

# AI4Gov

Trusted AI for Transparent Public Governance  
fostering Democratic Values

## Deliverable 5.2

# Input Papers to Facilitate the Workshops on Awareness Raising V2

<31-12-2024>


Version 1.0



Funded by  
the European Union

*Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Agency. Neither the European Union nor the granting authority can be held responsible for them.*



PROPERTIES	
<b>Dissemination level</b>	Public
<b>Version</b>	1.0
<b>Status</b>	Final
<b>Beneficiary</b>	
<b>License</b>	 <p>This work is licensed under a Creative Commons Attribution-No Derivatives 4.0 International License (CC BY-ND 4.0). See: <a href="https://creativecommons.org/licenses/by-nd/4.0/">https://creativecommons.org/licenses/by-nd/4.0/</a></p>

AUTHORS		
	Name	Organisation
<b>Document leader</b>	Panagiotidou Georgia	AUTH
<b>Participants</b>	Bouranta Vasiliki	AUTH
<b>Reviewers</b>	Kyrkou Danai	VIL
	Stroia-Vlad Iuliana	SIE
	Tanja Zdošek Draksler	JSI

<b>VERSION HISTORY</b>				
<b>Version</b>	<b>Date</b>	<b>Author</b>	<b>Organisation</b>	<b>Description</b>
0.1	08/12/2024	Panagiotidou Georgia, Bouranta Vasiliki	AUTH	ToC
0.5	10/12/2024	Panagiotidou Georgia, Bouranta Vasiliki	AUTH	Draft
0.6	16/12/2024	Kyrkou Danai	VIL	Reviewed Document
0.7	20/12/2024	Stroia-Vlad Iuliana	SIE	Reviewed Document
0.8	23/12/2024	Panagiotidou Georgia	AUTH	Reviewed Final Draft
0.9	23/12/2024	Tanja Zdolšek Draksler	JSI	Reviewed Document
1.0	23/12/2024	Panagiotidou Georgia	AUTH	Final

# Table of Contents

<b>Abstract .....</b>	<b>8</b>
<b>1 Introduction.....</b>	<b>9</b>
1.1 Purpose and scope.....	9
1.2 Document structure.....	10
<b>2 Understanding AI Bias and Discrimination .....</b>	<b>12</b>
2.1 Definition of AI Bias.....	12
2.2 Impact of AI Discrimination.....	12
<b>3 Preliminary Focus Groups with Pilot Cases Users .....</b>	<b>14</b>
3.1 Description.....	14
3.2 Main Findings.....	14
3.2.1 <i>AI and Sustainable Development Goals (SDGs)</i> .....	14
3.2.2 <i>Evaluation Methodologies</i> .....	14
3.2.3 <i>Addressing Biases in AI</i> .....	14
3.2.4 <i>AI Platform for Sustainable Tourism Policy</i> .....	15
3.2.5 <i>Community Engagement</i> .....	15
3.2.6 <i>Human Factors and Infrastructure</i> .....	15
3.2.7 <i>Ethical and Legal Considerations</i> .....	15
3.2.8 <i>Real-Time Data and Decision Support</i> .....	15
3.2.9 <i>Recommendations</i> .....	16
<b>4 1<sup>st</sup> Panel Discussions .....</b>	<b>17</b>
4.1 “Artificial Intelligence in Local Governance: Prospects and Challenges” - Key Findings and Recommendations .....	17
4.1.1 <i>Digital Transformation and AI Implementation in Local Governance</i> .....	17
4.1.2 <i>Challenges and Ethical Considerations</i> .....	18
4.1.3 <i>AI in Tourism Policy and Management</i> .....	18
4.1.4 <i>Recommendations</i> .....	19
4.2 “Accessible and Inclusive Artificial Intelligence for Citizens” - Key Findings and Recommendations .....	19
4.2.1 <i>Digital Access and inclusivity</i> .....	19
4.2.2 <i>The Role of Community Centres</i> .....	20
4.2.3 <i>Education and Training Programs</i> .....	20
4.2.4 <i>Digital Transformation in Municipal Services</i> .....	20
4.2.5 <i>Challenges and Ethical Considerations</i> .....	20
4.2.6 <i>Youth Engagement and Digital Skills</i> .....	21
4.2.7 <i>Recommendations</i> .....	21
<b>5 2<sup>nd</sup> and 3<sup>rd</sup> Panel Discussions.....</b>	<b>22</b>
5.1 2 <sup>nd</sup> Panel Discussion - Rethinking Healthcare with Digital Health Technology: Key findings .....	22
5.1.1 <i>AI in Healthcare and Regulatory Challenges</i> .....	23
5.1.2 <i>Transparency and Trust in AI Systems</i> .....	23
5.1.3 <i>Impact of AI on Doctor-Patient Relationship</i> .....	24
5.1.4 <i>Ethical and Privacy Concerns</i> .....	24
5.1.5 <i>Digital Literacy and Training</i> .....	24
5.1.6 <i>AI’s Role in Public Health and Social Services</i> .....	24
5.1.7 <i>Critical AI Literacy</i> .....	24
5.2 3 <sup>rd</sup> Panel Discussion – AI and Governance .....	25
5.2.1 <i>AI in Governance</i> .....	25
5.2.2 <i>Bias and Discrimination</i> .....	25
5.2.3 <i>Transparency and Inclusivity</i> .....	25
5.2.4 <i>Raising Awareness and Participation in Decision-Making and Democracy</i> .....	26
<b>6 4<sup>th</sup> and 5<sup>th</sup> Panel Discussions .....</b>	<b>27</b>
6.1 4 <sup>th</sup> Panel Discussion - Ethical AI and Democratic Governance: Challenges and Innovations in Political Methodology .....	27
6.1.1 <i>Ethical AI in Governance and Policy Making</i> .....	28

6.1.2	<i>The Role of Regulatory Frameworks</i> .....	29
6.1.3	<i>Public Engagement and Civic Participation</i> .....	29
6.1.4	<i>Addressing Misinformation and Disinformation</i> .....	30
6.1.5	<i>Challenges in Local Governance and Capacity Building</i> .....	30
6.1.6	<i>Speed of Technological Advancements vs. Policy Response</i> .....	30
6.1.7	<i>Recommendations</i> .....	31
6.2	4 <sup>th</sup> Panel Discussion - Ethical and Practical Dimensions of Artificial Intelligence in Public Administration and Society: Key Findings.....	32
6.2.1	<i>The Transformative Role of AI in Governance</i> .....	33
6.2.2	<i>Ethical Challenges in AI Implementation</i> .....	33
6.2.3	<i>Data Governance and Accessibility</i> .....	33
6.2.4	<i>AI in Healthcare and Public Safety</i> .....	34
6.2.5	<i>Regulatory and Policy Frameworks</i> .....	34
6.2.6	<i>Education and Public Awareness</i> .....	34
6.2.7	<i>Addressing Algorithmic Bias</i> .....	34
6.2.8	<i>Recommendations</i> .....	35
<b>7</b>	<b>Existing Awareness Strategies</b> .....	<b>36</b>
7.1	Educational Campaigns .....	36
7.1.1	<i>Public Awareness Campaigns</i> .....	36
7.1.2	<i>Social Media</i> .....	37
7.1.3	<i>Interventions in Educational Institutions</i> .....	37
7.2	Policies and Regulations .....	38
7.3	Corporate Initiatives .....	39
7.3.1	<i>Strategies of Addressing AI bias Internally</i> .....	39
7.3.2	<i>Examples of Corporate Responsibility in AI Ethics</i> .....	40
<b>8</b>	<b>Developing Effective Workshops</b> .....	<b>41</b>
8.1	Workshop Objectives .....	42
8.2	Target Audiences .....	43
8.3	Workshop Content.....	44
8.3.1	<i>Introduction to AI</i> .....	44
8.3.2	<i>Identifying Sources of Bias in AI</i> .....	44
8.3.3	<i>Techniques for Mitigating Bias in AI Systems</i> .....	45
8.3.4	<i>Ethical Considerations in AI Development</i> .....	45
8.3.5	<i>Hands-on Activities and Case Studies</i> .....	45
<b>9</b>	<b>Proposed Curriculum: Workshops on AI Bias and Discrimination</b> .....	<b>46</b>
9.1	Module 1: The Ethical Foundations of AI .....	46
9.2	Module 2: Understanding AI Bias and Its Societal Impact .....	47
9.3	Module 3: Identifying Sources of Bias in AI Systems.....	48
9.4	Module 4: Sources and Dynamics of AI Bias .....	48
9.5	Module 5: Mitigating AI Bias and Ensuring Inclusivity .....	49
9.6	Module 6: Tools and Techniques for Bias Mitigation.....	49
9.7	Module 7: Ethical AI Frameworks and Best Practices .....	50
9.8	Module 8: Advocating for Transparent and Accountable AI .....	50
9.9	Module 9: Policy Implications and Advocacy.....	51
9.10	Module 10: Collaborative Problem-Solving for Ethical AI .....	51
9.11	Evaluation and Certification.....	52
9.12	Learning Methods and Delivery .....	52
9.13	Target Audience .....	52
9.14	Learning Outcomes .....	52
<b>10</b>	<b>Conclusions</b> .....	<b>53</b>
<b>11</b>	<b>References</b> .....	<b>55</b>

## Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
DPB	Diputación Provincial de Badajoz
HRF	Holistic Regulatory Framework
JSI	Jozef Stefan Institute
KPI	Key Performance Indicator
MOOCs	Massive Open Online Courses
MT	Greek Ministry of Tourism
NGO	Non-governmental organization
SDG	Sustainable Development Goals
UNESCO	United Nations Educational, Scientific and Cultural Organization
VVV	Municipality of Vari-Voula-Vouliagmeni

## Abstract

This document results from the Task 5.1 of Work Package 5 (WP5) in the AI4Gov project, aimed at fostering citizen engagement and inclusiveness through ethical AI technologies. WP5 focuses on enhancing awareness and promoting ethical AI practices among stakeholders and the public. Task 5.1 involved organizing panel discussions on AI governance, inclusiveness, discrimination, biases, and transparency. These discussions, led by Aristotle University of Thessaloniki (AUTH), provided valuable insights into mitigating AI bias and discrimination and enhancing inclusiveness, representation, participation, openness, pluralism, and tolerance.

The document presents a comprehensive analysis based on three (3) focus groups and five (5) panel discussions held during the first 24 months of the AI4Gov project. It aims to inform and facilitate workshops designed to raise citizen awareness about AI bias and ethics. The findings and recommendations offer actionable insights for developing effective awareness-raising strategies, ultimately serving as a foundational resource for creating workshops that educate and empower citizens to advocate for ethical AI practices.



# 1 Introduction

## 1.1 Purpose and scope

The present document is an outcome of all the activities completed throughout the implementation of the Task 5.1 which is part of WP5 in AI4Gov project.

The AI4Gov project aims to foster citizen engagement and inclusiveness in democratic processes through the development and implementation of ethical and transparent AI technologies. WP5 specifically focuses on enhancing awareness and promoting ethical AI practices among stakeholders and the public. The goal of T5.1 has been to organize panel discussions focused on ethical aspects of AI when used in Governance and mainly inclusiveness, discriminations, biases and transparency.

According to T5.1 prerequisites several panel discussions with key actors from different sectors that use AI is expected to take place within the project's lifecycle. The Panel discussions enabled participants to exchange views, provide input depending on their expertise, area of activities, interests and participation in different policy areas.

Acknowledging that AI is more than a technical issue, these discussions examined the trade-offs related to fairness, bias, civic engagement and raise of awareness, the types of stakeholders to be targeted and the means to approach the latter towards supporting societal change (e.g., technicality of AI and ways to transfer relevant knowledge to non-experts).

The Lead partner (Aristotle University of Thessaloniki, AUTH) organized the panel discussions on Inclusive AI in conference venues. The workshops to be organized within the AI4Gov project should be based on input papers with data on existing awareness-raising strategies to mitigate AI bias and discrimination and to enhance inclusiveness, representation, participation, openness, pluralism, and tolerance.

As Artificial Intelligence (AI) is increasingly integrated into various aspects of daily life, it is promising significant advancements but is also raising critical ethical and social concerns. Among these concerns, AI bias and discrimination have emerged as pressing issues that can perpetuate existing inequalities and introduce new forms of injustice. Therefore, this document presents a comprehensive analysis based on focus groups and panel discussions that were organized until M24 of the AI4Gov project, centred on AI bias and ethics. The primary aim is to inform and facilitate the development and implementation of workshops designed to raise awareness among citizens about these critical issues in the life cycle of AI4Gov project.

The goals of this initiative are to enhance understanding of AI bias and its societal impacts, to promote ethical practices in AI development and deployment, and to foster an inclusive dialogue among stakeholders. By engaging diverse participants, including local community members, experts, and policymakers, these workshops aim to cultivate a well-informed public that can actively participate in shaping the ethical frameworks governing AI technologies.

The content of this document is structured to provide a detailed exploration of AI bias, its manifestations, and implications across different sectors, including healthcare, governance, and

sustainable development. It also synthesizes the findings and recommendations from various discussions, providing actionable insights for developing effective awareness-raising strategies. Ultimately, this document serves as a foundational resource for creating workshops that not only educate but also empower citizens to advocate for ethical AI practices.

## 1.2 Document structure

The document begins with an *Abstract*, providing a brief summary of the document's contents, goals, and key findings. Following the abstract, the *Introduction* outlines the purpose and scope of the document, explaining the goals and aims of the initiative and describing the content and structure of the document.

The next section, *Understanding AI Bias and Discrimination*, defines AI bias and discusses its impact across various societal domains. It elaborates on both conscious (explicit) and unconscious (implicit) biases, highlighting their potential to perpetuate social inequalities and injustices.

In *Preliminary Focus Groups with Pilot Case Users*, the document synthesizes insights from focus group discussions held in April 2024. This section describes the focus groups, their main findings, and recommendations related to AI and Sustainable Development Goals (SDGs), evaluation methodologies, and addresses biases in AI.

The *Panel Discussions* section summarizes key points from a series of in-depth discussions on the intersection of AI and governance. It includes detailed findings and actionable recommendations across various domains, exploring topics such as digital transformation, AI implementation in local governance, and challenges related to transparency, trust, and ethical considerations. Panels have also addressed the transformative potential of AI in fields such as tourism policy, where AI-driven tools can enhance sustainability and resource management, and the importance of inclusivity in AI systems to ensure accessibility for all citizens.

In addition to these foundational discussions, the section incorporates insights from panels focused on AI's role in governance and public services. These discussions highlighted the opportunities and challenges of AI in advancing policymaking, healthcare, and societal outcomes. Common themes included regulatory challenges, the need for transparency and trust, and the importance of ensuring that AI systems are both effective and equitable. Ethical concerns such as privacy, bias, and inclusivity were prominent throughout, emphasizing the need for robust frameworks and responsible implementation of AI technologies.

Recent panels further enriched these discussions by exploring how AI is shaping political methodology and data analysis. They emphasized the dual need for innovation and ethical accountability, underscoring the importance of interdisciplinary approaches to tackle the complex implications of AI on governance and society. Together, these panels provide a comprehensive overview of the potential and challenges of integrating AI into public administration and beyond, offering valuable recommendations for fostering responsible and impactful use of these technologies.

*Existing Awareness Strategies* provides an overview of current strategies aimed at mitigating AI bias and discrimination. It discusses educational campaigns, public awareness initiatives, policies,

regulations, and corporate efforts to promote fairness and accountability in AI systems. *Developing Effective Workshops* outlines guidelines for creating workshops to raise awareness about AI bias and discrimination. This section details the workshop objectives, content, and structure, emphasizing the importance of combining theoretical knowledge with practical exercises.

To address the growing need for awareness and education around the ethical implications of AI, another section outlines a *Proposed curriculum for workshops* aimed at raising awareness about AI bias and discrimination. These workshops target a diverse audience, including policymakers, educators, developers, and members of the public. The curriculum focuses on understanding the sources and consequences of AI bias, exploring case studies of discriminatory outcomes, and equipping participants with tools to identify and mitigate bias in AI systems. Through a combination of lectures, interactive activities, and group discussions, the workshops should foster critical thinking about the societal impact of AI and promote the development of fair and inclusive technologies. This initiative aims to empower stakeholders to make informed decisions and advocate for ethical practices in AI design and deployment

The *Conclusions* section summarizes the main findings from the focus groups and panel discussions, offering key recommendations for mitigating AI bias and enhancing inclusiveness, participation, and transparency in AI applications. References lists all sources and references cited throughout the document, ensuring that the information provided is well-supported and credible.

## 2 Understanding AI Bias and Discrimination

### 2.1 Definition of AI Bias

Bias, as defined in the AI4Gov project, refers to *an inclination or prejudice for or against an individual or group that is considered unfair. This unfair bias can arise from personal experiences, societal norms and expectations, or information absorbed from various sources such as media, education, and family (Greenwald & Krieger, 2006). Bias can be categorized into two primary types: conscious (explicit) bias and unconscious (implicit) bias (Banaji & Greenwald, 2013). Both types can significantly impact individuals and society by perpetuating social inequalities and injustices.*

Conscious bias, also known as explicit bias, involves prejudices that individuals openly and deliberately maintain. These biases manifest through direct remarks, discriminatory behaviour, or policies that favour certain groups over others (Dovidio et al., 2002; Dovidio & Gaertner, 2004). Unconscious bias, or implicit bias, operates at a subconscious level and often goes unrecognized by those who hold such biases. It involves automatic associations and stereotypes based on characteristics like race, gender, or ethnicity (Greenwald & Banaji, 1995). This type of bias can lead to unequal opportunities in various contexts, such as education, healthcare, and the workplace (Fitzgerald & Hurst, 2017).

Furthermore, one can identify cognitive bias as: a systematic pattern of deviation from norm or rationality in judgment, whereby individuals create their own subjective reality based on their perception of information (Kahneman, 2011). These biases arise from various cognitive processes, including heuristics, social influences, and emotional factors, leading to errors in reasoning, evaluation, and decision-making. Cognitive bias can affect AI decision-making, embedding human prejudices into machine processes. Understanding and addressing these biases is essential to develop fairer and more accurate AI systems, thereby reducing AI-driven discrimination and promoting equity in various societal domains.

### 2.2 Impact of AI Discrimination

The impact of AI bias extends to various societal domains, contributing to discrimination and inequality. In education, AI systems influenced by bias can perpetuate disparities in student assessments and opportunities. In healthcare, biased AI algorithms can result in misdiagnosis or inadequate treatment for certain groups, exacerbating health inequalities (Panch et al., 2019; Nazer et al., 2023). In employment, biased AI systems can affect hiring and promotion decisions, leading to a lack of diversity and inclusivity in the workplace. In general, cognitive biases in AI can lead to decisions that are not in the best interest of users, perpetuating existing inequalities and missing valuable opportunities.

Therefore, it is crucial to recognize the existence of these biases and implement strategies to counteract them. This includes ensuring diverse and representative training data, conducting regular bias audits, and fostering an environment where multiple perspectives are considered. By actively seeking information that contradicts pre-existing beliefs and questioning assumptions

(Lilienfeld, Ammirati, & Landfield, 2009; Stanovich & West, 2008), AI developers can mitigate the impact of cognitive biases.

Bias in AI systems can originate from the data used to train these systems. If the training data reflects existing societal biases, the AI will likely replicate and amplify these biases in its outputs. For instance, if an AI system is trained on historical hiring data that includes biases against certain demographics, the system may continue to favour certain groups over others in its recommendations. Addressing AI bias involves ensuring that training data is representative and free from biases, alongside developing algorithms that can identify and mitigate bias during the decision-making process.

## 3 Preliminary Focus Groups with Pilot Cases Users

### 3.1 Description

This chapter synthesizes the key insights and recommendations derived from three focus group discussions during April 2024. All users of the pilot cases (JSI, VVV, MT and DPB) were invited to participate in order to discuss in free context on ethical aspects of their pilot cases and the platforms which are to be developed in the frames of the AI4Gov project. The discussions primarily focused on the use of AI platforms for evaluating projects in terms of bias, ethics, and alignment with standards such as legislation, recommendations or for example Sustainable Development Goals (SDGs). The discussions provided valuable insights into the challenges and opportunities of using AI platforms for evaluating projects and developing sustainable tourism policies. By addressing biases, improving evaluation methodologies, fostering collaboration, and ensuring ethical compliance, AI can significantly contribute to sustainable development and informed decision-making. The recommendations outlined aim to guide the development and implementation of AI platforms that are ethical, effective, and inclusive.

### 3.2 Main Findings

#### 3.2.1 AI and Sustainable Development Goals (SDGs)

The participants acknowledged the significant potential of AI in addressing SDGs, highlighting the necessity of a comprehensive evaluation framework to measure the impact of AI projects on sustainable development. Effective communication among stakeholders emerged as crucial to ensure the evaluation process captures the intended outcomes of AI projects. The importance of cross-disciplinary collaboration and consideration of cultural and geographical contexts in AI applications were emphasized as essential components for successful AI integration.

#### 3.2.2 Evaluation Methodologies

The discussions revealed the limitations of current evaluation methodologies, such as the GPT-4 framework, which need to be addressed to better capture the intended outcomes of AI projects. Participants proposed the development of more comprehensive evaluation frameworks that take into account the intended purpose of the AI application, stakeholder input, and the specific context in which the AI will be used. Incorporating key performance indicators (KPIs) to track the impact of AI projects on SDGs and ethical considerations was suggested as a way to enhance these methodologies. Tailoring the evaluation process to individual projects was emphasized to ensure relevance and effectiveness.

#### 3.2.3 Addressing Biases in AI

A significant portion of the discussions focused on understanding and addressing biases in AI. Participants stressed the need to redesign evaluation forms to better address and mitigate biases. They discussed various frameworks for AI ethics, including the UNESCO AI Ethics Framework and

the IEEE AI Ethics Framework, as tools to address biases. The conversation highlighted the technical and philosophical aspects of mitigating biases, including issues related to data collection, model building, and the developers' contexts. Addressing these aspects is critical to developing countermeasures against potential biases in AI solutions.

#### 3.2.4 AI Platform for Sustainable Tourism Policy

The discussions on the AI platform for sustainable tourism policy, specifically with the partners from VVV and MT in the early stages of the AI4GOV project, revolved around its objectives and potential applications. Participants highlighted that the platform should support data-driven decision-making for sustainable tourism by enabling the collection, analysis, and visualization of relevant data. The integration of the platform with existing systems and databases was considered essential for seamless communication and data exchange.

#### 3.2.5 Community Engagement

Engaging the community in the development and implementation of the AI platform was deemed crucial. Participants emphasized involving local stakeholders such as residents, businesses, and NGOs in the decision-making process facilitated by the AI platform. Educating the community on the benefits and potential risks of using AI in sustainable tourism policy was highlighted as a key step towards ensuring transparency and accountability in the platform's decision-making process.

#### 3.2.6 Human Factors and Infrastructure

The management of human factors and infrastructure related to the AI platform was another major topic of discussion. Ensuring that the AI platform is accessible and user-friendly for various stakeholders, including municipal staff and policymakers, was considered essential. Providing training and support to stakeholders on using the AI platform effectively and responsibly was also emphasized. Continuous monitoring and evaluation of the AI platform's performance and impact on sustainable tourism policy were recommended to make necessary adjustments and improvements over time.

#### 3.2.7 Ethical and Legal Considerations

The ethical and legal considerations associated with AI were a recurring theme throughout the discussions. Participants highlighted the importance of anonymizing personal data and complying with AI regulations to address ethical concerns. Avoiding discrimination based on gender, age, nationality, and other factors was emphasized, with a focus on creating comprehensive data characterizations to prevent bias from entering the system.

#### 3.2.8 Real-Time Data and Decision Support

The discussions also covered the importance of real-time data for informed decision-making. The installation of sensor-based systems, like All Oran, to monitor various aspects such as traffic and waste management was considered crucial for providing real-time data. Participants stressed the

need to balance automation with human oversight, ensuring that while AI can provide valuable support through information analysis and insights, final decisions should be made by human beings to ensure ethical and contextual appropriateness.

### 3.2.9 Recommendations

To develop comprehensive evaluation frameworks, it is recommended to create robust frameworks that consider the purpose, stakeholder input, and context of AI applications. Using KPIs to track and evaluate the impact of AI projects on SDGs and ethical considerations is necessary. Enhancing bias mitigation strategies involves redesigning evaluation forms and implementing established ethics frameworks to address biases effectively. Cross-disciplinary collaboration should be fostered to incorporate diverse perspectives in AI project evaluations and consider cultural and geographical contexts.

Engaging and educating stakeholders by involving local communities, businesses, and NGOs in the decision-making process facilitated by AI platforms is crucial. Educating stakeholders on the benefits and risks of AI can foster transparency and accountability. Ensuring ethical and legal compliance by prioritizing data privacy and preventing discrimination is essential. This can be achieved by creating comprehensive data characterizations and anonymizing personal data. Balancing automation with human oversight is recommended to use AI to support decision-making with real-time data and insights while ensuring final decisions are made by humans. Providing training and support to stakeholders is necessary for effective and responsible use of AI platforms. Implementing real-time monitoring systems, like sensor-based systems, can enhance decision-making in areas such as traffic and waste management.



## 4 1<sup>st</sup> Panel Discussions

In November 2023, AI4Gov organized a project event with two panel discussions followed by an educational workshop on the use of AI in public governance, exclusively for a Greek audience (held in Greek language)<sup>1</sup>.

More than 160 people participated in the hybrid event, including members of municipal councils and the academic community, experts and civil associations, while the event was welcomed by representatives of the Ministry of Digital Governance, as well as regional and local government.

The first panel discussion ran under the theme "Artificial Intelligence in Local Governance: Prospects and Challenges," and the second one was titled "Accessible and Inclusive Artificial Intelligence for Citizens".

The first panel provided valuable insights into the challenges and opportunities of using AI in local governance and public service delivery. Addressing biases, improving evaluation methodologies, fostering collaboration, and ensuring ethical compliance are critical to maximizing the benefits of AI. The second panel highlighted the importance of making AI accessible and inclusive for all citizens. By addressing digital barriers, enhancing support services, and promoting transparency and accountability, municipalities can ensure that AI technologies are used ethically and effectively to improve public services. The recommendations outlined in this report aim to guide the development and implementation of AI initiatives that are inclusive, fair, and beneficial to all citizens. Moreover, by following these recommendations, municipalities can effectively leverage AI technologies to improve governance and enhance the quality of life for their residents.

The main findings and the recommendations raised during these discussions are summarized in the following subsections.

### 4.1 “Artificial Intelligence in Local Governance: Prospects and Challenges” - Key Findings and Recommendations

#### 4.1.1 Digital Transformation and AI Implementation in Local Governance

The discussion highlighted the significant strides made in digital transformation by municipalities. A key focus was on the development of digital tools and platforms designed to enhance public service delivery and citizen engagement. One municipality developed a mobile application, which allows residents to report issues, access information, and request services directly from their smartphones. This app, available on both Android and iOS platforms, facilitates direct communication between citizens and municipal authorities, streamlining service requests and improving responsiveness.

---

<sup>1</sup> <https://ai4gov-project.eu/2024/01/03/parousiaseis-imeridas/>

Other digital initiatives included an online appointment booking system to reduce waiting times at municipal offices, an e-services portal that centralizes all digital services offered by the municipality, and the implementation of electronic payment systems for municipal fees. Additionally, a bike rental application, was developed to simplify the process of renting public bicycles. Efforts were also made to address the management of stray animals through a dedicated website, promoting adoption and organizing volunteer efforts.

These initiatives demonstrate a strong commitment to leveraging technology to improve public services and enhance citizen engagement, making daily interactions with municipal services more efficient and user-friendly.

#### 4.1.2 Challenges and Ethical Considerations

A critical theme of the discussion was the potential impacts and challenges of AI in public administration. While AI can significantly improve efficiency and automate processes, several important issues need to be addressed. The effectiveness of AI systems heavily relies on the quality and comprehensiveness of the data used for training. The lack of systematic codification of legislation in Greece, for example, hinders the development of reliable AI models. Ensuring the protection of personal data was also emphasized as paramount. AI systems must be designed to comply with data protection regulations and avoid unauthorized use of personal information.

Transparency and accountability were identified as crucial elements for public sector AI applications. AI systems should be publicly available for scrutiny to ensure they operate without biases and uphold ethical standards. There was a strong call for developing AI models that are transparent and can be audited to ensure they do not make decisions based on biases that could lead to discrimination. This includes ensuring diversity in data and continuously monitoring AI decisions.

#### 4.1.3 AI in Tourism Policy and Management

The application of AI in tourism policy was another key topic. The municipality of Vari-Voula-Vouliagmeni, characterized by its extensive coastline and high-quality tourist services, aims to develop data-driven policies to enhance tourism management. The focus was on using AI to monitor and analyse parking violations, predict areas with high violation rates, and optimize the allocation of municipal resources. In waste management, an AI system is being developed to track waste collection data, propose optimal collection routes, and potentially expand the "pay-as-you-throw" system. These initiatives aim to reduce operational costs, promote sustainable practices, and improve the overall quality of services provided to residents and visitors.

Success in these initiatives relies on the availability of reliable data from various sources, including the Ministry of Tourism and local data. The integration of AI tools is expected to improve policy design, enhance service delivery, and increase citizen satisfaction.

#### 4.1.4 Recommendations

To enhance data management and quality, it is recommended to develop standardized frameworks for data collection and codification, especially in legislative areas, to support AI training and implementation. Ensuring that data used for AI applications is accurate, comprehensive, and representative of diverse populations is essential to minimize biases.

Promoting transparency and accountability in AI systems involves making AI algorithms and models used in the public sector publicly available for examination and validation. Implementing mechanisms to continuously monitor and evaluate AI systems is necessary to ensure they adhere to ethical standards and operate without biases.

Strengthening data protection measures is crucial. AI systems should comply with data protection regulations to safeguard personal information. Stakeholders should be educated on the importance of data privacy and the ethical use of AI technologies.

Addressing bias and discrimination in AI models requires training AI models using diverse datasets to prevent biases and discriminatory outcomes. Regular audits of AI systems should be conducted to identify and mitigate potential biases in decision-making processes.

Fostering community engagement and education means involving local communities and stakeholders in the development and implementation of AI initiatives. Educating citizens and municipal staff on the benefits, risks, and ethical considerations of AI is crucial to build trust and ensure the effective use of technology.

## 4.2 “Accessible and Inclusive Artificial Intelligence for Citizens” - Key Findings and Recommendations

### 4.2.1 Digital Access and inclusivity

The discussion highlighted the critical need for digital access and inclusivity in public services. Participants shared a poignant example of a man in the UK who faced significant challenges in accessing his entitled benefits due to digital barriers. This example underscored the importance of ensuring that digital platforms are user-friendly and accessible to all, regardless of their digital literacy levels. In Greece, similar issues were identified, with many citizens needing assistance to navigate digital platforms for accessing social benefits.

One key initiative discussed was the introduction of the "Minimum Guaranteed Income" in 2017, which aimed to provide a safety net for those in extreme poverty. Despite the sophisticated digital platform for these applications, a significant number of citizens still required in-person assistance to complete their applications. This disparity highlighted the necessity of community services that support citizens in accessing digital tools.

#### 4.2.2 The Role of Community Centres

Community centres play a vital role in bridging the digital divide. These centres provide essential services, such as assisting with applications for various social benefits, including disability allowances, housing benefits, and birth allowances. They also offer food distribution, counselling, and integration into other social programs. The panel emphasized the increasing importance of these centres as more services move online.

Moreover, community centres have been instrumental in offering psychological support, particularly in managing stress, which can be a barrier to learning new technologies. This support is not only crucial for older adults but also for younger people who may lack the confidence to use digital tools effectively.

#### 4.2.3 Education and Training Programs

The need for ongoing education and training was a recurrent theme. Community centres and municipalities have initiated various programs to teach basic computer skills and digital literacy. For instance, there are initiatives aimed at training elderly citizens on the use of computers and mobile devices. Additionally, there are programs to help citizens understand and navigate digital government services.

For refugees and migrants, language barriers present an additional challenge. Programs have been developed to teach Greek to non-native speakers, helping them integrate into society and access digital services. These programs have been recognized for their effectiveness and have received accolades for their contribution to social inclusion.

#### 4.2.4 Digital Transformation in Municipal Services

Municipalities have made significant strides in digital transformation. Efforts include developing applications for reporting issues like potholes and fallen trees, as well as platforms for managing urban planning information. These digital tools aim to improve efficiency and make it easier for citizens to interact with municipal services.

One notable project is the creation of a digital assistant that can handle frequently asked questions from citizens, reducing the burden on municipal staff and improving service delivery. This tool uses AI to provide accurate and timely information, enhancing the overall user experience.

#### 4.2.5 Challenges and Ethical Considerations

The panel also discussed the challenges and ethical considerations of implementing AI in public services. Ensuring data privacy and protection is paramount, especially when dealing with sensitive personal information. There is also a need to address biases in AI algorithms to prevent discrimination and ensure fairness in decision-making processes.

Participants highlighted the importance of transparency and accountability in AI systems. Public sector AI applications should be designed to be transparent, with clear explanations of how

decisions are made. This transparency helps build trust among citizens and ensures that AI is used ethically and responsibly.

#### 4.2.6 Youth Engagement and Digital Skills

Engaging the youth and enhancing their digital skills was another significant focus. Municipal youth councils have been established to involve young people in local governance. These councils provide a platform for young citizens to voice their concerns, propose solutions, and participate in decision-making processes.

Youth councils have also embraced digital tools to facilitate their activities. For example, e-petitions and live chat platforms have been introduced to gather input from young people and address their concerns in real-time. These tools help ensure that the voices of young citizens are heard and considered in municipal policies.

#### 4.2.7 Recommendations

To enhance digital access and inclusivity, it is recommended to develop more user-friendly digital platforms and provide comprehensive support services for those who need assistance. Community centers should be strengthened to offer a broader range of digital literacy programs and support services.

Ensuring transparency and accountability in AI systems is crucial. Public sector AI applications should be designed with clear explanations of decision-making processes and mechanisms for regular auditing and monitoring to prevent biases and ensure fairness.

Ongoing education and training programs are essential to equip citizens with the necessary digital skills. These programs should be tailored to meet the needs of different demographic groups, including elderly citizens, refugees, and migrants.

Youth engagement should be a priority, with continued support for youth councils and the development of digital tools that facilitate their participation in local governance. Providing young people with opportunities to contribute to decision-making processes helps build a more inclusive and responsive local government.

## 5 2<sup>nd</sup> and 3<sup>rd</sup> Panel Discussions

In June 2024 the AI4Gov partners organized and participated in two discussion panels in Ljubljana, Slovenia, within two different events – international conferences, one in the domain of healthcare and medicine, and the other in the domain of law and regulation:

The 2<sup>nd</sup> discussion panel took place during the international conference in Ljubljana: "Dialogues in Neurodegenerative Disorders: Care, communication and biomedical challenges" (Neurocare)<sup>2</sup>, 6<sup>th</sup> - 7<sup>th</sup> June 2024. The AI4Gov partners (JSI and UPRC) organized a panel discussion titled "Rethinking Healthcare with Digital Health Technology," where AI in healthcare and trustworthiness were the main focus of the discussion.

The 3<sup>rd</sup> discussion panel was organized under the international conference titled "Global Conference on AI And Human Rights"<sup>3</sup>, which took place at the Faculty of Law, University of Ljubljana, 13<sup>th</sup> and 14<sup>th</sup> June 2024, organized under the patronage of UNESCO. Specifically, the AI4Gov discussion panel was held after panel 3 on the first day of the conference, which was titled "AI and governance". After the presentations of the invited speakers the floor was open for a discussion between the speakers and the audience. Speakers received questions from the panel chair (former Slovenian Minister for Public Administration) and from the audience. There were five panelists from different domains: an AI researcher, a social sciences researcher, an UNESCO representative, a researcher from the domain of human rights and democracy, and a researcher from the domain of law. The audience was mixed, most representatives being from the domain of law and regulation, but also from education, policy institutions and also some from the industry.

The discussion brought to light several critical issues and considerations regarding the integration of AI in various aspects of society, with a particular focus on governance, bias, discrimination, transparency, inclusivity, and democratic participation.

### 5.1 2<sup>nd</sup> Panel Discussion - Rethinking Healthcare with Digital Health Technology: Key findings

The panel discussion underscored the transformative potential of AI in enhancing public services and healthcare delivery. However, realizing this potential requires addressing regulatory, ethical, and educational challenges. By following the recommendations outlined in this report, stakeholders can ensure that AI technologies are implemented in a manner that is safe, effective, and inclusive, ultimately benefiting all citizens. More specifically the points raised were:

- Enhancement of Regulatory Frameworks: Develop comprehensive regulatory frameworks that ensure AI tools used in healthcare are certified for accuracy and reliability. These frameworks should be dynamic to keep pace with rapid technological advancements.

---

<sup>2</sup> <https://www.neurocare.si/>

<sup>3</sup> <https://www.ai-right-to-life.si/en/2014-conference>

- Promotion of Transparency and Explainability: Implement AI systems that offer clear explanations of their decision-making processes. This transparency builds trust among users and facilitates wider adoption of AI tools in critical areas like healthcare.
- Strengthening Ethical and Privacy Protections: Ensure that AI applications comply with stringent data protection regulations. Develop AI tools specifically for medical use, validated through rigorous testing to ensure their safety and reliability.
- The Need to Improve Digital Literacy: Invest in training programs for healthcare professionals to enhance their understanding of AI. Similarly, develop initiatives to increase patients' digital literacy, enabling them to engage effectively with AI-driven healthcare solutions.
- Fostering Critical AI Literacy: Encourage the development of critical AI literacy to help users understand the strengths and limitations of AI. This education is crucial for professionals who rely on AI for decision-making, ensuring they can use these tools responsibly and effectively.
- Ensuring Inclusive AI Applications: Design AI systems with accessibility and inclusivity in mind to ensure they meet the needs of all citizens, particularly marginalized and vulnerable populations. This approach helps in addressing inequalities and enhancing the overall effectiveness of public health and social services.

### 5.1.1 AI in Healthcare and Regulatory Challenges

The discussion emphasized the increasing integration of AI in healthcare and the regulatory challenges that accompany this trend. AI tools are being used for various purposes, such as predicting medical outcomes, diagnosing diseases, and managing administrative tasks. However, the deployment of these tools requires stringent regulation to ensure their safety and effectiveness. AI models used in medicine need to be certified to guarantee they meet specific accuracy and reliability standards. This certification process is crucial because medical solutions are classified as high-risk applications under the AI Act, necessitating external evaluation and stringent oversight.

### 5.1.2 Transparency and Trust in AI Systems

A significant concern discussed was the need for transparency and trust in AI systems, especially in medical contexts. Participants highlighted that while AI can assist in diagnostic processes, the final decision should always rest with a human (healthcare expert). This ensures that AI serves as a supportive tool rather than a replacement for human judgment. The importance of explainable AI was stressed, where the AI system's decision-making process is transparent and understandable to its users, particularly healthcare professionals. This transparency helps in building trust among users and ensures that AI tools are adopted more widely and effectively.

### 5.1.3 Impact of AI on Doctor-Patient Relationship

The panel discussed the potential impact of AI on the doctor-patient relationship. There is a concern that doctors might become too reliant on AI systems without fully understanding their underlying mechanisms. This reliance could undermine the doctor-patient relationship if not managed properly. Ensuring that doctors are well-informed about how AI tools work and also about their limitations is crucial. This knowledge empowers them to use AI effectively while maintaining their critical role in patient care.

### 5.1.4 Ethical and Privacy Concerns

Ethical and privacy concerns were prominent in the discussion. The use of AI in healthcare involves handling sensitive personal data, which must be protected to comply with data protection regulations like GDPR. There was a debate on the appropriateness of using general AI tools, such as ChatGPT, for medical diagnostics due to concerns about data privacy and the accuracy of the information provided by such tools. Participants stressed the need for AI tools specifically designed and validated for medical use to ensure they provide reliable and safe outcomes.

### 5.1.5 Digital Literacy and Training

The panel highlighted the importance of digital literacy and ongoing training for both healthcare professionals and patients. Educating healthcare professionals about AI and its applications can help them integrate these tools into their practice effectively. Similarly, increasing patients' digital literacy ensures they can understand and engage with AI-driven healthcare solutions. This dual approach helps in maximizing the benefits of AI while mitigating risks associated with its use.

### 5.1.6 AI's Role in Public Health and Social Services

AI's potential in public health and social services was also discussed. AI can enhance the efficiency of these services by automating routine tasks, analysing large datasets for public health insights, and providing personalized care recommendations. However, the integration of AI in these areas must be done thoughtfully to ensure it addresses the needs of all citizens, including marginalized and vulnerable populations. Ensuring accessibility and inclusivity in AI applications is critical to avoid exacerbating existing inequalities.

### 5.1.7 Critical AI Literacy

Developing critical AI literacy among users was deemed essential. Understanding the capabilities and limitations of AI helps users make informed decisions and use these tools effectively. This literacy is particularly important for those in decision-making roles, such as doctors and policymakers, who need to critically assess AI tools and their outputs.



## 5.2 3<sup>rd</sup> Panel Discussion – AI and Governance

The discussion during the conference highlighted several major findings:

- First, there is a need for adaptable regulations to prevent the misuse of AI and ensure its ethical implementation.
- Second, continuous and multidisciplinary education is crucial for understanding and responsibly handling of AI.
- Third, safeguarding digital rights is essential to empower users and maintain trust in AI systems. Fourth, ensuring AI systems are transparent and explainable is necessary to prevent biases and foster inclusivity.
- Lastly, raising public awareness and involving the public in AI decision-making processes are vital for aligning AI developments with democratic principles and societal values.

### 5.2.1 AI in Governance

The conference underscored the importance of ethical impact assessments, illustrated by a pilot project in Chile focused on social security and pensions. This project highlighted the need for training authorities to ensure they are competent in using AI tools effectively. The discussion also touched upon the manipulation of political data and elections through algorithms and computational propaganda. Notable instances, such as the 2016 US elections and Brexit, serve as stark reminders of how AI can be used to polarize societies. To address these challenges, there is a pressing need for regulations that can swiftly adapt to the fast-paced advancements in AI technology.

### 5.2.2 Bias and Discrimination

One of the major concerns discussed was the potential for AI to perpetuate bias and discrimination. The impact of biases, deepfakes, and disinformation in critical sectors such as healthcare and justice were emphasized. The conference highlighted the importance of multidisciplinary education and continuous learning to predict and understand the impact of AI, which is essential for addressing these biases effectively.

### 5.2.3 Transparency and Inclusivity

The conference identified five key dimensions of AI ethics: legal, socio-cultural, scientific/educational, economic, and technological/infrastructure. Understanding these dimensions is crucial for addressing the ethical implications of AI. Additionally, the discussion on digital rights emphasized the need for the right to be forgotten, anonymity, self-determination, and encryption to ensure user privacy and control. The challenge of making AI systems explainable across different levels of education was also highlighted. Ensuring transparency in AI systems is necessary to build trust and foster inclusivity.

#### 5.2.4 Raising Awareness and Participation in Decision-Making and Democracy

Education was a recurring theme, with recommendations to integrate it closely with governance and human rights. Developing educational curricula that keep pace with technological advancements is vital for preparing both authorities and the general public for responsible AI use. Digital literacy programs for children and schools were emphasized as essential for fostering future generations capable of managing AI responsibly. The role of users, both critical and technical, in understanding and managing AI systems was also discussed. Furthermore, raising awareness about the impact of AI on democracy and encouraging active public participation in AI-related decision-making processes are crucial for ensuring that AI developments align with democratic values.

## 6 4<sup>th</sup> and 5<sup>th</sup> Panel Discussions

All AI4Gov partners working under the T5.1 continued their active participation in international conferences by organizing two additional discussion panels in 2024, that explored the intersection of AI with governance, politics, and data analysis. These panels, the fourth and fifth in the series of discussions under the AI4Gov project, were held in Dublin and Thessaloniki, respectively, and brought together experts, researchers, and practitioners to address the ethical, methodological, and practical dimensions of AI in public administration and society.

The 4<sup>th</sup> panel, titled "Ethical AI and Democratic Governance: Challenges and Innovations in Political Methodology," was organized by the Aristotle University of Thessaloniki (AUTH) under the ECPR General Conference 2024, hosted by the University College Dublin from 12<sup>th</sup> to 15<sup>th</sup> August 2024. Held on the third day of the conference, this panel was part of the section on Political Methodology. It featured four discussants and an engaged audience and aimed to explore how AI technologies are reshaping governance and the conduct of political affairs. The panel delved into topics such as transparency, bias, inclusivity, and the integration of AI in public policy and administration, highlighting both the challenges and opportunities posed by these emerging technologies.

The 5<sup>th</sup> panel took place under the 12<sup>th</sup> Panhellenic Conference for Data Analysis with international participation, held in Thessaloniki from 3<sup>rd</sup> to 6<sup>th</sup> October 2024. Organized again by AUTH, this panel was conducted on the second day of the conference and was titled "Methods and Tools of Analysis in Artificial Intelligence: Implications from their Application." With four discussants, the panel examined the technical and ethical implications of AI methods and tools, focusing on their impact in real-world applications. Discussions spanned the development of AI methodologies, their practical use in various domains, and the broader societal effects of these technologies, emphasizing the need for responsible and ethical deployment.

Both panels provided unique platforms to engage with critical issues at the intersection of AI, governance, and society. They underscored the importance of interdisciplinary dialogue in addressing the challenges of transparency, accountability, and fairness in AI applications. These discussions not only advanced the academic and practical understanding of AI in public administration but also highlighted the essential role of inclusivity and ethical considerations in shaping the future of AI governance and technology.

### 6.1 4<sup>th</sup> Panel Discussion - Ethical AI and Democratic Governance: Challenges and Innovations in Political Methodology

The 4<sup>th</sup> panel discussion explored the transformative role of AI in governance, policymaking, and political communication. Participants had the opportunity to discuss the ethical and technical dimensions of integrating AI into political methodologies, emphasizing the profound implications for transparency, inclusivity, and public trust in democratic processes. The session addressed both the opportunities AI presents and the critical challenges it raises, ranging from regulatory complexities to the potential risks of misuse in undermining democratic values.

One of the central themes of the discussion was the application of AI in enhancing policymaking efficiency and effectiveness. AI technologies can analyse vast amounts of data, forecast outcomes, and provide evidence-based insights to guide decisions. However, these advancements are accompanied by ethical concerns, including issues of bias, data privacy, and the accountability of AI-driven systems. The panellists underscored the importance of developing robust regulatory frameworks, such as the Holistic Regulatory Framework (HRF) being developed under the AI4Gov Horizon project, to ensure that AI tools adhere to principles of fairness, transparency, and non-discrimination.

Another key focus was the use of AI in managing misinformation and disinformation, particularly in political contexts. While AI-driven tools like large language models (LLMs) offer innovative solutions for fact-checking and combating false information, challenges remain regarding biases in training data, inconsistent performance across topics, and the transparency of AI operations. The panel also highlighted the need for continuous research and refinement to address these limitations.

The discussion further emphasized the critical role of public engagement and civic participation in the successful implementation of AI in governance. Engaging citizens in the design and evaluation of AI tools can enhance trust and ensure that these systems reflect societal values. Participants also discussed the disparities in the capacity of local governments to adopt AI technologies and the necessity of building local governance capacities to foster equitable access and innovation.

Finally, the panel examined the broader implications of AI for democracy. The rapid pace of technological advancements often outstrips the ability of policymakers to respond effectively, creating an urgent need for proactive governance strategies. By addressing these challenges collaboratively, stakeholders can leverage AI to enhance democratic participation, improve transparency, and safeguard ethical standards in governance.

### 6.1.1 Ethical AI in Governance and Policy Making

The panel emphasized the increasing integration of AI into political methodologies, particularly in governance and policymaking, discussing both its transformative potential and the ethical challenges it presents. AI has proven to be an asset in analysing vast datasets, generating actionable insights, and forecasting outcomes with improved accuracy. These capabilities enable policymakers to design evidence-based policies that address societal needs more effectively. For example, AI tools have been utilized to predict economic trends, manage urban planning, and even identify patterns in public health data to pre-empt crises.

Despite these advancements, panellists underscored the ethical dilemmas surrounding AI use in governance. Privacy concerns emerged as a recurring theme, particularly in the collection, storage, and anonymization of data. Many governments collect data from citizens to inform policy decisions, but inadequate measures to secure this data can lead to breaches or misuse. Furthermore, the process of obtaining informed consent from citizens was identified as a persistent challenge, especially in sensitive contexts where individuals may not fully understand

how their data will be used. This lack of clarity often results in distrust toward AI-driven governance tools.

Transparency in AI systems was highlighted as an essential component of ethical implementation. Participants noted that AI systems used in governance often operate as "black boxes," making it difficult for users—even those deploying the systems—to understand how decisions are made. This opacity undermines public trust and raises concerns about accountability. To address this, the panel called for the adoption of explainable AI frameworks, ensuring that decision-making processes are clear and accessible to both policymakers and the public.

### 6.1.2 The Role of Regulatory Frameworks

Discussion emphasized the critical role of regulatory frameworks in ensuring the ethical deployment of AI in governance. The "Holistic Regulatory Framework" (HRF), being developed under the AI4Gov project was introduced to the audience as a key initiative designed to address fairness, inclusivity, and non-discrimination in AI tools. This framework seeks to ensure that AI systems are developed and deployed in compliance with legal standards, such as the General Data Protection Regulation (GDPR). Panellists stressed that maintaining human oversight in AI-driven decision-making processes is fundamental to preserving accountability.

Participants acknowledged that regulatory frameworks must be dynamic, adapting to the rapid evolution of AI technologies. Traditional, static approaches to regulation are insufficient in a landscape where new capabilities and challenges emerge frequently. The HRF proposes a model that emphasizes regular auditing and revision of regulations, fostering a collaborative environment between technologists, legal experts, and policymakers. By prioritizing transparency and explainability, the framework aims to bridge the gap between technological innovation and ethical governance.

### 6.1.3 Public Engagement and Civic Participation

AI's role in enhancing civic participation was a central topic of discussion. The panel explored how well-designed AI tools can foster greater transparency and trust in governance, leading to increased public engagement in political processes. For instance, AI-powered platforms that aggregate public opinion or facilitate participatory budgeting have shown promise in creating more inclusive governance structures.

However, challenges remain in achieving meaningful civic participation, particularly in regions with centralized governance systems and limited local autonomy. In such contexts, citizens often lack the channels or resources to influence decision-making processes meaningfully. The panel suggested that governments invest in training programs to improve citizens' digital literacy and understanding of AI tools. Such initiatives can empower individuals to engage with AI-driven platforms confidently and responsibly, ensuring their voices are heard. The discussion also touched on the importance of inclusive design in AI systems. Tools must be accessible to diverse user groups, including those with limited technical expertise. By prioritizing usability and inclusivity, AI can serve as a bridge between citizens and their governments, fostering a more engaged and informed electorate.

#### 6.1.4 Addressing Misinformation and Disinformation

The proliferation of misinformation and disinformation in political communication was identified as a pressing issue. AI tools, particularly large language models (LLMs) like GPT-4, offer innovative solutions for fact-checking and identifying false information. These systems can analyse vast amounts of text, flagging inconsistencies or inaccuracies that may indicate disinformation campaigns. Despite these capabilities, limitations remain. Panellists pointed out that biases in training data can affect the accuracy and reliability of these tools. Moreover, LLMs often lack the contextual understanding necessary to navigate complex or nuanced topics, resulting in inconsistencies in performance. The opacity of these systems also complicates efforts to validate their findings, raising questions about their trustworthiness.

Participants called for continued research and development to refine AI tools for combating disinformation. Collaborative efforts between technologists, media organizations, and policymakers are essential to create robust systems capable of addressing the dynamic nature of misinformation campaigns. Additionally, public education on identifying and avoiding false information was recommended as a complementary strategy.

#### 6.1.5 Challenges in Local Governance and Capacity Building

The discussion highlighted significant disparities in the capacity of local governments to adopt and utilize AI tools. Many municipalities rely heavily on central governments for funding and technical expertise, limiting their ability to innovate independently. This dependency often results in a lack of localized policymaking strategies, reducing opportunities for civic engagement at the community level.

Panellists emphasized the need for targeted capacity-building initiatives to address these challenges. Providing local governments with access to training, funding, and technical support can empower them to implement AI solutions tailored to their specific needs. Decentralizing governance structures was also proposed as a means of fostering innovation and enabling more responsive policymaking at the local level.

#### 6.1.6 Speed of Technological Advancements vs. Policy Response

A recurring concern was the rapid pace of AI advancements compared to the slower response of policymakers and regulatory bodies. Unregulated AI tools pose significant risks to democracy and governance, including potential misuse for surveillance, bias in decision-making, and erosion of public trust. The panel called for accelerated policy development processes to address these risks proactively. Collaboration between technologists and policymakers was identified as a critical factor in bridging this gap. By fostering open dialogue and shared understanding, stakeholders can develop regulations that balance innovation with ethical considerations, ensuring that AI serves as a force for positive change in governance.

### 6.1.7 Recommendations

The panel discussion highlighted the immense potential of AI to revolutionize governance and enhance democratic processes. However, realizing this potential requires addressing significant ethical, technical, and social challenges. The following recommendations are designed to guide policymakers, technologists, and stakeholders in implementing AI systems that are transparent, inclusive, and aligned with democratic values:

- Dynamic Regulatory Frameworks: Regulatory frameworks must evolve to keep pace with technological advancements. Initiatives like the Holistic Regulatory Framework (HRF) should be continuously updated to incorporate new legal and technological developments. These frameworks must ensure that AI tools adhere to principles of human oversight, fairness, inclusivity, and transparency, safeguarding individual rights and promoting public trust.
- Public Engagement and Education: It is essential to invest in educational and training programs that empower citizens to understand and engage with AI tools effectively. Public participation in the design and evaluation of AI systems should be encouraged, fostering a sense of ownership and ensuring that these tools reflect societal values and needs.
- Capacity Building in Local Governance: Many local governments face limitations in adopting and utilizing AI tools independently. Strengthening the capacity of municipalities through funding, training, and decentralization of governance structures can enable them to develop tailored, context-specific solutions that address local challenges.
- Advancing Research in AI and Misinformation: AI tools hold promise in combating misinformation, but their limitations must be addressed. Ongoing research should focus on improving the reliability of these tools, reducing biases, and ensuring their performance across diverse contexts and topics. This will enhance their effectiveness in supporting informed decision-making and public communication.
- Fostering Multidisciplinary Collaboration: The integration of AI in governance requires collaboration between political scientists, technologists, and policymakers. Such interdisciplinary efforts ensure that AI tools are technically robust while remaining politically ethical and aligned with democratic ideals. Regular feedback from citizens and end-users should guide the design and refinement of these systems.
- Accelerating Policy Adaptation to Technology: Policymakers need to adopt agile approaches to regulatory development to match the rapid pace of AI advancements. Proactive strategies and international cooperation can help create a unified response to the global implications of AI, ensuring its responsible and ethical use in governance.

By focusing on these areas, stakeholders can harness the power of AI to strengthen democratic institutions, enhance transparency in decision-making, and foster greater civic participation. These recommendations serve as a foundation for integrating AI in governance in a manner that prioritizes ethical considerations and the public good.

## 6.2 4<sup>th</sup> Panel Discussion - Ethical and Practical Dimensions of Artificial Intelligence in Public Administration and Society: Key Findings

The panel discussion, titled "The Ethical and Practical Dimensions of Artificial Intelligence in Public Administration and Society," explored the transformative potential of AI across various domains, including governance, healthcare, justice, and public safety. Panelists examined the profound opportunities AI offers to improve efficiency, enhance decision-making, and optimize resource allocation in public administration, while addressing the ethical, social, and technical challenges it poses.

A central theme of the discussion was the role of AI in modernizing public services, with a particular emphasis on automating routine tasks and leveraging predictive analytics to design data-driven policies. These advancements were recognized as critical to enhancing the speed and effectiveness of public service delivery. However, the panelists stressed that such progress must be accompanied by robust mechanisms to ensure transparency, accountability, and the safeguarding of citizen rights.

The ethical challenges surrounding AI systems, including algorithmic bias, data privacy concerns, and the lack of explainability in decision-making processes, were at the forefront of the conversation. Participants underscored the importance of establishing regulatory frameworks that promote fairness and inclusivity while enabling innovation. The European Union's AI Act was discussed as a pioneering model for regulating high-risk AI applications, though its implementation on a global scale presents significant challenges.

Another key focus was the importance of data governance, with panelists emphasizing the need for high-quality, representative datasets that mitigate bias and promote equitable outcomes. Disparities in access to resources and technological capabilities between regions and communities were highlighted, revealing the potential for AI to exacerbate existing inequalities if not implemented inclusively.

In healthcare and public safety, AI's transformative capabilities were evident in areas such as diagnostics, robotic surgery, and surveillance systems. While these tools promise substantial benefits, their deployment raises critical questions about security, reliability, and ethical use. Panelists emphasized the need for rigorous oversight and the development of sector-specific guidelines to ensure responsible implementation.

Education and public awareness emerged as essential elements for bridging the gap between AI developers, policymakers, and society at large. Enhancing digital literacy and fostering an informed citizenry were seen as key to building trust and ensuring that AI technologies are adopted in ways that align with democratic values.

The panel concluded with a call for multidisciplinary collaboration and proactive governance to address the challenges posed by AI while harnessing its potential to transform public administration and society for the better. This report captures the diverse insights and actionable recommendations shared during the discussion.



### 6.2.1 The Transformative Role of AI in Governance

AI's influence in public administration is undeniable, offering tools that streamline tasks like document management, resource allocation, and public service delivery. Participants shared examples of AI applications in areas such as contract management, predictive analytics for urban planning, and automation of routine government workflows. These technologies reduce administrative burdens and allow governments to focus on higher-value activities. However, several panellists warned of the risks associated with over-reliance on AI in governance, particularly when it comes to ensuring human oversight and contextual understanding in decision-making processes.

Moreover, while AI has the potential to make services more accessible and efficient, concerns about its uneven implementation were raised. Smaller municipalities and regions often lack the resources to adopt AI at scale, leading to a digital divide that risks deepening existing inequalities in public service delivery.

### 6.2.2 Ethical Challenges in AI Implementation

A recurring theme was the ethical dilemmas posed by AI. Algorithmic decisions are increasingly impacting citizens' lives, from eligibility for social benefits to judicial outcomes. Panellists discussed the "black box" nature of many AI systems, where the internal logic of decision-making is not accessible or understandable to users, even those deploying the systems. This lack of transparency erodes public trust and creates significant accountability gaps.

In addition, ethical concerns surrounding data privacy were extensively debated. Governments and private entities deploying AI are often entrusted with sensitive personal data. Without robust safeguards, there is a risk of data misuse, leaks, or unauthorized surveillance. These risks are particularly acute in contexts such as healthcare and law enforcement, where data breaches can have serious implications for individuals and communities.

### 6.2.3 Data Governance and Accessibility

The discussion repeatedly returned to the critical role of high-quality, representative data in the success of AI systems. Panellists highlighted the importance of democratizing access to public datasets to enable innovation and equity in AI development. However, concerns were raised about the dominance of proprietary models and datasets controlled by large corporations, which limit opportunities for smaller players to participate in AI development. Participants also emphasized that datasets must reflect the diversity of populations they aim to serve. Non-representative datasets risk perpetuating biases, leading to outcomes that disadvantage already marginalized communities. This was illustrated with examples of facial recognition systems that perform poorly on certain demographic groups, underscoring the need for rigorous standards in data collection and validation.

#### 6.2.4 AI in Healthcare and Public Safety

Healthcare emerged as one of the most promising yet challenging domains for AI application. Panellists discussed its transformative potential in areas like diagnostic imaging, robotic surgery, and patient monitoring. For example, AI tools can help identify diseases earlier and with greater accuracy than traditional methods, saving lives and reducing healthcare costs. However, integrating these tools into medical practice requires addressing concerns about data security, model reliability, and the implications of machine error in high-stakes scenarios.

In public safety, AI-driven surveillance systems and predictive policing tools were highlighted as controversial applications. While these technologies can enhance crime prevention efforts, they also raise significant ethical concerns regarding their potential misuse for profiling and discrimination. Panellists agreed that such tools must be carefully regulated and deployed only with strict oversight and accountability mechanisms in place.

#### 6.2.5 Regulatory and Policy Frameworks

The importance of regulatory frameworks that balance innovation with protection of citizen rights was a central focus. The European Union's AI Act was discussed as a pioneering effort to regulate AI comprehensively, categorizing applications by risk levels and imposing stricter requirements on high-risk systems. However, the panellists acknowledged the difficulty of implementing such frameworks globally, given the varying capacities and priorities of different nations. Speakers advocated for adaptive regulations that evolve alongside AI technologies, emphasizing the need for continuous monitoring and revision of policies. This approach ensures that regulations remain relevant and effective without stifling innovation.

#### 6.2.6 Education and Public Awareness

Panellists stressed that education is critical to bridging the gap between AI developers, policymakers, and the general public. Public understanding of AI's capabilities and limitations is often limited, leading to mistrust or unrealistic expectations. To address this, participants called for targeted initiatives to enhance digital literacy, such as workshops for government officials, training programs for educators, and public awareness campaigns. Educational efforts should also focus on equipping policymakers with the technical knowledge needed to evaluate and oversee AI systems. Without this understanding, the risk of poorly informed regulations increases, potentially hindering the beneficial adoption of AI technologies.

#### 6.2.7 Addressing Algorithmic Bias

Algorithmic bias was a key concern throughout the discussion. Panellists shared examples of AI systems that unintentionally reinforce existing stereotypes or systemic inequalities, such as recruitment algorithms that disadvantage women or minorities. To combat these issues, the development of fairness metrics and diversity audits was recommended. Regular evaluations of AI systems can help identify and mitigate biases, ensuring equitable outcomes.

### 6.2.8 Recommendations

AI's integration into public administration and society must be guided by principles of fairness, transparency, and inclusivity. The panel proposed several actionable steps to achieve this:

- Ensure Transparency and Accountability: AI systems must include mechanisms for explainable decision-making to foster trust and enable meaningful oversight. Transparency is especially vital in high-stakes applications, such as healthcare and criminal justice.
- Improve Data Quality and Access: Governments should promote open data initiatives, ensuring datasets are diverse, accurate, and accessible to a wide range of developers. This approach reduces the risk of bias and democratizes AI innovation.
- Develop Adaptive Regulatory Frameworks: Regulations should be flexible and responsive to the rapid pace of AI advancements. The inclusion of multidisciplinary advisory boards can help ensure balanced and informed policy development.
- Enhance Public and Policymaker Education: Investment in digital literacy and technical training is essential for both citizens and policymakers. These programs should emphasize critical thinking about AI's societal impact and ethical considerations.
- Encourage Ethical AI Practices: Incentivize the development of AI systems that prioritize social good, with ethical impact assessments integrated into the design and deployment processes.
- Address Socioeconomic Disparities: Targeted initiatives should support under-resourced municipalities and communities in adopting AI, ensuring equitable access to its benefits.

## 7 Existing Awareness Strategies

Awareness strategies to address AI discrimination and mitigation of bias are crucial for fostering a fairer and more equitable deployment of AI technologies. These strategies encompass a range of educational, policy-oriented, and technological approaches designed to identify, understand, and mitigate biases in AI systems. These strategies include educational programs and workshops, formulation of ethical guidelines and frameworks, conduct of regular bias audits and impact assessments, transparency and explainability in AI systems promotion by corresponding initiatives, encouraging diversity and inclusion within AI development teams, public awareness campaigns and collaborative initiatives between governments, academia, industry, and civil society. These strategies aim to create a comprehensive framework for understanding and mitigating AI bias, ensuring that AI technologies are developed and deployed in a manner that promotes fairness, inclusivity, and accountability. By continuing to expand and refine these strategies, stakeholders can work towards minimizing AI-driven discrimination and fostering a more equitable technological landscape.

### 7.1 Educational Campaigns

Educational initiatives are fundamental in raising awareness about AI discrimination and bias. Universities, research institutions, and organizations offer courses, workshops, and training programs aimed at AI developers, policymakers, and the general public. These programs focus on the ethical implications of AI, the sources of bias, and techniques for mitigating bias in AI systems. By enhancing understanding and skills, these educational efforts empower stakeholders to recognize and address bias effectively. To name some of these initiatives, the Oxford University and The Oxford Internet Institute offer courses on the ethical and governance aspects of AI, focusing on understanding and addressing biases in AI systems. The courses cover a wide range of topics, including the social implications of AI and strategies for mitigating bias. Delft University of Technology in the Netherlands offers a course on responsible innovation, which includes modules on AI ethics and fairness. The University of Edinburgh offers a course that explores the impact of AI on society, including issues of fairness, bias, and discrimination. There are also workshops and training sessions organized by various institutions and tech companies focusing on identifying and mitigating biases in AI systems and covering best practices for ensuring fairness and transparency in AI applications such as the ETH Zurich – AI Ethics Lab, SAP (German software technology company), Google and Microsoft.

#### 7.1.1 Public Awareness Campaigns

Public awareness campaigns play a vital role in educating society about the potential risks and biases associated with AI. These campaigns utilize various media platforms to disseminate information, highlight real-world examples of AI bias, and promote informed discussions on the ethical use of AI. By raising public consciousness, these campaigns encourage a collective effort to advocate for fair and unbiased AI systems. Two prominent examples of public awareness campaigns are the Algorithmic Justice League (AJL) which was designed to raise awareness about

the social implications of AI bias through advocacy, art, and research, and the AI for Good Global Summit which was organized by the International Telecommunication Union (ITU) and XPRIZE with the aspiration to bring together AI innovators, ethicists, and policymakers to discuss and address the ethical challenges of AI, including bias and discrimination.

### 7.1.2 Social Media

To effectively address AI discrimination and mitigate bias on social media, a multi-faceted approach is required. Awareness strategies should focus on educating both developers and users about the inherent biases that AI systems can perpetuate. Training programs for AI developers and educational campaigns for users can increase awareness of how AI-driven content recommendations might reinforce existing biases. Transparency in AI algorithms is crucial; platforms should disclose how their AI systems function and the data they utilize, enabling independent audits and public scrutiny (Binns, 2018).

Social media platforms can implement features that allow users to understand why certain content is being recommended, promoting greater transparency. Additionally, platforms should employ bias detection tools that continuously monitor AI systems for discriminatory patterns and adjust algorithms to minimize bias (Gillespie, 2020).

There are several awareness campaigns and initiatives currently active to address AI discrimination and mitigate bias on social media such as Reclaim Your Face, a European civil society initiative which aims to ban biometric mass surveillance. It focuses on raising awareness about the misuse of facial recognition technology and other AI systems that can perpetuate discrimination and bias. The campaign involves a coalition of over 60 organizations advocating for stronger regulations to protect fundamental rights and ensure transparency in AI applications (D-CENT). Another example of an initiative to address discrimination on social media was implemented by the social media platform Facebook. In response to concerns about racial bias in its ad targeting algorithms, Facebook conducted a civil rights audit. This audit led to changes in how the platform handles targeted advertising to minimize discrimination based on race and ethnicity. The audit is part of broader efforts to increase transparency and accountability in AI-driven ad targeting (Brookings, 2021).

These campaigns and initiatives are crucial in promoting transparency, accountability, and fairness in AI systems, particularly those used on social media platforms. They highlight the importance of diverse and inclusive datasets, ethical AI design, and the active involvement of civil society in shaping AI policies and practices.

### 7.1.3 Interventions in Educational Institutions

Several educational institutions have designed interventions to promote the mitigation of bias and discrimination in AI. These initiatives cover a wide range of activities such as implicit bias training programs for students and faculties and incorporating modules on bias in data science and machine learning courses. Other activities included development of collaborative research projects with industry and the government to conduct research on AI bias, development of collaborations with policymakers to promote regulations and standards that address AI bias, as

well as promoting diversity by encouraging the recruitment and support of underrepresented groups in AI fields, thus ensuring diverse perspectives are included in the development of AI technologies.

Here are some notable examples: The Leverhulme Centre for the Future of Intelligence (CFI) at the University of Cambridge runs the "AI: Ethics and Society" research program which specifically investigates the ethical challenges posed by AI, including bias and discrimination. The Institute for Ethics in Artificial Intelligence at the Technical University of Munich - amongst its other initiatives regarding the development of skills to identify and mitigate bias in AI systems- has developed an interdisciplinary approach which ensures that ethical considerations are integrated into AI research and development processes. The Centre for Digital Ethics and Policy at UCL conducts research and provides education on the ethical challenges of digital technologies, including AI. Carnegie Mellon University's AI Ethics and Society Initiative integrates ethical considerations into AI education and research. The initiative offers interdisciplinary courses that address bias in AI and promote the development of fair and accountable AI systems. CMU also conducts research projects aimed at understanding the root causes of AI bias and developing methods to reduce it.

## 7.2 Policies and Regulations

Recent policies and regulations have been designed to address and mitigate bias and discrimination in AI. In the United States, significant steps have been taken under the Biden-Harris Administration to ensure responsible AI development. A key initiative is the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, signed in October 2023. This Executive Order establishes new standards for AI safety and security, with specific directives for federal agencies to protect privacy, advance equity, and mitigate bias in AI systems. It emphasizes the need for transparency and accountability in AI deployment and promotes international collaboration on AI governance to address global challenges associated with AI technologies.

Additionally, federal agencies such as the Federal Trade Commission (FTC), the Department of Justice (DOJ), the Equal Employment Opportunity Commission (EEOC), and the Consumer Financial Protection Bureau (CFPB) have issued a joint statement pledging to increase enforcement efforts against bias in automated systems. This joint statement highlights the potential for AI systems to perpetuate unlawful bias and discrimination and reiterates the application of existing legal authorities to these technologies. The statement underscores the importance of using representative and balanced datasets to prevent skewed outputs and illegal discrimination.

In Europe, the European Union (EU) is also actively addressing AI bias through the Artificial Intelligence Act (AI Act), which was entered into force on the 1st of August 2024 and establishes a comprehensive legal framework to ensure the safe and ethical development and deployment of AI systems within the EU. This legislation mandates rigorous testing, documentation, and oversight to ensure AI systems are transparent, non-discriminatory, and respectful of fundamental rights. The AI Act also establishes a risk management framework for AI applications, requiring regular assessments to identify and mitigate potential biases and discriminatory

outcomes. The Act adopts a potential risk-based to health, safety and fundamental rights approach, categorizing AI systems into four levels of risk: minimal, limited, high, and unacceptable, with corresponding obligations for each category. General-Purpose AI (GPAI) Models are addressed by imposing transparency obligations and, for those posing systemic risks, stricter requirements such as risk management and reporting of serious incidents. A governance framework is established involving national competent authorities and the European AI Office to oversee implementation and enforcement. The implementation has a very specific timeline and impacts various stakeholders, such as providers and deployers, Member states and the public. Entities involved with high-risk AI systems must ensure compliance with the Act's requirements, including conducting conformity assessments and implementing risk management measures. The Member states are required to designate national authorities responsible for supervising AI systems and establish AI regulatory sandboxes to facilitate innovation while ensuring compliance. Regarding the public, the main aim of the AI Act is to protect fundamental rights, ensuring that AI systems are developed and used in a manner that is safe, transparent, and respects individual freedoms. The AI Act represents a significant step in regulating AI technologies, setting a precedent for AI governance globally. Its phased implementation allows stakeholders to adapt to new requirements, promoting the development of trustworthy AI systems within the EU.

Both the U.S. and EU approaches emphasize the need for comprehensive frameworks that integrate ethical considerations into AI development and deployment. These efforts reflect a broader commitment to fostering fair, inclusive, and accountable AI systems that safeguard individuals' rights and promote social equity.

### 7.3 Corporate Initiatives

Several leading tech companies have taken proactive steps to address and mitigate bias in their AI systems. These initiatives range from developing ethical guidelines and tools to fostering diverse and inclusive work environments. These corporate initiatives highlight the commitment of leading tech companies to address and mitigate bias in AI. By developing ethical guidelines, creating fairness-focused tools, and fostering diverse and inclusive environments, these companies are working to ensure that AI technologies are fair, transparent, and accountable. Such efforts are essential for building trust in AI systems and promoting their responsible use across various applications.

#### 7.3.1 Strategies of Addressing AI bias Internally

There are corporations that have demonstrated a commitment to addressing AI bias in their activities through comprehensive strategies that involve regular testing, transparency, ethical guidelines, and promoting diversity within their AI development teams. Their efforts contribute to the broader goal of developing fair and trustworthy AI systems.

For example, Microsoft (<https://www.microsoft.com/en-us/ai/principles-and-approach/>) and Google (<https://ai.google/responsibility/responsible-ai-practices/>) have published extensive guidelines on the responsible use of AI, highlighting the importance of ethical considerations in

AI development. Both of them advocate for transparency, fairness, and accountability, and have implemented practices to test and mitigate bias in their AI systems.

IBM has, also, been at the forefront of advocating for and implementing measures to mitigate AI bias. They have developed a comprehensive policy framework that emphasizes accountability, transparency, fairness, and security (<https://www.ibm.com/impact/ai-ethics>). IBM calls for regular bias testing, documentation of assessment processes, and ongoing monitoring of high-risk AI systems. They also promote AI literacy and diversity in AI development teams to reflect a broader range of perspectives and minimize bias.

Another example is Deloitte which has developed tools and techniques to detect and remediate bias in AI systems. They train their AI developers to recognize and address bias and promote transparency by explaining how AI algorithms make decisions. Deloitte's approach includes integrating control structures and processes to manage the risks associated with AI bias, fostering an ethical AI development environment. (<https://www.deloitte.com/global/en/about/story/purpose-values/commitment-to-responsible-business-practices.html>)

### 7.3.2 Examples of Corporate Responsibility in AI Ethics

Google has established a set of AI principles that emphasize fairness, transparency, and accountability. These principles guide the company's AI development and deployment, ensuring that AI systems do not perpetuate bias or discrimination. They also developed the What-If Tool, a visual interface that helps AI developers analyse their machine learning models. The tool allows users to explore model performance across different data subsets, identify potential biases, and make adjustments to mitigate them.

Microsoft in cooperation with several companies and academic institutions has, also, participated in a community driven project to introduce an open-source toolkit, Fairlearn, which is designed to help developers assess and improve the fairness of their AI models. Fairlearn provides tools to visualize and mitigate disparate impacts, promoting fairer AI outcomes. (<https://fairlearn.org/>)

Another example is IBM's AI Fairness 360, an open-source library that provides metrics to check for bias in datasets and machine learning models. It also includes algorithms to mitigate identified biases, helping developers create fairer AI systems. (<https://aif360.res.ibm.com/>)

Social media have also launched tools to address exclusion and discrimination such as Facebook (Meta). Facebook has launched the Inclusive AI program, which focuses on building AI systems that work fairly across diverse populations. This initiative includes efforts to collect diverse datasets and develop algorithms that perform well for all users. They also developed Fairness Flow; an internal tool used to evaluate the fairness of machine learning models. The tool helps engineers identify and mitigate biases during the development process, ensuring more equitable outcomes.

Finally, Amazon has initiatives to ensure that its AI systems are trained on diverse and representative datasets. This helps reduce biases that can arise from homogeneous data.



## 8 Developing Effective Workshops

As presented in previous sections of this report, there were five panel discussions conducted on AI and governance which provided the project with insightful findings and recommendations on topics such as digital transformation, AI implementation in local governance, challenges and ethical considerations, and inclusivity in AI for citizens. Based on the experience of participating in and conducting such panel discussions, certain suggestions and recommendations regarding the structure and organization of effective workshops on AI and bias discrimination and mitigation have emerged.

Developing an effective workshop on awareness regarding bias and discrimination in AI requires a well-structured approach that combines theoretical knowledge with practical exercises. Some guidelines have to be followed to ensure the workshop is impactful and informative.

- Define clear objectives such as awareness, identification, mitigation and ethics and fairness.
- Develop a comprehensive agenda which must include an introduction to AI bias, definition of key concepts such as bias, discrimination, fairness, and ethics in the context of AI and provide to the participants case studies highlighting incidents of AI bias and their consequences.
- Include in the workshop interactive sessions and hands-on exercises where participants can work with datasets and AI models to identify and address biases. Use tools like Google's What-If Tool or IBM's AI Fairness 360. Also, facilitate small group discussions to explore different types of biases and their effects. Encourage sharing of personal experiences and insights. Include exercises that involve ethical decision-making scenarios, helping participants apply ethical principles in practical contexts.
- Incorporate diverse perspectives by inviting guest speakers, experts from diverse backgrounds, including ethicists, data scientists, and representatives from affected communities, to share their perspectives on AI bias and discrimination. Organize panels to discuss ethical dilemmas and best practices for promoting fairness in AI.
- Provide educational resources such as reading materials, key documents of regulatory and legal framework, such as GDPR and the EU AI Act, and ethical guidelines from companies like Google and Microsoft.
- Include feedback mechanisms towards the end of the workshop. Collect feedback from participants through surveys and discussions to understand the workshop's effectiveness and areas for improvement.
- Create a safe and inclusive environment by ensuring that the workshop environment is inclusive and respectful, encouraging open dialogue and diverse viewpoints. Offer support for participants who might find the discussions around bias and discrimination personally challenging.

## 8.1 Workshop Objectives

As mentioned before, an effective workshop on bias and discrimination in AI should have clear and comprehensive objectives to ensure participants gain a deep understanding of these issues and are equipped to address them in their work.

### **Objective No. 1. Raise Awareness of AI Bias and Discrimination**

- **Understanding Bias:** Educate participants on the various types of bias (e.g., cognitive bias, data bias, algorithmic bias) and how they manifest in AI systems.
- **Impact on Society:** Highlight the real-world implications of AI bias and discrimination on different communities, especially marginalized and underrepresented groups.

### **Objective No. 2. Identify Sources and Types of Bias**

- **Data Bias:** Teach participants how biases in training data can lead to biased AI outcomes. Discuss examples where data collection, selection, or labelling introduced bias.
- **Algorithmic Bias:** Explain how algorithmic design choices and model selection can perpetuate or amplify biases. Provide case studies illustrating these issues.

### **Objective No. 3. Develop Skills to Detect and Mitigate Bias**

- **Bias Detection:** Introduce tools and techniques for identifying bias in datasets and AI models, such as fairness metrics and bias detection algorithms.
- **Bias Mitigation Strategies:** Provide practical strategies for mitigating bias, including data augmentation, re-sampling, algorithmic adjustments, and fairness-aware machine learning techniques.

### **Objective No. 4. Foster Ethical and Responsible AI Development**

- **Ethical Principles:** Promote understanding of key ethical principles in AI, such as fairness, accountability, transparency, and inclusivity.
- **Regulatory Compliance:** Ensure participants are aware of relevant laws and regulations, such as GDPR in Europe and the AI Bill of Rights in the US, and their implications for AI development and deployment.

### **Objective No. 5. Encourage Critical Thinking and Ethical Decision-Making**

- **Scenario Analysis:** Engage participants in analysing ethical dilemmas and decision-making scenarios involving AI bias and discrimination.
- **Discussion and Reflection:** Facilitate discussions that encourage participants to reflect on their own biases and the ethical implications of their work.

### **Objective No. 6. Promote Diversity and Inclusion in AI Development**

- **Inclusive Practices:** Highlight the importance of diverse and inclusive teams in AI development and the role they play in mitigating bias.
- **Case Studies:** Present successful case studies where diverse teams and inclusive practices led to more equitable AI outcomes.

### **Objective No. 7. Facilitate Collaboration and Knowledge Sharing**

- **Networking Opportunities:** Create opportunities for participants to connect and collaborate with each other, fostering a community of practice around ethical AI.
- **Ongoing Support:** Offer follow-up sessions, online forums, or mentorship programs to support participants in their ongoing efforts to address AI bias and discrimination.

## 8.2 Target Audiences

When organizing a workshop to train and educate on AI discrimination and bias mitigation, it is essential to include a diverse array of target audiences to ensure comprehensive understanding and effective solutions. The primary target audiences should include:

**AI Developers and Data Scientists:** These are the individuals who design, build, and maintain AI systems. Training them on recognizing and mitigating bias is crucial, as they directly influence the creation of algorithms and the selection of training data (Binns, 2018).

**Policy Makers and Regulators:** Government officials and regulatory bodies play a critical role in establishing guidelines and laws that govern AI use. Educating them on AI discrimination helps in forming policies that promote fairness and accountability (Fung, 2006).

**Social Media Platform Executives and Managers:** Decision-makers in social media companies need to understand the impacts of AI bias on their platforms. Their involvement ensures that corporate policies align with ethical AI practices and support unbiased content moderation.

**Civil Society Organizations and Advocacy Groups:** Organizations that advocate for human rights and digital equity should be included to ensure the voices of marginalized communities are heard. These groups can provide valuable insights and push for inclusive AI practices.

**Academic Researchers and Educators:** Scholars and educators who study AI ethics and bias can contribute their research findings and help develop educational materials that raise awareness and understanding of AI discrimination (Noble, 2018).

**Journalists and Media Professionals:** As influencers of public opinion, journalists and media professionals should be educated on AI bias to accurately report on related issues and raise public awareness.

**General Public and End-Users:** Including everyday users of AI-driven platforms ensures that they are aware of potential biases and their rights. Educated users can better advocate for fairer AI systems and contribute to the discourse on ethical AI.

By engaging these diverse groups, the workshop can foster a holistic approach to addressing AI discrimination, ensuring that multiple perspectives are considered and that the resulting solutions are robust and inclusive.

### 8.3 Workshop Content

A workshop designed to raise awareness about bias and discrimination in AI systems should have defined and clear objectives on which the structure and content of the workshop is based. This structure should ensure a comprehensive and interactive approach to understanding, detecting, and mitigating bias in AI. By combining theoretical knowledge with practical exercises and discussions, participants will be well-equipped to promote fairness and ethical practices in AI development.

#### 8.3.1 Introduction to AI

Definition of AI, Key Concepts in AI, Model Building and Evaluation.

#### 8.3.2 Identifying Sources of Bias in AI.

**Overview of AI Bias:** Define AI bias and discrimination, types of bias (data bias, algorithmic bias, societal bias), and their sources.

**Impact on Society:** Discuss the societal implications of AI bias using case studies from various domains such as healthcare, criminal justice, hiring, and finance.

**Data Bias:** Explain how data collection, selection, and labelling can introduce bias. Provide examples and interactive exercises to identify bias in datasets.

**Algorithmic Bias:** Discuss how biases can be introduced through algorithm design and deployment. Include case studies and hands-on exercises to detect biases in AI models.

### 8.3.3 Techniques for Mitigating Bias in AI Systems.

**Data-Level Mitigation:** Techniques such as data augmentation, re-sampling, and synthetic data generation.

**Algorithm-Level Mitigation:** Fairness-aware algorithms, re-weighting, and adversarial debiasing.

**Practical Exercises:** Interactive sessions where participants apply these techniques to mitigate bias in sample datasets and models.

**Introduction to Bias Detection Tools:** Present and demonstrate tools such as IBM's AI Fairness 360, Microsoft's Fairlearn, and Google's What-If Tool.

### 8.3.4 Ethical Considerations in AI Development.

**Ethical Principles and Frameworks:** Discuss ethical principles such as fairness, accountability, transparency, and inclusivity. Reference guidelines from organizations like the IEEE, European Union, and other ethical frameworks.

**Legal and Regulatory Frameworks:** Overview of relevant laws and regulations such as GDPR, the AI Act, and the US Executive Order on AI.

**Compliance Strategies:** How to ensure AI systems comply with legal and regulatory requirements.

### 8.3.5 Hands-on Activities and Case Studies.

**Hands-On Exercises:** Provide practical sessions where participants use these tools to analyse datasets and models for bias.

**Case Studies:** Analyse ethical dilemmas and decision-making scenarios involving AI bias and discrimination.

**Group Projects:** Collaborative exercises where participants work in teams to identify, analyse, and mitigate biases in AI systems.

**Presentations:** Each team presents their findings and solutions, followed by group discussions and feedback.

## 9 Proposed Curriculum: Workshops on AI Bias and Discrimination

This chapter provides an indicative design of a curriculum for workshops aimed at raising awareness about AI bias and discrimination. It serves as an example derived from the findings of the discussion panels reported in this deliverable. The proposed curriculum highlights key themes and insights from all five panel discussions and is intended to guide the development of similar educational initiatives. Following the structure presented in chapter 8, the topics/content of such workshops are in this chapter.

The workshops aim to raise citizen awareness about AI bias and ethics, empowering them to advocate for fair and inclusive AI practices. Designed for a diverse audience, including policymakers and regulators, public sector officials, educators and students in crucial scientific fields like technology, law and social sciences, AI developers and technologists, and civil society stakeholders, the curriculum provides participants with the tools and knowledge needed to address AI bias and promote ethical implementation.

The curriculum for workshops on AI bias and discrimination is designed to provide participants with a comprehensive understanding of the sources, implications, and mitigation strategies associated with biased AI systems. These workshops aim to equip attendees with both theoretical knowledge and practical tools to identify and address AI bias, fostering the development of fairer and more inclusive technologies. The curriculum is structured as a series of interactive sessions, each with lectures, case studies, interactive discussions, and hands-on activities.

While the proposed curriculum offers an indicative example based on the findings of Task 5.1, it is important to note that a comprehensive curriculum has already been developed under Tasks 5.2 and 5.3 of the AI4Gov project. That curriculum has been applied in the development of AI4Gov MOOCs and provides a detailed framework for organizing workshops on awareness-raising and education around ethical AI practices. The content in this chapter is complementary, serving as an example derived from the discussion panels and offering additional inspiration for designing workshops in alignment with the guidelines outlined in Chapter 8.

### 9.1 Module 1: The Ethical Foundations of AI

This module integrates the ethical frameworks discussed across the panels, including the European Union's AI Act and UNESCO's guidelines. Participants will explore concepts like explainability, accountability, and fairness in AI design. Case studies, such as healthcare and public safety scenarios, will allow participants to evaluate AI systems against these ethical benchmarks.

Key Topics:

- Introduction to AI Ethics: Overview of ethical principles in AI (Fairness, Accountability, Transparency, and Human-Centered Design).
- Legal and Ethical Frameworks: Deep dive into the EU AI Act, UNESCO guidelines, and the IEEE's Ethical AI guidelines.

- Concepts of Explainability and Interpretability: Why it is important for AI systems to provide clear and understandable outputs, especially in public administration.
- Case Studies (examples):
  - Healthcare: Ethical implications of AI in medical diagnosis (e.g., disparities in medical image recognition for different demographic groups).
  - Public Safety: Discussion of AI use in predictive policing and the risks of wrongful profiling.

#### Activities:

- Group discussion on key ethical dilemmas in AI applications.
- Role-play: Participants act as policymakers reviewing the ethics of a proposed AI system for healthcare.

## 9.2 Module 2: Understanding AI Bias and Its Societal Impact

This introductory module explores the nature of AI bias and its implications for governance, public services, and societal equity. Drawing from the discussions on local governance and inclusivity, participants will examine real-world examples, such as biased decision-making in social benefits and predictive policing. The module will also highlight the ethical challenges raised during the panels, including data privacy, discrimination, and the transparency of AI systems.

#### Key Topics:

- What is AI Bias?: Definition and types (selection bias, confirmation bias, automation bias, etc.).
- Societal Impact of AI Bias: How AI bias affects access to social services, employment, and justice.
- Transparency and Ethical Challenges: Balancing transparency, fairness, and privacy in AI systems.
- Case Studies (examples):
  - Local Governance: Examples of biased decision-making in social benefit allocation.
  - Predictive Policing: Analysis of cases where predictive policing systems led to racial profiling.

#### Activities:

- Brainstorming session on "How AI Bias Impacts My Community".
- Interactive exercise where participants identify biases in fictional AI-based governance scenarios.

### 9.3 Module 3: Identifying Sources of Bias in AI Systems

Participants will delve into the root causes of AI bias, including skewed or incomplete datasets, lack of diversity in training data, and design choices that unintentionally embed systemic biases. The module includes hands-on exercises where participants analyse sample datasets and algorithms to identify potential biases.

Key Topics:

- Data-Centric Sources of Bias: Incomplete, unrepresentative, and skewed datasets.
- Algorithm-Centric Sources of Bias: The impact of model design choices on fairness.
- Design-Centric Bias: How developer assumptions and cultural contexts influence AI outputs.
- Interactive Examples: Identifying biases in real-world systems (e.g., gendered language in AI chatbots).

Activities:

- Hands-on Exercise: Participants review and analyze sample datasets to identify instances of imbalance or misrepresentation.
- Algorithm Audit: Participants identify flaws in a provided algorithm and propose corrections.

### 9.4 Module 4: Sources and Dynamics of AI Bias

Leveraging findings on governance and political methodology, this module explains how bias infiltrates AI systems at various stages, including data collection, algorithm design, and deployment. Practical exercises will allow participants to identify sources of bias, such as skewed datasets or lack of diversity in training data, using examples like the disparities discussed in local governance and healthcare applications.

Key Topics:

- Dynamic Nature of Bias: How bias can be compounded or amplified at various stages of AI development.
- Data Labeling Bias: Human errors or assumptions that influence training data annotations.
- Algorithmic Amplification: How machine learning algorithms reinforce existing inequalities.
- Case Studies:
  - Disparities in healthcare AI systems.
  - Bias in credit scoring and loan approvals.



#### Activities:

- Real-world exercise: Participants analyze the stages of an AI system and identify "bias entry points".
- Problem-solving session on developing better data collection processes to reduce bias.

### 9.5 Module 5: Mitigating AI Bias and Ensuring Inclusivity

Building on the discussions about transparency and accountability in AI, this module focuses on actionable strategies to reduce bias and promote inclusivity. Participants will learn to use fairness metrics, preprocess data for diversity, and design algorithms that account for underrepresented groups. Tools and techniques will be drawn from the insights shared about auditing AI systems and fostering community engagement.

#### Key Topics:

- Techniques for Bias Mitigation: Techniques like data rebalancing, data augmentation, and adversarial debiasing.
- Promoting Diversity and Inclusion in AI Systems: Ensuring underrepresented groups are included in model design and training data.
- Fairness Metrics: Overview of fairness metrics, such as demographic parity, equal opportunity, and equalized odds.

#### Activities:

- Hands-on Exercise: Participants apply fairness metrics to sample datasets.
- Collaborative Design Exercise: Groups design an AI system while following principles of fairness and inclusivity.

### 9.6 Module 6: Tools and Techniques for Bias Mitigation

This module focuses on practical strategies for reducing bias in AI systems. Topics include:

- Fairness metrics for evaluating bias.
- Techniques for preprocessing data to improve diversity and representation.
- Algorithmic adjustments and post-processing methods to ensure equitable outcomes.
- Participants will practice applying these techniques using open-source tools and software.

#### Key Topics:

- Fairness Toolkits: Overview of popular tools like IBM AI Fairness 360, Google's What-If Tool, and Microsoft's Fairlearn.
- Data Preprocessing: Techniques for improving data quality before training.
- Post-Processing Techniques: Methods to adjust model outputs to improve fairness.

Activities:

- Practical Lab: Participants use AI fairness tools to analyze a dataset and reduce bias.
- Coding Exercise: Participants build a small AI model and apply bias mitigation techniques.

## 9.7 Module 7: Ethical AI Frameworks and Best Practices

Participants will learn about existing ethical frameworks, such as those developed by UNESCO, the IEEE, the European Commission and the AI4Gov's HRF. The module explores how these frameworks can guide the design and deployment of AI systems, ensuring compliance with ethical and legal standards. Interactive case studies will allow participants to evaluate AI systems against these guidelines.

Key Topics:

- Global Ethical Standards: Principles from UNESCO, IEEE, and the European Commission.
- Implementing Ethical Guidelines: How to apply guidelines during the AI lifecycle.
- Case Study Analysis: Ethical dilemmas in AI deployment.

Activities:

- Scenario Analysis: Participants evaluate an AI system using ethical guidelines from UNESCO.
- Group Discussion: How ethical frameworks can be integrated into day-to-day AI development.

## 9.8 Module 8: Advocating for Transparent and Accountable AI

Highlighting recommendations about public participation and democratic engagement, this module emphasizes the role of citizens in shaping AI governance. Participants will learn advocacy strategies to promote fair practices within their communities, including tools for engaging with policymakers and raising awareness about the impact of AI on democracy and human rights.

Key Topics:

- Transparent AI Systems: Explainable AI (XAI) and tools for transparency.
- Mechanisms for Accountability: Audit trails, third-party evaluations, and model documentation.
- Public Participation: Engaging communities in AI design and evaluation.

Activities:

- Advocacy strategy development: Participants draft an action plan to promote AI accountability in their community.

- Role-playing: Participants act as community leaders advocating for transparent AI in a fictional town hall meeting.

## 9.9 Module 9: Policy Implications and Advocacy

This module emphasizes the broader societal and policy dimensions of AI bias. Participants will explore the role of governance in promoting fair AI, including regulatory challenges and opportunities. Advocacy strategies for addressing AI bias within organizations and communities will also be covered, with examples of successful interventions from around the world.

Key Topics:

- Overview of AI Policy: Analysis of key policies like the European Union’s AI Act.
- Policy Advocacy: Strategies to advocate for fair AI policies.
- Global Case Studies: Successful AI policy interventions in different countries.

Activities:

- Group Exercise: Participants create a policy brief for a hypothetical AI regulatory body.
- Interactive Discussion: Best practices for engaging policymakers and regulatory bodies.

## 9.10 Module 10: Collaborative Problem-Solving for Ethical AI

Reflecting the interdisciplinary nature of the panels, this module fosters collaboration by engaging participants in group activities. Teams will tackle real-world AI challenges, such as addressing misinformation or ensuring inclusivity in digital platforms, using the knowledge and tools acquired throughout the workshop.

Key Topics:

- Problem-Solving Methodologies: Design Thinking and Systems Thinking approaches.
- Group Collaboration: Teams analyze real-world problems and develop solutions.
- Interdisciplinary Collaboration: Working across technical, social, and policy fields.

Activities:

- Team Challenge: Teams work on a case study (e.g., redesigning a biased AI system for credit scoring).
- Presentation: Teams present their proposed solutions, including ethical guidelines, technical changes, and advocacy recommendations.

## 9.11 Evaluation and Certification

### **Final Assessment:**

- Participants present a final group project demonstrating their understanding of AI bias, the tools for mitigation, and ethical advocacy.
- Successful participants receive a "Certificate in Ethical AI and Bias Mitigation".

This 10-module curriculum offers a comprehensive framework for promoting ethical, fair, and inclusive AI systems. By addressing both technical and policy dimensions, participants gain the skills needed to identify, mitigate, and advocate against AI bias.

## 9.12 Learning Methods and Delivery

The curriculum employs a mix of lectures, case studies, hands-on exercises, and collaborative activities to engage participants. Workshops can be delivered in-person or virtually, using interactive platforms and materials such as annotated datasets, open-source tools, and real-world AI case studies.

## 9.13 Target Audience

The curriculum is suitable for a wide range of participants, including policymakers and regulators, public sector officials, educators and students in crucial scientific fields like technology, law and social sciences, AI developers and technologists, community leaders, civil society organisations and activists. By tailoring content to different groups, the workshops ensure that all stakeholders gain the skills and insights needed to contribute to ethical AI development.

## 9.14 Learning Outcomes

Upon completion, participants will have:

- A clear understanding of the causes and types of bias in AI systems.
- A clear understanding of the sources and impacts of AI bias.
- Practical experience in identifying and mitigating bias in AI systems.
- Gain insight into regulatory frameworks and ethical standards for AI.
- Develop strategic skills to advocate for fair and inclusive AI practices within their areas of influence.
- Learn strategies for mitigating bias in AI development and deployment.
- Explore case studies of AI discrimination in various sectors (e.g., healthcare, criminal justice, employment).
- Develop the capacity to critically analyse AI decision-making processes.

## 10 Conclusions

The focus groups and panel discussions conducted under Task 5.1 of WP5 in the AI4Gov project have underscored several critical findings regarding AI bias and discrimination. These activities highlighted that AI bias can manifest in various domains, such as healthcare, governance, and public service delivery, impacting marginalized and vulnerable populations disproportionately. The discussions emphasized the need for ethical and transparent AI technologies to foster inclusiveness and mitigate discrimination.

Key recommendations from these discussions include the following:

Develop Comprehensive Evaluation Frameworks: Create robust frameworks that consider the purpose, stakeholder input, and context of AI applications. Incorporate key performance indicators (KPIs) to track the impact of AI projects on sustainable development goals (SDGs) and ethical considerations.

Enhance Bias Mitigation Strategies: Redesign evaluation forms and implement established ethics frameworks to address biases effectively. Regular bias audits and the use of diverse training data are crucial for mitigating biases during AI development.

Promote Transparency and Accountability: Ensure AI algorithms and models are publicly available for scrutiny and validation. Continuous monitoring and evaluation of AI systems are necessary to ensure they adhere to ethical standards and operate without biases.

Foster Community Engagement and Education: Involve local communities, businesses, and NGOs in AI decision-making processes to enhance transparency and accountability. Educating stakeholders on the benefits and risks of AI promotes informed participation and trust in AI technologies.

Ensure Ethical and Legal Compliance: Prioritize data privacy and prevent discrimination through comprehensive data characterizations and anonymization. Compliance with data protection regulations is essential to safeguard personal information.

Balance Automation with Human Oversight: Use AI to support decision-making with real-time data and insights, but ensure final decisions are made by humans to maintain ethical and contextual appropriateness.

Provide Training and Support: Equip stakeholders with the necessary skills to use AI platforms effectively and responsibly. Training programs should be tailored to meet the diverse needs of different demographic groups, including elderly citizens, refugees, and migrants.

To translate these recommendations into actionable steps, the workshops developed under WP5 should have to focus on enhancing understanding of AI bias, its societal impacts, and promoting ethical practices in AI development and deployment. These workshops incorporate both theoretical knowledge and practical exercises, such as using tools to identify and mitigate bias in AI models. By engaging diverse participants, these workshops aim to cultivate a well-informed public capable of actively participating in shaping ethical AI frameworks.

Ultimately, this document serves as a foundational resource for developing strategies to raise awareness on AI issues and offers an exemplar structure and curriculum for creating workshops that not only educate but also empower citizens to advocate for ethical AI practices. By implementing these recommendations, stakeholders can develop and deploy AI technologies that are fair, inclusive, and beneficial to all citizens, contributing to a more equitable technological landscape.

## 11 References

- Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press.
- Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy*. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency
- Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82(1), 62–68
- Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, Vol. 36, pp. 1–52). Elsevier Academic Press.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4–27
- Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, 945-967.
- FitzGerald, C., Hurst, S. (2017). Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics* 18, 19
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Lilienfeld, S.O., Ammirati, R.J., & Landfield, K. (2009). Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science*, 4, 390 - 398.
- Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. (2023) Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health* 2(6): e0000278.
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019 Dec;9(2):010318.
- Stanovich, Keith & West, Richard. (2008). On the Relative Independence of Thinking Biases and Cognitive Ability. *Journal of personality and social psychology*. 94. 672-95.