lack of representation of different stakeholders, privacy and data protection risks, and political systems that prioritize political goals over the interests of underrepresented populations. Mitigation strategies include stakeholder mapping, assembling interdisciplinary teams, and regular evaluations of the system’s outputs. The importance of participatory design and compensating victims of errors was emphasized.

Figure 3 depicts a summary of the workshop discussions, where participants used a color to indicate the areas in the AI pipeline prone to biases. The figure reveals that biases in the public and financial sectors can arise at various stages of the AI lifecycle, with particular emphasis on the early stages, including problem formulation and data collection/generation. The feedback stage was identified as a critical opportunity to assess the impacts and address any errors resulting from the deployed system. These findings underscore the importance of employing participatory methods and using representative data to mitigate biases in the public and financial sectors.

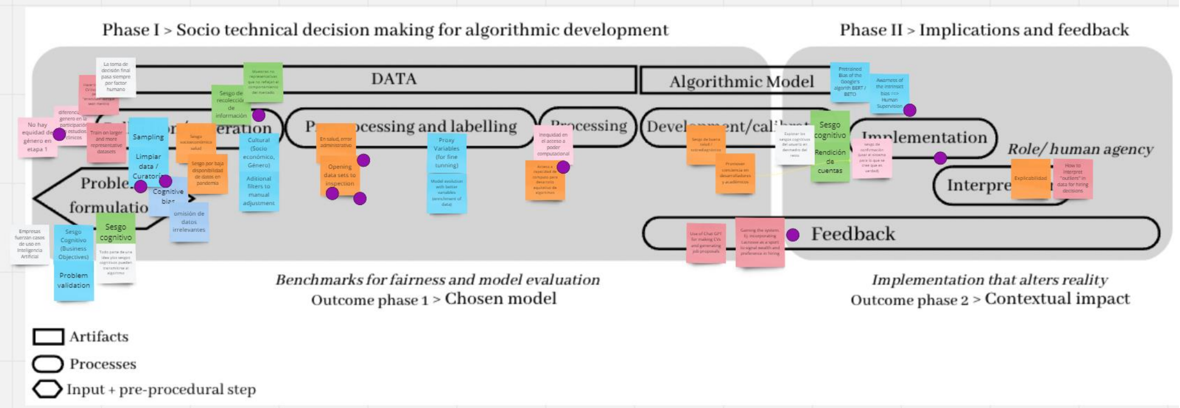
Figure 3. Biases and mitigation strategies identified throughout the AI pipeline for public and financial sectors


Phase I > Socio technical decision making for algorithmic development
Phase II > Implications and feedback
DATA
Algorithmic Model
Collection/generation
Pre processing and labelling
Processing
Development/calibration
Implementation
Role/ human agency
Interpretation
Feedback
Problem formulation
Strong data protection laws
value sensitive design
Interdisciplinary teams
The data could be biassed
Determine the "fairness" in AI
pilot projects
Definition of success
Sandboxing, pre-testing
Do external human rights impact assessments
Benchmarks for fairness and model evaluation
Outcome phase 1 > Chosen model
Implementation that alters reality
Outcome phase 2 > Contextual impact
Artifacts
Processes
Input + pre-procedural step

- Recruitment and healthcare sector: The discussion on recruitment processes and the healthcare sector focused on identifying harmful biases in AI systems and recommending mitigation strategies. Biases such as socio-economic, gender, cognitive, and sampling biases were identified by workshop participants across different stages of the AI pipeline. Suggested mitigation strategies include human supervision, open data sets for inspection, fair access to computing capacity for algorithm development, and promoting awareness among developers and academics. The importance of transparency without compromising intellectual property was also highlighted.

  Figure 4 summarizes the workshop discussions, using the same methodology as described earlier. The figure demonstrates that biases in recruitment processes and the healthcare sector can manifest at different stages of the AI lifecycle. Similar to other sectors discussed previously, biases were found to be more prominent in the early stages, including problem formulation, data collection/generation, and pre-processing and labeling. The development/calibration and algorithmic modeling stages were also identified as important areas for implementing mitigation measures. These findings underscore the importance of continuous human supervision in addressing biases in recruitment processes and the healthcare sector.

Figure 4. Biases and mitigation strategies identified throughout the AI pipeline for recruitment and healthcare sectors

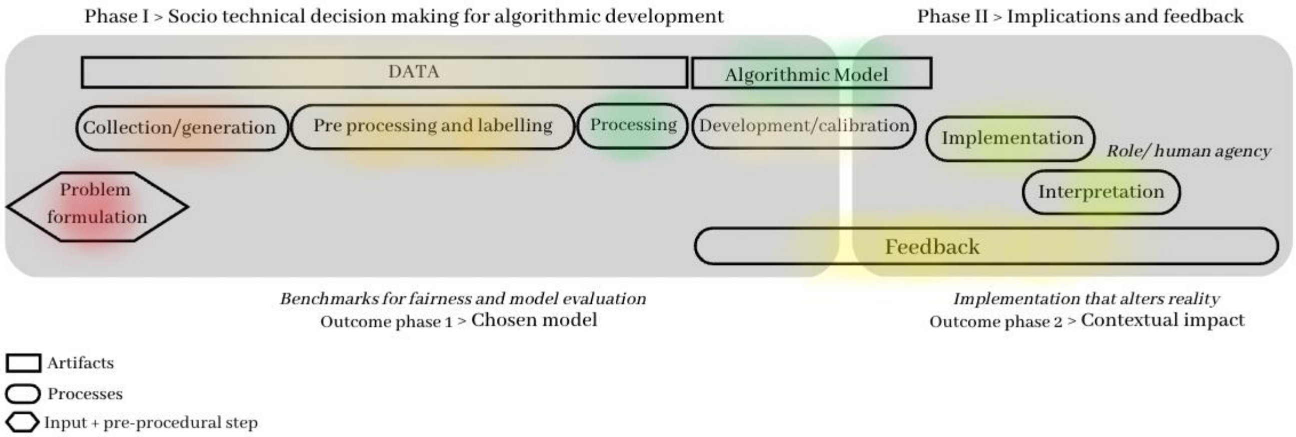

Phase I > Socio technical decision making for algorithmic development
Phase II > Implications and feedback
DATA
Algorithmic Model
Pre-processing and labelling
Processing
Development/calibration
Implementation
Role/ human agency
Interpretation
Feedback
Problem formulation
Benchmarks for fairness and model evaluation
Outcome phase 1 > Chosen model
Implementation that alters reality
Outcome phase 2 > Contextual impact
Artifacts
Processes
Input + pre-procedural step
La toma de decisión final pasa siempre por factor humano
No hay equidad de género en etapa 1
diferencia genero en la participación estudios clínicos
Train on larger and more representative datasets
Sampling
Limpiar data / Curatoría
Cognitive bias
omisión de datos irrelevantes
Sesgo socioeconómico salud
Sesgo por baja disponibilidad de datos en pandemia
Sesgo de recolección de información
Muestras no representativas que no reflejan el comportamiento del mercado
Cultural (Socio económico, Género)
Aditional filters to manual adjustment
En salud, error administrativo
Opening data sets to inspection
Proxy Variables (for fine tunning)
Model evolution with better variables (enrichment of data)
Inequidad en el acceso a poder computacional
Acceso a capacidad de cómputo para desarrollo equitativo de algoritmos
Empresas fuerzan casos de uso en Inteligencia Artificial
Sesgo Cognitivo (Business Objectives)
Problem validation
Sesgo cognitivo
Todo parte de una idea y los sesgos cognitivos pueden transmitirse al algoritmo
Sesgo de buena salud / sobrediagnóstico
Promover conciencia en desarrolladores y académicos
Explorar los sesgos cognitivos del usuario en desmedro del resto
Sesgo cognitivo
Rendición de cuentas
Pretrained Bias of the Google's algorith BERT / BETO
Awarness of the intrinsict bias ==> Human Supervision
sesgo de confirmación (usar el sistema para lo que se cree que es verdad)
Explicabilidad
How to interpret "outliers" in data for hiring decisions
Use of Chat GPT for making CVs and generating job proposals
Gaming the system. Ej. incorporating Lacrosse as a sport to signal wealth and preference in hiring

Overall, it is possible to identify common trends from the experts' perspective across the different topics. One significant concern is the type of biases resulting from direct human intervention, including biased problem formulation, lack of awareness, data labeling, and cognitive biases such as confirmation bias. As a result, mitigation strategies tend to focus on human-in-the-loop approaches, such as forming diverse and multidisciplinary teams, implementing participatory designs, conducting data process audits, and establishing feedback mechanisms to continuously monitor the performance of the deployed AI systems.

Therefore, biases and mitigation strategies were predominantly identified around the borders of the AI pipeline, with particular emphasis on earlier stages such as problem formulation, data collection or generation, pre-processing, and data labeling. The final stage of feedback is also crucial.

A heat map was created based on the consolidated number of biases and mitigation strategies identified by the experts in the three breakout rooms during the workshop, highlighting the critical importance of these stages in the pipeline. As shown in Figure 5, red corresponds to the highest concentration of interactions (the stage where the most biases and mitigation strategies were identified): "Problem formulation" had the highest concentration, followed by "Collection/generation" and "Pre-processing and labeling". In fact, according to the experts' observations, the first three stages accounted for nearly 60% of the identified biases and mitigation strategies. On the other hand, "Processing" and "Algorithmic Model" had the lowest number of biases and mitigation strategies identified, representing less than 5% of the occurrences each.

Figure 5. Heat map of biases and mitigation strategies identified throughout the AI pipeline by experts during the online workshop


Phase I > Socio technical decision making for algorithmic development
Phase II > Implications and feedback
DATA
Algorithmic Model
Collection/generation
Pre processing and labelling
Processing
Development/calibration
Implementation
Role/ human agency
Interpretation
Problem formulation
Feedback
Benchmarks for fairness and model evaluation
Outcome phase 1 > Chosen model
Implementation that alters reality
Outcome phase 2 > Contextual impact
Artifacts
Processes
Input + pre-procedural step

Additionally, certain technical strategies have been frequently mentioned, such as ensuring data availability, data curation, and avoiding unnecessary information that could introduce undesirable biases. However, these strategies come with their own challenges, such as the trade-off between data availability and data protection, as well as issues related to accessing technical resources that may not be affordable for everyone. These challenges will be further discussed in this report as future considerations.

2. Key Recommendations Based on Successful Case Studies

As a socio-technical system, the successful development of Artificial Intelligence requires a combination of institutional and technical perspectives to effectively address its challenges. The research conducted in the first phase of this project has demonstrated that both dimensions can be developed concurrently and independently; however, these dimensions can be further strengthened when aligned. A robust institutional environment, supported by a clear framework, provides certainty for the development of AI systems by establishing guidelines and goals that can inform the modeling process.

In the following, key recommendations based on successful case studies covered during the research and online workshop will be presented. These recommendations are categorized into institutional and technical recommendations, reflecting the different aspects that need to be considered in fostering the advancement of AI systems.

Institutional recommendations

Institutional recommendations encompass the macro-level regulatory environment and policies necessary to create ecosystem enablers that foster the innovation and adoption of safe and trusted AI systems.

1) Creation and adoption of domestic frameworks: Successful case studies have highlighted the importance of establishing domestic frameworks as a critical first step in AI governance. These frameworks help define principles, values, and goals in alignment with local regulations, providing a foundation for innovation in the AI development ecosystem. They bridge the gap between high-level ethical principles and practical implementation, guiding the actions of various stakeholders. While adherence to international frameworks is desirable for creating common rules globally, fostering cross-board interoperability, and learning from good practices, local adaptations of foreign examples are crucial to ensure relevance within the vernacular context.
2) The principle of *learning-by-doing*: As AI is a disruptive and evolving technology, regulations face the challenge of striking a balance between protecting customer rights through adequate regulation and promoting innovation by avoiding over-regulation. Successful case studies demonstrate the use of flexible regulations and regulatory experimentation, allowing the gradual establishment of mandatory mechanisms. Soft regulations, such as recommendations, can evolve into hard laws, such as bills or acts, over time. Guidelines, recommendations, and self-assessments serve as initial monitoring and assessment tools for AI systems, raising awareness and encouraging organizations to support specific regulatory frameworks. Initiatives may include recommendations for operational management, the promotion of testing frameworks, and the use of testing technologies.
3) Participatory instances in early phases: To ensure strong engagement, it is crucial to consider multiple perspectives during the creation of the institutional environment. Successful case studies have incorporated participatory instances in the early phases of policy development, involving scholars, institutes, companies, and international organizations. Participatory methodologies enhance the robustness of frameworks and policies, and foster engagement among participants in the AI ecosystem.

4) Multistakeholder approach: Given the complexity of ecosystems in which AI systems are developed, involving multiple stakeholders is essential. This typically includes the government as the primary regulator; the academia for research, use, and development of AI systems; the industry as users, developers, and providers of AI solutions; and civil society and customers as major users. A multistakeholder approach facilitates the exchange of cutting-edge knowledge, practical applications, trustworthiness, and engagement, resulting in positive impacts on the AI ecosystem.
5) Proactive mechanisms to address harmful biases: Effective governance and robust measures enable the adoption of proactive mechanisms to detect, prevent, and mitigate harmful biases as they arise. These mechanisms are often embedded in sector-specific agencies or regulatory entities, with a focus on prioritized sectors for monitoring. An example of proactive mechanisms found in successful case studies is proactive risk assessment, which helps identify gaps and potential risks in specific AI systems. Implementing proactive mechanisms requires an enforcement structure with the authority to collect data, assess AI systems, and enforce suggested measures. In cases of non-compliance with safety measures, regulators may have the authority to order the cessation of a particular AI system.

Technical recommendations

Technical recommendations encompass micro-level practices and mechanisms aimed at creating safe and trusted AI systems. Taking a tool-agnostic approach, these recommendations focus on the AI lifecycle and suggest practices to counteract biases resulting from direct human intervention.

1) AI solutions and defining fairness: Biases can arise when AI systems are used without a clear problem to solve. Problem-oriented AI systems should be employed, ensuring that strictly necessary measures are taken to address the original problem. Furthermore, a definition of fairness should be adopted that allows for quantitative measurement. Different approaches to fairness metrics, such as metric, causal, and similarity-based, should also be considered.
2) Formulating the problem and assessing impacts through interdisciplinary teams and participatory designs: It is essential to include communities affected by AI systems in the formulation of problems and assessment of impacts during the feedback stage.

This practice aligns with the institutional recommendation of participatory instances and helps prevent biases resulting from the under-representation of specific groups. Gender criteria should also be considered in these instances.

3) Creating supervisor tools for model auditing: It is important to develop tools that enable the auditing of AI models. These tools should not be restricted to specific technologies but should also establish inclusive work teams and committees to audit collected data. Mechanisms for gathering representative data and ensuring that the collected data passes non-bias filters should be implemented.
4) Transparency and explainability of AI systems: There is a significant concern regarding the transparency of AI systems, particularly in cases where explainability is necessary to determine eventual responsibilities. This concern is heightened by the complexity of machine-learning systems, which can make traceability during data processing challenging. To address this transparency problem, data governance should be established, registering individuals responsible for each stage of the process and enhancing traceability throughout the entire lifecycle.
5) Cleaning data and implementing a proper data curation process during the early stages of data collection and generation: One common bias that can occur during this stage is non-representative sampling, leading to unreliable results by over-representing certain groups and under-representing others. It is recommended to build theoretical-based AI models and avoid including unnecessary features and data that could introduce undesirable biases.
6) Opening datasets for inspection and auditing errors during the pre-processing and labeling stage: Biases can emerge during these stages due to rough labeling processes, which create proxies that result in biased outcomes. Incomplete data can also contribute to biases, and mitigation mechanisms such as the Synthetic Minority Oversampling Technique (SMOTE) can be employed to address imbalanced data. However, it is crucial to have human monitoring available as SMOTE is also susceptible to errors.
7) Ensuring equal access to computing capacity for algorithm development: Unequal access to computing capacity during the data processing stage raises concerns about fairness. Computing capacity plays a vital role in enabling the development of AI systems, and disparities in access can affect participation and data representativeness. Measures to address this issue include investing in computing

infrastructure and ensuring data availability for the creation of more representative datasets.

3. Key Challenges for Future Discussion

Key challenges for future discussions have been identified by experts during interviews and the online workshop. They have identified several dilemmas that need to be addressed in the future development of AI systems, particularly in sensitive fields crucial to the successful deployment of fair AI systems. These challenges include personal data protection and data availability, the regulatory environment for the public and private sectors, and the implementation of global principles in local contexts.

Addressing these challenges will require ongoing discussions, collaborations, and interdisciplinary efforts to ensure the development of AI systems that are fair, safe, and trusted.

Data protection and data availability

The use of data in AI systems poses a dilemma between data availability and privacy. While more data enhances the precision and development of AI systems, incorporating data from various sources can raise privacy concerns. Therefore, striking a virtuous balance is essential.

The experts' discussion revealed that while system precision is desirable, the purpose of the system must be carefully evaluated to determine if the privacy concerns outweigh its benefits. In other words, considering data privacy and autonomy as inalienable rights, they should take precedence over an AI system that risks violating them.

However, algorithms and sensitive information are still at risk of exposure due to the extraction and disclosure of private data. Even in cases where the purpose has been carefully considered, there are several avenues through which this risk persists. Firstly, an increasing number of AI systems are being developed using open sources, making it challenging to enforce data privacy regulations. Secondly, general-purpose systems like Large Language Models (LLMs) can be used for harmful purposes, even if their initial intent was noble, making it difficult to anticipate their future uses. Thirdly, data protection relies not only on legal compliance but also on cybersecurity technology, which is crucial for ensuring the actual safeguarding of data against criminal attempts.

On the other hand, transparency promotes the representation of minority groups in AI datasets, enables the backup of important data, fosters unbiased software frameworks, and ensures public accountability, including explainability of AI systems. Therefore, compensatory mechanisms are indispensable for minimizing privacy breaches and fostering the trustworthy development of precise AI systems. Some examples of these mechanisms include privacy-preserving techniques such as differential privacy, federated learning, or secure multi-party computation; anonymization and de-identification processes; data minimization and purpose limitation by collecting only necessary data for specific purposes and not retaining it longer than necessary; and secure data sharing frameworks like secure data enclaves, data trusts, or federated learning frameworks.

In summary, regulating general-purpose systems and AI development is challenging due to the lack of clear purpose and the complexities of navigating multiple regulations. Reviewing the purpose of a general-purpose system, such as an LLM, as an initial filter is not sufficient since it is difficult to foresee its various uses. Furthermore, regulating AI development is challenging as open code and open data communities are increasingly thriving in the AI ecosystem. The dilemma between privacy/data protection and transparency/data availability remains significant, and advancements in cybersecurity are becoming increasingly crucial for effective data protection and mitigating the risks of potential privacy breaches.

General regulatory principles and local principles

Currently, there are general principles aimed at promoting the development of trustworthy AI ecosystems, supported by international organizations that foster collaboration among various economies. For example, the OECD, fAIr LAC initiative (Inter-American Development Bank), and UNESCO have established principles of trust and fair AI systems. However, for these principles to be effective and appropriate for each culture, they need to be applied locally. The challenge lies in finding a balance between general principles that provide a common framework and principles that are specific enough to be implemented in a local context, respecting the diversity and autonomy of communities.

On one hand, universal principles are valuable for facilitating collective efforts, establishing shared frameworks for technology use, and promoting consistency and equivalences among regulatory environments in different economies. Compliance with universal principles brings advantages such as knowledge exchange, shared experiences, adoption of good practices, and the development of global systems. On the other hand, universal principles carry the risk of exacerbating the North-South gap, as they may be influenced by the perspectives of the Global North, potentially overshadowing the values and principles of the Global South, overlooking cultural and political differences.

This bias of over-representation is not limited to institutional fields but can also manifest in AI systems. For example, language processing tools often prioritize English, rendering them useless for regions where English is not widely spoken. Consequently, universal principles may conflict with local identity and needs.

Two aspects were explored as potential mechanisms to address this challenge. Firstly, adopting a "Glocal approach," which involves thinking globally and acting locally. Efforts should focus on building inclusive AI systems by incorporating diverse groups that can contribute to and enhance the system's design process. To achieve this, it is important to seek compatibility and integration of global and local principles, consider local needs while incorporating foreign advancements, and embrace local characteristics to promote algorithmic diversity and design processes.

Secondly, the ability to adapt to constant technological changes must be embraced. This underscores the importance of the tool-agnostic principle, allowing for flexibility and adaptation to different contexts and specific tools, while keeping the overarching goals in mind. Therefore, universal principles should be able to accommodate various scenarios and tools as needed. Additionally, promoting local discussions within the AI ecosystem and considering diverse algorithms can lead to cultural adaptation specific to each economy.

The main conclusions highlight the need for inclusive AI systems, global-local compatibility and integration, cultural adaptation, diverse algorithms, and stakeholder representation. The challenges identified include biased language processing tools/algorithms, inadequate representation of stakeholders' needs, uncertainty in technological advancements, and the integration of local characteristics while maintaining a global perspective. Furthermore, an additional challenge will be to establish effective cross-border interoperability among different economies.

Therefore, APEC serves as a valuable forum for economies from emerging regions to have representation in global discussions and the development of AI systems. It facilitates the convergence of global perspectives into local and specific contexts, promoting the establishment of fair and trustworthy AI ecosystems.

Private regulation, public regulation, and regulatory experimentation

Economies typically employ different regulations to oversee the performance of AI systems in the public and private sectors, despite these systems sharing common fundamentals. Consequently, the same algorithm might be assessed by different authorities and held to different standards. The workshop explored whether it is necessary to bridge this gap or whether having two parallel regulations is desirable.

Experts acknowledge that the public and private sectors have different objectives, justifying the existence of distinct regulations. The public sector aims to protect rights and provide legal certainty, while the industry seeks to foster innovation and typically operates based on market needs. Furthermore, the enforcement of regulations in these sectors differs, with the public sector generally having easier enforcement mechanisms than the industry. Thus, it is natural to have different regulations or institutions responsible for enforcement, with soft regulations often associated with the private sector and stricter regulations with the public sector.

However, both sectors also have complementary needs, which presents an opportunity to bridge the gap between them. On one hand, private sector involvement in AI development requires legal certainty and a common framework for all stakeholders, as this facilitates the feasibility of initiatives and promotes social acceptance. On the other hand, the public sector requires greater flexibility to adapt to a rapidly changing environment, given that technological advancements often outpace the pace of regulatory development. Consequently, regulatory divergence may impede innovation.

Technology governance requires stakeholder engagement, alignment with public values, and clear enforcement rules. The current situation has disadvantages, including a lack of alignment with public values, ambiguity in regulations due to the rapid development and adoption of AI systems, and knowledge gaps between regulators and AI developers. Therefore, achieving partial or total integration of regulatory systems is a desirable goal.

To address this challenge, potential alternatives were explored. Given the need for flexibility in the public sector and the need for shared and clear rules in the private sector, regulatory experimentation emerges as a strong alternative to bridge the existing regulatory gaps. Mechanisms such as regulatory sandboxes, pilots, and policy prototyping can offer a balanced solution that protects rights while providing the necessary flexibility to avoid stifling innovation. Active collaboration between the public and private sectors is also essential to bridge the gaps between regulatory bodies and AI developers. Regulatory experimentation can serve as a participatory platform for this collaboration and knowledge exchange.

In summary, the coexistence of parallel regulatory and enforcement systems can be seen as a natural equilibrium between the public and private sectors, considering their distinct goals and characteristics. However, recognizing their complementary needs and finding ways to bridge the gap between these systems remains important. Establishing participatory instances with the potential for regulatory experimentation offers a promising solution to this challenge. Alternatively, the partial or complete integration of both systems remains a challenge that needs to be overcome.