AI4Gov

Trusted AI for Transparent Public Governance fostering Democratic Values

# Deliverable 2.1

# *AI4Gov Holistic Regulatory Framework V1*

**30-06-2023**

**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 | |
|---|---|
| **Dissemination level** | Public |
| **Version** | 1.0 |
| **Status** | Final |
| **Beneficiary** | ViLabs |
| **License** | This work is licensed under a Creative Commons Attribution-No Derivatives 4.0 International License (CC BY-ND 4.0). See: https://creativecommons.org/licenses/by-nd/4.0/ |

| AUTHORS | | |
|---|---|---|
| | **Name** | **Organisation** |
| **Document leader** | Mpampis Chatzimallis | ViLabs |
| **Participants** | Danai Kyrkou | ViLabs |
| | Prof.Theodore Chadjipadelis, Georgia Panagiotidou, PhD | AUTH |
| **Reviewers** | Spiros Borotis, Sotiris Athanasopoulos, Nikos Achilleopoulos | MAG |
| | Nechifor Cosmin-Septimiu, Raluca-Maria Repanovici | SIE |
| | George Manias | UPRC |

| VERSION HISTORY | | | | |
|---|---|---|---|---|
| **Version** | **Date** | **Author** | **Organisation** | **Description** |
| 0.1 | 02.06.2023 | Mpampis Chatzimallis | ViLabs | ToC |
| 0.2 | 21.06.2023 | Danai Kyrkou, Prof.Theodore Chadjipadelis, Georgia Panagiotidou, | ViLabs, AUTH | First draft released |
| 0.3 | 26.06.2023 | Nechifor, Cosmin-Septimiu, Raluca-Maria Repanovici, George Manias, Sotiris Athanasopoulos, Spyros Borotis, Nikos Achilleopoulos | MAG, SIE, UPRC | Internal reviewing process and updates |
| 1.0 | 30.06.2023 | Mpampis Chatzimallis, Danai Kyrkou, Nechifor, Cosmin-Septimiu, Raluca-Maria Repanovici, George Manias, Sotiris Athanasopoulos, Spyros Borotis, Nikos Achilleopoulos | ViLabs, MAG, SIE, UPRC | Final contributions and review for submission |

# Table of Contents

**Abstract ............ 6**

**1. Introduction ............ 7**

1.1. Purpose and scope ............ 7

1.2. Document structure ............ 7

1.3. Target audience of the deliverable ............ 8

**2. Identification of Traditional Bias Impacting Rights and Values ............ 9**

2.1. WP2 Structure ............ 9

2.2. Definition of Traditional Bias and types of bias ............ 10

*2.2.1. Conscious bias ............ 10*

*2.2.2. Unconscious Bias ............ 11*

2.3. Impact of bias on individuals and society. ............ 13

2.4. Underrepresented groups ............ 14

*2.4.1. Identification of Underrepresented Groups ............ 14*

*2.4.2. Challenges Faced by Underrepresented Groups ............ 16*

*2.4.3. Introduction to the sources of bias ............ 16*

*2.4.4. Common environments where individuals may perceive discrimination ............ 17*

*2.4.5. Conclusion ............ 19*

**3. The Impact of Bias in AI4Gov Pilot Cases: Addressing Ethical and Societal Implications ............ 20**

*3.1.2. Detecting and addressing bias: Provide tips and strategies for recognizing and mitigating bias, including self-reflection, education, and diversity and inclusion initiatives. ............ 22*

*3.1.3. Exploring Bias in AI: Insights and Reflections from the 1st AI4Gov Training Workshop ............ 23*

*3.1.4. "Workshop Summaries and Expert Insights: Addressing Human Rights, AI, Travel Categories, and Biases - Insights from a ViLabs workshop" ............ 25*

*3.1.5. Conclusion: Summarize the key points and emphasize the importance of addressing bias for promoting fairness, inclusivity, and equality. ............ 26*

**4. The Interview and Survey Process ............ 28**

4.1. Overview ............ 28

4.2. Focus and Topics of the Interviews ............ 28

4.3. Compliance with EU Regulations on Human Rights ............ 29

4.4. Identifying Areas of Overlooked Discrimination ............ 30

4.5. Methodology of the Interviews and Surveys ............ 31

**5. Initial analysis for the AI-based Democracy Holistic Regulatory Framework (HRF) ............ 34**

5.1. Research Design ............ 34

5.2. Methodology ............ 34

5.3. Results ............ 35

*5.3.1. Strengths ............ 35*

*5.3.2. Weaknesses ............ 39*

*5.3.3. Opportunities ............ 41*

*5.3.4. Threats ............ 43*

**6. Conclusions ............ 47**

**7. References ............ 48**

**ANNEX 1. Ai4GOV Vocabulary ............ 52**

## Abbreviations

| Abbreviation | Description |
|---|---|
| HRF | Holistic Regulatory Framework |
| AI | Artificial Intelligence |
| ML | Machine Learning |
| BIE | Blockchain-based Information Exchange |
| DGF | Data Governance Framework |
| XAI | Explainable AI |
| HRF | Holistic Regulatory Framework |
| GDPR | General Data Protection Regulation |
| XAI Library | eXplainable AI Library |

# Abstract

This report investigates the scope and impacts of traditional bias, paying particular attention to underrepresented groups and the challenges they encounter. It sheds light on the far-reaching influence of both conscious and unconscious biases on individuals and societies, and how such biases permeate decision-making processes, behaviours, and outcomes across a range of sectors.

The research specifically delves into bias present within Artificial Intelligence (AI) applications in governmental instances (AI4Gov pilot use-cases), drawing attention to the ethical and societal implications as well as potential adverse consequences. A comprehensive overview of these cases furnishes a more profound understanding of the manifestation of bias in AI and its implications in real-world scenarios.

Recommendations for mitigating bias are proffered, with a focus on self-reflection, diversity and inclusion initiatives, and continuous education. Insights are offered into the practical applications and strategies for addressing bias, gathered from an in-depth exploration of AI4Gov Training Workshops.

Additionally, the research employs an interview and survey process to identify underrepresented groups and neglected areas of discrimination, ensuring adherence to EU Regulations on Human Rights. Finally, the report presents an initial analysis of the AI-based Democracy Holistic Regulatory Framework, taking into account its strengths, weaknesses, opportunities, and threats.

The research underscores the crucial role of acknowledging and addressing bias, especially in the realm of AI implementations, as a strategy to foster fairness, inclusivity, and equality across all societal sectors.

# 1. Introduction

## 1.1. Purpose and scope

This deliverable of Work Package 2 (WP2) of the AI4Gov project aims to provide a comprehensive analysis and specification of key aspects related to fundamental rights, bias, discrimination, and the integration of AI technologies in order to support the upcoming design of the project's Holistic Regulatory Framework (HRF). The purpose of this document is to align technical and social-science definitions, identify the impact of bias on citizens' groups and support the grounds for the upcoming design and development of a regulatory framework. It serves as the first version of the deliverable and lays the groundwork for addressing bias and discrimination, ensuring compliance with EU regulations, and facilitating the practical application of AI4Gov technologies.

## 1.2. Document structure

The deliverable is structured into **six chapters**. It begins with the introduction chapter **(Chapter 1)** outlining the purpose and scope of the deliverable and highlighting its structure. The report then proceeds with **Chapter 2** with information on WP2 and the identification of traditional bias and their impact on rights and values. **Chapter 3** consists of a depiction of the impact of Bias In the Ai4Gov project pilot cases, and supports the scope of the deliverable with results from relevant research and participatory activities, while **Chapter 4** then provides information on the interview and surveys process that will be followed in the upcoming project period, while **Chapter 5** consists of an initial analysis of the ai-based democracy HRF, presenting the research design, methodology, and results categorized into strengths, weaknesses, opportunities, and threats. The report concludes with **Chapter 6** conclusions, summarising the key findings and insights. **The References section** provides the list of sources used for D2.1 for further reading, and an appendix may be included for additional relevant information or data. Finally, to aid third-party readers in familiarising themselves not only with the terms used in this paper, but more importantly, with those of the overall project, an AI4Gov 'Vocabulary' is provided as an **ANNEX.** This is a 'live' document that will be regularly updated and made publicly available, serving as a constant source of support for understanding the evolving language of the project. Simultaneously, it serves an essential role within this deliverable by offering support to our project partners. The 'Vocabulary' acts as an initial guide to the complex subject of bias and its related terminologies, providing a baseline understanding necessary for deeper exploration and discourse in subsequent stages of the project."

## 1.3. Target audience of the deliverable

The target audience of the deliverable in Work Package 2 (WP2) includes various stakeholders involved in the AI4Gov project and those interested in the intersection of AI, governance, and fundamental rights, including:

- Project Consortium Members: The deliverable is primarily intended for the project consortium members involved in WP2 and other related work packages. with the document offers detailed insights related to the qualitative analysis and regulatory framework, in order to support the technical partners in the development of the AI4Gov platform and tools.
- Project Stakeholders: Other stakeholders involved in the AI4Gov project, such as external advisors, experts, and policymakers, may be interested in this deliverable since it provides valuable information on the analysis of fundamental rights and the development of a regulatory framework.
- Researchers and Academics: The deliverable can be relevant to researchers and academics working in the fields of AI, governance, ethics, and fundamental rights. The included qualitative analysis methods and regulatory frameworks, offer potential areas for further research and study.
- Policy and Decision-Makers: Policymakers and decision-makers involved in governance, ethics, and AI regulation may be interested about the outcome of this deliverable to understand the analysis of fundamental rights and the development process of a regulatory framework. This knowledge can help policy makers to connect the dots among AI, Bias, and policy.
- General Public: Although the deliverable contains technical details as well, it can also be of interest to whoever is concerned about the ethical implications of AI and the protection of fundamental rights. D2.1 offers insights into the analysis of bias and introduces different types of biases and the environments that a person may experience.

# 2. Identification of Traditional Bias Impacting Rights and Values

## 2.1. WP2 Structure

Work Package 2 plays a crucial role in the project by focusing on several key objectives. Firstly, it aims to contribute to a comprehensive understanding of bias, discrimination, unfairness, and non-inclusiveness within the context of AI and Big Data. This involves aligning technical and social science definitions and analysing the interconnections between these aspects during the design and deployment of AI technologies. Secondly, WP2 aims to establish a reference architecture that outlines the practical application and operation of the project's technologies. This architecture serves as a framework for implementing the AI4Gov environment effectively. Furthermore, this work package focuses on integrating the various technological building blocks from WP3 and WP4 to create a cohesive AI4Gov environment. This involves integrating data management solutions, AI tools, and other necessary enablers to form an integrated platform. Lastly, WP2 conducts research analysis and assessment to ensure compliance with current EU regulations on fundamental rights and values. By addressing these objectives, WP2 contributes to the overall success of the project and supports the development of an ethical and legally compliant AI4Gov platform**.** **WP2 breaks down into four tasks that cover various aspects related to bias, discrimination, fundamental rights, values, and the design and integration of the AI4Gov platform:**

- T2.1 Qualitative analysis on fundamental rights & values: focuses on identifying the extent of traditional bias impairing the rights and values of specific citizen groups through interviews and surveys. It also assesses compliance with EU regulations on human rights protection.
- T2.2 Specification and design of the AI-based Democracy Holistic Regulatory Framework: involves the specification and design of the AI-based Democracy Holistic Regulatory Framework. This task analyzes existing processes, maps them to the policy management lifecycle, and ensures alignment with legal activities, ethical protocols, and applicable laws and regulations.
- T2.3 Reference Architecture Specification: specifies the overall architecture of AI4Gov, describing the components, relations, and functionality. It defines open interface specifications, communication patterns, and information flows. This task also considers design principles and new processes for digital policy making pipelines.
- T2.4 Integration of AI4Gov Platform and Tools: focuses on integrating the mechanisms designed in WP2 to WP5 based on the architecture specification. It includes the integration of AI tools as services to the AI4Gov platform, adopting collaborative development tools and methodologies to ensure smooth integration and high-quality results.

These tasks collectively contribute to the development of a comprehensive framework and platform that address bias, protect fundamental rights, and enable effective policy making in the AI4Gov project.

In the context of identifying traditional bias impacting rights and values, it is essential to have a clear understanding of what traditional bias entails and the different types of bias that exist. The following subsections are provided to support this identification.

## 2.2. Definition of Traditional Bias and types of bias

Above all, it is important to define what is bias. There are a lot of different understandings of the term, but in AI4Gov project, the following definition has been chosen: Bias is a term used to describe an inclination or prejudice for or against an individual or group in a way that is considered unfair. It can result from personal experiences, societal norms and expectations, or information we have absorbed from various sources such as media, education, and family (Greenwald & Krieger, 2006). Bias can be categorised into two primary types: conscious (or explicit) bias and unconscious (or implicit) bias (Banaji & Greenwald, 2013). Both types of bias can have a significant impact on individuals and society, perpetuating social inequalities and injustices.

Considering all of these, the relevance of bias to human rights is profound. Human rights are inherent to all individuals, regardless of their background, identity, or characteristics. However, bias can undermine these rights, impeding equal access to opportunities, resources, and fair treatment. It can also perpetuate inequality, discrimination, and marginalisation, particularly for underrepresented groups. When bias comes into play, it can have far-reaching consequences for human rights. Bias can lead to discrimination, inequality, and the violation of basic rights. Ultimately, it can result in denial of opportunities or equal access to resources and services.

### 2.2.1. Conscious bias

Conscious bias, also known as Explicit, conscious, or overt Bias (Dovidio, Kawakami, & Gaertner, 2002) (Dovidio & Gaertner, Aversive racism, 2004): describes prejudices that people openly and deliberately believe. These biases are founded on overt prejudices, attitudes, or beliefs towards specific populations. Direct remarks, discriminatory behaviour, or discriminatory policy are all examples of explicit prejudice expression. Unlike implicit biases, explicit biases are maintained actively and might represent discriminating or prejudiced ideas. Explicit socialization processes, personal beliefs, and ideologies are more likely to have an impact on implicit biases than other factors. They can be quantified via self-report questionnaires or seen in actions and interactions.

To address conscious bias, education and awareness programs can be implemented to increase understanding and empathy towards different groups of people (Czopp et al., 2006). It is also important to hold individuals and organizations accountable for their actions and to promote diversity and inclusivity in all aspects of society (Kalev et al., 2006).

### 2.2.2. Unconscious Bias

Unconscious bias, also known as implicit bias, refers to the automatic and unconscious preference or discrimination against certain individuals or groups based on their race, gender, or other characteristics (Greenwald & Banaji, 1995). Unconscious bias operates at a subconscious level and can be difficult to recognize, even by those who hold them. Examples of unconscious bias include assuming that a woman is less competent than a man in a leadership position or associating certain ethnic groups with negative stereotypes.

Unconscious bias can perpetuate inequalities and hinder diversity in various contexts, such as education, healthcare, and workplaces (Fitzgerald & Hurst, 2017). To address unconscious bias, individuals can take implicit bias tests to identify their own biases and implement strategies to reduce bias in decision-making, such as blind hiring or diverse hiring committees (Devine et al., 2012).

Unconscious bias is prevalent in various settings, such as healthcare, education, and workplaces, and can have significant consequences for individuals and society (Fitzgerald & Hurst, 2017). For instance, unconscious bias in healthcare can lead to misdiagnosis and inadequate treatment for certain groups of people (Burgess et al., 2019). In education, unconscious bias can affect grading and teacher-student interactions, leading to unequal opportunities for students (Staats et al., 2016). In workplaces, unconscious bias can affect hiring and promotion decisions, leading to a lack of diversity and inclusivity.

Recognising and addressing unconscious bias is crucial for creating a fair and inclusive environment. Successful initiatives to address unconscious bias include diversity and inclusion training, implicit bias testing, and implementing strategies to reduce bias in decision-making (Devine et al., 2012; Kalev et al., 2006). In conclusion, bias can have a significant impact on individuals and society. Conscious and unconscious bias are two types of bias that can lead to discrimination, stereotyping, and social injustice. Recognizing and addressing unconscious bias is crucial for creating a fair and inclusive environment in various contexts such as workplaces, education, and healthcare. It is important for individuals and organizations to reflect on their own biases and take action to reduce them to promote diversity and inclusivity in all aspects of society.

#### *2.2.2.1. Cognitive Bias*

Cognitive bias is a term used to describe the systematic errors in thinking that occur when people process information (Kahneman, 2011). These biases can affect our decision-making, leading us to make flawed judgments or draw incorrect conclusions (Kahneman, 2011). Understanding cognitive bias is crucial because it can help us make better decisions, avoid errors, and improve our overall cognitive abilities (Banaji & Greenwald, 2013).

Cognitive bias can take many forms and can be met in different contexts including confirmation bias, anchoring bias, availability heuristic bias and many others.

Confirmation bias is a cognitive bias that occurs when people seek out information that confirms their pre-existing beliefs or opinions, while ignoring information that contradicts them (Nickerson, 1998). This bias can lead to the reinforcement of false beliefs and can make it difficult for people to change their minds. Anchoring bias occurs when people rely too heavily on the first piece of information they receive when making a decision, even if that information is irrelevant or inaccurate (Tversky & Kahneman, 1974). This bias can lead to errors in judgment and can prevent people from considering other relevant information. Availability heuristic is a cognitive bias that occurs when people rely on easily accessible information to make decisions, rather than gathering all the relevant data (Tversky & Kahneman, 1973). This bias can lead to errors in judgment and can prevent people from considering all the available information.

Other types of cognitive biases include the bandwagon effect, where people tend to follow the opinions of others, even if they are incorrect or irrational; the framing effect, where people's decisions are influenced by the way information is presented to them; and the sunk cost fallacy, where people continue to invest in a project or decision, even if it is no longer rational to do so (Arkes & Blumer, 1985).

Real-life examples of cognitive bias can be seen in many different contexts. For example, in politics, people may be more likely to support a candidate who shares their beliefs, even if that candidate has a history of unethical behaviour. In investing, people may be more likely to invest in a company that has performed well in the past, even if there is no evidence that it will continue to do so in the future (Shefrin, 2002).

The impact of cognitive bias on decision-making can be significant. It can lead people to make decisions that are not in their best interests, or to miss good opportunities. It can also lead to groupthink, where people within a group tend to conform to the opinions of others, even if those opinions are incorrect or irrational (Janis, 1982).

To overcome cognitive bias, it is important to be aware of its existence and to actively seek out information that contradicts our pre-existing beliefs (Lilienfeld, Ammirati, & Landfield, 2009). It is also important to consider multiple perspectives and to question our assumptions (Stanovich & West, 2008). By doing so, we can make better decisions and avoid the pitfalls of cognitive bias.

In conclusion, cognitive bias is a pervasive issue that affects human decision-making in many different contexts (Kahneman, 2011). By understanding the different types of cognitive biases and their impact on decision-making, we can take steps to overcome them and make better decisions.

## 2.3. Impact of bias on individuals and society.

The impact of bias on individuals and society is significant, leading to both obvious and subtle forms of discrimination and stereotyping (Steele, 2010). This impact can be seen in many aspects of people's lives. These aspects are presented below:

By restricting their possibilities, skewing their perspectives, and maintaining inequity, bias may harm individuals. When someone is biased, they could experience exclusion, discrimination, or unjust treatment because of their colour, gender, religion, sexual orientation, or other traits. This may result in low self-esteem, a sense of exclusion, and an awareness of unfairness. Additionally, bias may influence people's attitudes and beliefs, which can result in the internalization of stereotypes and biases that serve to reinforce bias.

Furthermore, prejudice can affect how decisions are made, resulting in less-than-ideal results in processes including employment, promotion, policy-making, and media portrayal. It can bolster preconceptions, propagate stereotypes, and obstruct the quest of justice, fairness, and equality. Society may lose out on the many views, skills, and contributions of excluded groups by maintaining biased narratives and uneven power relations.

In order to combat bias, efforts must be made to promote inclusion, diversity, and equal opportunity. It entails identifying and overcoming one's own prejudices, creating empathy and understanding amongst other groups, or putting into practice laws and procedures that uphold justice and equality. People and society may endeavour to create a more inclusive, egalitarian, and peaceful environment by decreasing prejudice.

Besides the examples above, the project is also use-case oriented and seeks to take feedback from the pilot partners. In this regard, some impact examples inspired by the Use Cases are presented below:

Using AI for Sustainable Development and the European Green Deal: Bias can significantly influence the effectiveness of AI in sustainable development initiatives and large-scale environmental plans like the European Green Deal. For instance, if the data used to train AI models is biased, it could lead to skewed predictions and recommendations, which could hamper progress towards sustainability goals. On the other hand, addressing and mitigating bias in AI can enhance its potential in driving sustainable practices and influencing environmentally-friendly policies (Holstein, Wortman Vaughan, Daumé III, Dudik, & Wallach, 2019).

Policies for Sustainable Water Cycle Management at a Large Scale: Bias, particularly in the form of systemic or structural bias, can influence the effectiveness and fairness of water management policies. For example, biases in the allocation of water resources can lead to disparities in access to clean water, affecting certain communities more than others. Acknowledging and addressing these biases is critical for developing and implementing equitable and sustainable water management policies (Cook, & Spray, 2012).

Trustworthy Data-Driven Touristic Policies: In the context of tourism, bias can impact how data is collected, interpreted, and used in policy-making. Biased data can lead to policies that favour certain regions or types of tourism, which could have various economic and environmental implications. Ensuring that data-driven policies are based on unbiased data can contribute to more equitable and sustainable tourism practices (Zamfir & Corbos, 2015).

## 2.4. Underrepresented groups

Going a bit deeper into the characteristics of the people experiencing biases and discrimination, usually they belong to underrepresented groups, or communities who are not adequately represented in public administration and governance. These groups may include ethnic minorities, migrants, religious groups, persons with disabilities, and others who face systemic barriers to equal opportunities and representation. In this section, an overview of the concept of underrepresented groups is provided, including how they are identified, the challenges they face in public administration and governance, and several strategies for addressing bias and increasing diversity and inclusivity.

### 2.4.1. Identification of Underrepresented Groups

Underrepresented groups can be identified through various means, including demographic data, representation in decision-making positions, and experiences of discrimination and marginalization (Government Accountability Office, 2020). For example, ethnic minorities and migrants may be underrepresented in political decision-making positions, while persons with disabilities may face barriers to employment and education (United Nations, 2020). Identifying underrepresented groups is essential for addressing systemic issues of bias and promoting diversity and inclusivity.

Online public services have become increasingly prevalent in modern societies, offering convenience and efficiency in accessing various governmental resources and services. However, it is crucial to acknowledge that not all individuals have equal access to, or proficiency in, using these services. Certain underrepresented groups face unique challenges and barriers that can hinder their ability to fully utilise online government services. In order to address these issues effectively, it is essential to identify and include these underrepresented groups in surveys examining problems associated with using such services (UNDP, 2021). AI4Gov aims to explore some of the key underrepresented groups that should be included in the surveys that will be conducted, highlighting their specific challenges and potential solutions.

One of the most prominent underrepresented groups towards technological developments is the elderly population. The elderly population often faces difficulties in navigating online government services due to limited digital literacy skills and unfamiliarity with technology. Many older adults may lack access to computers, smartphones, or reliable internet connections. Moreover, age-related visual impairments, cognitive decline, and mobility limitations can further hinder their ability to use online platforms effectively. Surveying the elderly population can help identify the specific challenges they encounter, such as complex user interfaces, small font sizes, or insufficient assistance options. Addressing these issues may involve designing user-friendly interfaces, providing accessible support, and promoting digital literacy programs tailored to seniors.

Another group experiencing discrimination when it comes to online services are the low-income households. Households with low incomes often face financial barriers to accessing the necessary digital devices and internet services required for online government interactions. Moreover, they may lack the resources to acquire the skills needed to navigate complex online systems effectively. Surveying this group can shed light on challenges related to affordability, limited connectivity options, and inadequate training opportunities. Potential solutions could involve expanding access to affordable internet services, providing subsidized devices, and offering digital literacy programs targeted at low-income households.

Online services can be also challenging for people with disabilities, such as vision disparities, or cognitive disparities. They encounter unique barriers when accessing online government services, such as inaccessible websites, lack of alternative formats for information, or inadequate support for assistive technologies. It is crucial to include this group in surveys to understand the specific barriers they face and to identify opportunities for improvement. Implementing accessibility standards, conducting regular accessibility audits, and providing alternative means of access (e.g., phone or in-person assistance) can significantly enhance the usability of online government services for people with disabilities.

Another important underrepresented group to be included in the surveys are people from linguistic and cultural Minorities. Linguistic and cultural minorities may face language barriers when interacting with online government services. Inadequate translation services, insufficient multilingual support, and cultural insensitivity can impede their access to vital information and services. Surveying this group can help identify the linguistic and cultural challenges they encounter and explore solutions such as expanding language options, improving translation services, and promoting culturally sensitive user experiences.

Inclusive surveys examining the problems associated with using online government services must consider the needs and challenges faced by underrepresented groups. By including the elderly population, low-income individuals, people with disabilities, and linguistic and cultural minorities, policymakers and service providers can gain valuable insights into the barriers these groups face. The findings from such surveys can inform the development of user-friendly interfaces, improved accessibility features, and targeted support programs that ensure equitable access to online government services for all citizens. Ultimately, addressing the concerns of these underrepresented groups will contribute to a more inclusive and accessible digital government landscape.

In AI4Gov, the results of this survey will be used as inputs and provisions to the HRF, in order to create a solid bias-free framework that will guide the technical partners of the project, towards structuring innovative and inclusive AI tools. As a starting point, and in order to align this work with the pilot activities of the project, the first target groups will be selected based on the specific characteristics of the pilots, focusing on the underrepresented groups they are more prominent in their specific use-cases. Moving forward from this, more types of underrepresented groups will be included according to the availability and accessibility of the project.

### 2.4.2. Challenges Faced by Underrepresented Groups

Underrepresented groups face a range of challenges in public administration and governance, including discrimination, marginalization, and unequal opportunities. These challenges can manifest as conscious or unconscious biases, which can have significant consequences for individuals and society. For example, unconscious bias against ethnic minorities and migrants can lead to discriminatory hiring practices and unequal opportunities in education and healthcare (European Network Against Racism, 2020). Similarly, unconscious bias against persons with disabilities can lead to stigmatization and discrimination in the workplace and society as a whole.

### 2.4.3. Introduction to the sources of bias

Bias can originate from various sources, including personal experiences, cultural influences, and social norms. Personal experiences can shape our attitudes and beliefs towards others, based on our past interactions with them (Greenwald & Banaji, 2017). Cultural influences can also contribute to bias, as we are often exposed to certain beliefs and values that are prevalent in our society (Jost & Kay, 2010). Social norms, such as stereotypes and prejudices, can also contribute to bias, as they can influence our perceptions and attitudes towards others (Rudman & Ashmore, 2007). A complicated interaction of several elements that affect people's perceptions, beliefs, and attitudes can result in bias. The following are descriptions of some of the major bias-causing factors:

Personal Experiences (Schug, Alt, & Klauer, 2016): Bias may be greatly influenced by personal experiences. A person's opinions and attitudes about other groups of people can be shaped by the experiences, interactions, and observations they have throughout their lives. This can lead to generalizations and prejudices with particular groups, whether positive or bad.

Cultural Influences (Markus & Kitayama, 2010): Culture has a big impact on how biased people are. People's perceptions and assessments of others can be influenced by cultural norms, values, and beliefs. Stereotypes and prejudices can be strengthened and perpetuated by cultural messages that are spread through media, literature, education, and family. This includes associating ethnic groups with specific characteristics (e.g., black people are more violent and have a tendency for crime), or it can be connected to religion (e.g., distrust towards Muslims due to terrorism).

Socialization Procedures (Aboud & Doyle, 1996): Socialization procedures, including upbringing, family relationships, and peer pressure, are very important in the formation of biases. From their families, friends, and communities, children pick up societal conventions, stereotypes, and prejudices. Individuals may absorb and display prejudices if certain views are prominent in their social surroundings.

Social Norms and Conformity (Asch, 1951): Social norms and a drive for conformity both have the potential to promote bias. For the sake of gaining acceptability or avoiding social repercussions, people may conform to the prevalent prejudices within their social groupings. This uniformity has the potential to replicate discriminatory attitudes and prejudices across generations.

Cognitive Processes (Hamilton & Gifford, 1976): Some cognitive processes may boost biases develop. Heuristics and other cognitive shortcuts help our brains digest information more quickly, but they can also introduce biases. Biased ideas and judgements can be formed and reinforced by cognitive processes including classification, stereotyping, and availability heuristics.

Systemic variables (Pager & Shepherd, 2008): Institutional practices, rules, and power structures are only a few examples of systemic variables that might have an impact on bias. Systemic biases may spread throughout many society institutions, such as the healthcare, employment, and education sectors, resulting in inequities and uneven treatment based on specific traits or group membership.

It is crucial to understand how these variables interact and how different people and circumstances may be affected by them. For the sake of raising awareness, combating prejudices, and creating inclusion and equality in society, it is crucial to recognize and deal with these contributing elements.

### 2.4.4. Common environments where individuals may perceive discrimination

Discrimination can occur in various environments and can have a significant impact on individuals and communities. Discrimination can occur based on race, gender, sexual orientation, age, religion, and disability, among other factors. At this point it is useful to go a bit deeper on some indicative common environments where individuals may perceive discrimination, along with specific examples of discrimination that may occur in each of these environments. These environments include education, healthcare, and employment, where bias may have a major influence on decisions, behaviours, and results.

Education (Nieto, 2000) (Sue, Rasheed, & Rasheed, 2016): Teacher Prejudice: Expectations, assessments, and relationships between instructors and students can all be impacted by bias. There may be differences in educational achievements for students from underrepresented groups because of lower expectations, less opportunities, and biased disciplinary measures.

Standardized Test Bias: Standardized tests may contain bias that is either culturally prejudiced or that favours some groups over others. Inaccurate evaluations of pupils' ability and uneven access to educational opportunities may come from this.

Threat from stereotypes: Students from underrepresented groups may struggle academically, contributing to success disparities, since they are afraid of confirming unfavourable perceptions.

Healthcare (Burgess, van Ryn, Dovidio, & Sah, 2007) (Smedley, Stith, & Nelson, 2003): Diagnostic Bias: Bias can affect the diagnostic procedures used by medical personnel, resulting in differences in the identification and treatment of certain illnesses across various demographic groups.

Treatment Bias: Bias can influence how treatments are chosen and how healthcare is delivered. A subpar level of treatment, lengthier wait times, or underuse of pain management may be experienced by some demographics.

Bias can lead to health inequalities, which are situations in which underprivileged people have less access to healthcare, an unequal allocation of resources, and hurdles to receiving high-quality treatment.

Employment (Pager, Western, & Bonikowski, 2009) (Greenwald & Krieger, 2006): Bias in hiring: Bias can affect the recruiting and selection procedures, resulting in differences in hiring results. Inequalities in the workplace can be sustained through unconscious prejudices that have an impact on judgments based on traits like ethnicity, gender, age, or perceived beauty.

Promotional bias: When bias enters the picture, the chances for progress for members of underrepresented groups are reduced. Wage and Pay Equity Marginalized groups may receive less for the same labour or are underrepresented in higher-paying professions, which can lead to wage gaps and pay disparities.

Bias may prevent social mobility, reinforce entrenched inequities, and reduce chances for disadvantaged people and groups in all of these contexts. Promoting fairness, equitable chances, and social justice requires recognizing prejudice and correcting it via training, legislative reforms, and the creation of inclusive settings.

Factors that contribute to discriminatory practices include prejudice, stereotypes, and systemic inequalities. Discrimination can have a significant impact on those affected, including emotional distress, reduced opportunities, and physical harm. Discrimination can also perpetuate systemic inequalities and contribute to social and economic disparities. It is important to address discrimination in all environments and work towards creating inclusive and equitable communities.

### 2.4.5. Conclusion

This chapter covered the idea of prejudice and how it affects both individuals and society. Unfair prejudice or predisposition in favour of, or against a certain individual, group, or item is referred to as bias. Different types of prejudice, such as cognitive bias (systematic patterns of deviance from reasoned judgment), implicit bias (unconscious biases influencing perceptions and behaviours), and explicit bias (conscious and overt biases) were analysed.

Furthermore, it has be examined how prejudice might influence choices, actions, and results in a variety of contexts, including employment, healthcare, and education. Bias in education can affect students' expectations and assessments of instructors, lead to inequities in standardized test results, and maintain success inequalities. Bias in healthcare can have an influence on diagnosis, choices for treatments, and health inequalities across various demographic groups. Bias in the workplace can impact recruiting and promotion choices, and can contribute to pay disparities and income gaps.

Various causes of prejudice, such as individual experiences, cultural influences, societal norms, cognitive processes, and systemic issues, were also investigated. These elements interact to shape people's prejudices and uphold societal injustices.

Promoting justice, inclusion, and equality in society depends on recognizing prejudice and dealing with it. By combating bias, people and organizations may endeavour to build inclusive workplaces that appreciate diversity, guarantee fair treatment, and promote equitable opportunity. This entails fostering empathy, knowledge, and understanding as well as putting into practice strategies and regulations that combat systemic prejudices.

# 3. The Impact of Bias in AI4Gov Pilot Cases: Addressing Ethical and Societal Implications

The use of Artificial Intelligence (AI) in government processes has the potential to revolutionize the sustainability approach, water management, and tourism policies. However, AI systems are not immune to bias, which can have significant consequences for the effectiveness and fairness of these pilot cases. Research has shown that bias in AI systems can have significant consequences. For example, a study by the National Institute of Standards and Technology found that facial recognition algorithms were more likely to misidentify people of colour and women (Grother et al., 2019). Another study by the AI Now Institute found that predictive policing algorithms were more likely to target marginalized communities (Hao, 2019).

In this section, the impact of bias in three AI4Gov pilot use-cases is introduced, with a specific focus on the potential ethical and societal implications of such bias.

### *3.1.1.1. Overview of Pilot Cases*

- Pilot #1: Utilizing AI for Sustainable Development and the European Green Deal with partners at Institut "Jožef Stefan": The objective of this pilot is to use AI to identify areas of sustainable development and implement policies that align with the European Green Deal.
- Pilot #2: Implementing Policies for Sustainable Water Cycle Management at a Large Scale with partner at Diputación provincial de Badajoz (DPB): The objective of this pilot is to use AI to manage the water cycle in a sustainable and efficient manner, while addressing the needs of various stakeholders.
- Pilot #3: Creating Trustworthy Data-Driven Touristic Policies with partners at the Municipality of Vari, Voula, Vouliagmeni (VVV) and Greek Ministry of Tourism (MT): The objective of this pilot is to use AI to analyse tourist data and create policies that promote sustainable tourism while ensuring the safety and satisfaction of tourists.

### *3.1.1.2. Potential Consequences of Bias*

Bias in AI systems can have significant ethical and societal implications for each of these pilot cases. For example, in Pilot #1, biased data could result in policies that disproportionately impact certain communities or fail to address important sustainability issues. This would be the production of policies that would be western-oriented, neglecting the specific characteristics of the countries fighting to achieve SDGs and the European Green Deal.

In Pilot #2, bias in the data could result in inefficient water management practices or policies that do not consider the needs of all affected areas. This can lead to poorer water quality and water treatment in certain neighbourhoods or municipalities.

In Pilot #3, bias could lead to discriminatory policies that exclude certain groups of tourists or fail to address important safety concerns. This can lead to both lower quality of the touristic services of the municipality, as well as it can significantly affect the residents' lives in a negative way.

#### *3.1.1.3. Digital bias and digital exclusion*

Digital exclusion is the lack of access or ability to use digital technologies such as internet, computers, and mobile devices. It is prevalent among socially disadvantaged groups such as the elderly, low-income households, people with disabilities, and individuals living in rural or remote areas (Ragnedda & Ruiu, 2019). Digital exclusion has far-reaching consequences, including a lack of access to information, education, employment opportunities, and health services.

The causes of digital exclusion are numerous and complex. Income, education, age, geographic location, disability, and language are some of the factors that contribute to digital exclusion (Van Dijk, 2013). Insufficient infrastructure, lack of basic digital skills and education, and the high cost of data plans and devices are also significant barriers (Selwyn, 2004).

The consequences of digital exclusion range from a lack of access to information and services to the perpetuation of social inequalities. Digital exclusion can also lead to social isolation and contribute to economic disparities. Being unable to access the internet limits individuals' ability to participate in online communities, which can have a detrimental impact on mental health (Kim, 2017).

On the other hand, Digital bias refers to the biases that are embedded in digital technologies. Digital bias can manifest in various forms, such as algorithmic bias and data bias. Digital bias differs from digital exclusion in that it affects individuals who have access to digital technologies but are still subject to prejudice and discrimination (Caliskan et al., 2017).

For instance, biased algorithms can reinforce negative stereotypes and perpetuate systemic inequalities. Similarly, data bias can manifest when data used for analysis is skewed or incomplete, resulting in biased decision-making processes. Digital bias can lead to unfair outcomes and the perpetuation of social injustices.

Digital exclusion and digital bias are significant issues that need to be addressed to promote equality and access to digital technologies. Understanding their causes and consequences is crucial to developing effective solutions.

#### *3.1.1.4. Recommendations for Addressing Bias*

To address and mitigate bias in AI systems, it is important to take a proactive and comprehensive approach. Some recommendations for addressing bias in the AI4Gov pilot cases include:

- Diversifying the data: Ensuring that the data used in AI systems is diverse and representative of all communities and stakeholders.
- Testing for bias: Conducting regular testing to identify and address any biases in the AI systems.
- Promoting transparency: Providing transparency in the development and implementation of AI systems to ensure accountability and trust.
- Encouraging collaboration: Encouraging collaboration between stakeholders to ensure that the AI systems are developed and implemented in a way that is fair and equitable.

### 3.1.2. Detecting and addressing bias: Provide tips and strategies for recognizing and mitigating bias, including self-reflection, education, and diversity and inclusion initiatives.

To address and mitigate bias, it is important to understand its sources and effects and implement targeted strategies and interventions. Some strategies that have been shown to be effective in addressing bias include:

- Increasing awareness and education: Providing education and awareness programs that highlight the sources and effects of bias can help individuals recognize and address their own biases (Greenwald & Banaji, 2017).
- Encouraging diversity and inclusivity: Promoting diversity and inclusivity in various contexts, such as schools and workplaces, can help reduce bias by exposing individuals to different perspectives and experiences (Pettigrew & Tropp, 2006).
- Using objective criteria: Using objective criteria in decision-making processes, such as hiring and promotion, can help reduce bias by removing subjective factors that can contribute to bias (Rudman & Ashmore, 2007).
- Encouraging empathy: Encouraging empathy towards others can help reduce bias by promoting understanding and compassion towards individuals who may be different from us (Jost & Kay, 2010).
- Providing feedback and accountability: Providing feedback and accountability can help individuals recognize and address their own biases by holding them accountable for their actions and decisions (Paluck & Green, 2009).

In conclusion, bias is an inherent part of human nature that can influence our perceptions, decision-making, and interactions with others. To address and mitigate bias, it is important to understand its sources and effects and implement targeted strategies and interventions. In the subsections below, examples and input from experts is provided, after their participation in workshops and discussions that explore Bias in AI.

### 3.1.3. Exploring Bias in AI: Insights and Reflections from the 1st AI4Gov Training Workshop

The AI4Gov project held its 1st Training Workshop, "Bias In AI", on May 16, 2023, in Ljubljana, Slovenia, kindly hosted by the Jožef Stefan Institute. The first session focused on the fundamentals of AI and bias, as well as the impact of bias on human rights, especially for underrepresented groups. In the second session, participants attended case study presentations and participated in practical exercises, divided into groups. Almost 20 participants joined us in person and another 30 via remote access, representing the European Commission, Academia, IT Industry and Public Organisations. The subsections below, describe the scenarios and the process that was followed during the training activities.

#### *3.1.3.1. SCENARIO 1: social services automation of child care subsidised.*

Aim: social service workers are overwhelmed by the amount of paperwork they have to deal with. The idea is to build an automated system that will ingest the data users put in the system through a custom-made user interface and build an algorithm that will automatically decide the eligibility and the amount of childcare subsidized based on historical data.

Types of bias in the given scenario: Historical biases/User generated biases: Race, discrimination, socioeconomic, gender, religion

**How the bias can affect the users:**

**Citizens**

1. Citizens might not apply in the first place, because the biases make them believe they are not eligible;
2. Citizens might not be able to apply due to lack of access to information and digital skills;
3. Citizens might be rejected due to false applications;
4. The rejection might lead to financial issues;
5. Disruption of social cohesion/social integration/tension among natives and immigrants.

**Policy makers**

1. Discriminating policies due to lack of evidence informed tools to support the decision-making processes.

**Social workers**

1. More workload to detect and manage the issues that will arise;
2. Untrust towards the tool due to the biased information;
3. Rise of complaints affecting their mental health/distress.

#### *3.1.3.2. SCENARIO 2: computer vision recognition on a metro.*

Aim: lots of thefts are happening on the metro stations. City wants to prevent them by placing cameras on the metro stations and by utilizing a private company-developed computer vision algorithm that can recognize the faces and match them with the official records. What could go wrong?

**Types of bias in the given scenario:**

- Bias in the training datasets:
  - Some population groups may be underrepresented. In general, white male faces are overrepresented. This will lead to higher performance rates for these groups.
- Historical bias:
  - Datasets used to build the algorithm is based on datasets from the past; thus, not referring to present situations.
- Anchoring bias (Reliance on the first piece of information encountered when making decisions):
  - What if the algorithm makes a false identification? Should we rely on the first identification of the suspect?
- Bias in deployment:
  - Overreliance on results (automation bias): where a particular result is given, law enforcement authorities may over rely on them;
  - Differences in performance (related to the unbalanced datasets) may affect different populations differently. Errors rates can be higher for underrepresented groups;
  - Attribution of behaviours to specific types/categories/featured individuals.

#### *3.1.3.3. SCENARIO 3: recommender system on municipality website.*

Aim: Municipality of a smal city is trying to reach out to their inhabitants. They have updated their website (desktop and mobile friendly) and want to develop a recommendation system that provides news and information based on browsing- history. Their idea is to classify people into seniors, adults, adults with young kids, and teenagers. For this job, they hired a recent CS graduate.

**Bias can creep into the recommendation system at various stages:**

Data Collection: The browsing history of users can be a source of bias. If users have been predominantly exposed to specific types of content due to previous website structure or content availability, it may not reflect their actual interests or needs (Scheufele & Krause, 2019).

Algorithm Design: When designing the algorithm, the CS graduate might inadvertently introduce biases. For instance, a common pitfall could be confirmation bias, where the designer unconsciously structures the algorithm to reinforce their own pre-existing assumptions about each user group (Nickerson, 1998).

User Classification: The choice to classify users into specific categories (seniors, adults, adults with young kids, and teenagers) could itself be a source of bias, known as representational bias. This approach may lead to overgeneralization, where people within each category are assumed to have the same preferences, which is not always the case. Additionally, the algorithm may also exclude those who don't neatly fit into these categories (Bowker & Star, 2000).

Evaluation Metrics: The choice of metrics to evaluate the performance of the recommendation system could also introduce bias. For example, if the primary metric is to increase user engagement, the system may become biased towards recommending more polarizing or sensational content, which may not necessarily be the most beneficial or relevant to the user (Ekstrand et al., 2018).

These biases can affect users in several ways. They might feel stereotyped or misunderstood if they continually receive recommendations that don't resonate with their personal interests or needs. On a broader scale, this might lead to an information bubble or echo chamber, where users are only exposed to a narrow range of content, reinforcing their current views, and limiting exposure to diverse perspectives (Pariser, 2011).

Mitigating these biases is essential for a fair and effective recommendation system. Strategies might include diverse data collection, user-informed algorithm design, flexible user categories, and a broad range of evaluation metrics that value user satisfaction and diversity of content, as well as engagement.

### 3.1.4. "Workshop Summaries and Expert Insights: Addressing Human Rights, AI, Travel Categories, and Biases - Insights from a ViLabs workshop"

Workshop on Travellers' Categories and Biases: Addressing Ethical and Legal Considerations in Data Collection at Borders: The Workshop on Travellers' Categories and Biases brought together experts to delve into the ethical implications of AI-based data collection at borders. Esteemed speakers included the President of the Policy and Citizens' Observatory, specializing in migration, refugee, and border policies. A Professor of Gender, Migration, and Citizenship at Middlesex University provided insights on the intersection of gender, migration, and AI technologies. An expert on innovation and technology policies, discussed the impact of AI and big data in border control.

The Coordinator of ESWA European Sex Workers' Rights Alliance, shed light on the concerns and biases faced by sex workers.

Challenges in developing unbiased AI technologies: AI systems can discriminate against specific categories of people, which can result in violations of fundamental human rights and erode trust in AI technologies.

**According to the workshops conducted, the following key challenges were highlighted:**

1. Potential underrepresentation of certain target groups, particularly minorities, in the data collection process. The lack of diversity can result in biased outcomes and inaccurate decision-making, negatively affecting individuals who fall outside the dataset.
2. The quality of data used is paramount as unreliable or inaccurate data can have severe consequences, hindering the effectiveness of AI applications.
3. Protecting the privacy and fundamental rights of underrepresented groups is essential to ensure AI technologies align with principles such as non-discrimination, equal treatment, and the right to privacy.

The workshops also uncovered the following practices in order to address bias in AI development:

1. Input from the general public: Incorporating public feedback through workshops to amplify the voices of minority groups and address their concerns.
2. Data Quality Assurance: Using disaggregated data to evaluate how different individuals perceive border control technologies, ensuring accurate and reliable outcomes.
3. Identify and mitigate bias at every stage of AI development: Implementing a comprehensive plan to identify and mitigate bias at every stage of AI development, promoting reliability and objectivity.
4. Generate diverse AI development teams: Forming diverse AI development teams with representation from various backgrounds, including women and individuals with disabilities, to foster inclusivity and minimize prejudice.
5. Transparency and accountability: Adopting transparent and accountable practices in the development and use of AI technologies, prioritising data protection, privacy, and non-discrimination standards.

### 3.1.5. Conclusion: Summarize the key points and emphasize the importance of addressing bias for promoting fairness, inclusivity, and equality.

Discrimination can occur in various environments, including employment, education, healthcare, housing, public services and institutions, and social interactions and community settings. Discrimination can take many forms, including unequal pay, lack of access to resources, bullying, and exclusion. Discrimination can occur based on race, gender, sexual orientation, age, religion, and disability. Factors that contribute to discriminatory practices include prejudice, stereotypes, and systemic inequalities. Discrimination can have a significant impact on those affected, including emotional distress, reduced opportunities, and physical harm (Pager & Shepherd, 2008; Smedley, Stith, & Nelson, 2003; Williams, 2012).

Discrimination can also perpetuate systemic inequalities and contribute to social and economic disparities (Alexander, 2010; Bielby & Baron, 2003). Discrimination can have a significant impact on those affected, including emotional distress, reduced opportunities, and physical harm (Pager & Shepherd, 2008; Smedley, Stith, & Nelson, 2003; Williams, 2012). It is important to address discrimination in all environments and work towards creating inclusive and equitable communities. By addressing bias, fairness, inclusivity, and equality for all individuals and communities is promoted (Dovidio, Gaertner, & Kawakami, 2003; National Fair Housing Alliance, 2019; Russell & McGuire, 2008).

In addition to addressing bias and discrimination through education and policy changes, technology can also play a role in promoting fairness, inclusivity, and equality. One potential solution is represented by the outcome of the AI4Gov project, an artificial intelligence platform that aims to eliminate bias in government decision-making processes. The platform uses machine learning algorithms to analyse data and identify patterns of bias and discrimination. By identifying these patterns, AI4Gov can help government agencies make more informed and equitable decisions.

Furthermore, AI4Gov can also help promote transparency and accountability in government decision-making processes. By providing insights into decision-making processes, AI4Gov can help identify areas where bias and discrimination may be occurring and provide opportunities for intervention and correction. Additionally, AI4Gov can help ensure that government decisions are based on objective data and evidence, rather than subjective biases or assumptions.

Overall, AI4Gov has the potential to play a key role in promoting fairness, inclusivity, and equality in government decision-making processes. By leveraging the power of artificial intelligence and addressing bias and discrimination in all environments, a more just and equitable society can be developed for all individuals and communities (Dovidio, Gaertner, & Kawakami, 2003; National Fair Housing Alliance, 2019; Russell & McGuire, 2008).

# 4. The Interview and Survey Process

This chapter presents the methodological steps that will be followed to conduct the surveys on the targeted underrepresented populations to support the HRF of AI4Gov.

## 4.1. Overview

In the context of WP2 and the creation of the HRF, a survey will be conducted on underrepresented groups, in order to include their perspective in the HRF and minimise as much as possible potential biases and discrimination at the source. This survey will complement the literature review and research that will feed the HRF by implementing a bottom-up approach. Surveys to underrepresented groups (ethnic minorities; migrants; religious groups; persons with disabilities etc.) will tackle the question of how and in which environments individuals perceive discrimination and difficulties accessing various services or information. Furthermore, the surveys will focus on the perceived causes of discrimination (lack of interest, lack of information and raising awareness, less skills etc.), but also on the specific form of discrimination (understood as outcome such as the denial of a credit, a higher price for a certain service or longer waiting times). The survey will also assess whether existing services or platforms comply with existing EU regulations on human rights protection, by interviewing people to identify in which areas discrimination might be overlooked. The methodology of the interviews and surveys orients itself to earlier quantitative studies on discrimination as well as theoretical literature on factors of discrimination, based on what is acknowledged by the EU as fundamental rights and values.

## 4.2. Focus and Topics of the Interviews

AI4Gov will set the aims and goals of the survey to guide the process and better address the Objectives of the Survey on Underrepresented Groups. This will facilitate the process of creating the appropriate questions in order to map the input from the underrepresented groups in a representative way. The objectives are presented below:

Identify Discrimination Perception:

- Understand how individuals from underrepresented groups perceive discrimination in accessing various services or information through online government platforms.
- Determine the specific environments or contexts in which discrimination is perceived, such as government websites, online application processes, or service portals.

Explore Difficulties in Accessing Services or Information:

- Investigate the challenges faced by underrepresented groups when accessing online government services or information.
- Determine the specific barriers encountered, such as limited digital literacy, lack of internet access, or inaccessible user interfaces.

Examine Perceived Causes of Discrimination:

- Investigate the factors perceived by individuals from underrepresented groups as causes of discrimination in accessing online government services.
- Explore potential causes, including lack of interest or commitment to addressing the needs of underrepresented groups, insufficient information and awareness, or inadequate digital skills training.

Identify Specific Forms of Discrimination:

- Understand the specific outcomes of discrimination experienced by underrepresented groups in their interactions with online government services.
- Investigate instances of denial of credit, higher pricing for certain services, longer waiting times, or other forms of discrimination.

By structuring the survey around these objectives, the project team can gain valuable insights into the discrimination perception, access difficulties, perceived causes, and specific forms of discrimination experienced by underrepresented groups when using online government services. Such findings can guide policymakers, service providers, and advocates in developing targeted interventions and improvements to ensure equitable access and address discriminatory practices.

## 4.3. Compliance with EU Regulations on Human Rights

When conducting surveys on underrepresented populations, there are several regulations and human rights provisions that need to be taken into account. The specific regulations and provisions may vary depending on the country or region, and these will be thoroughly investigated when the specific target groups are selected. However, as general considerations, AI4Gov will take into account the following:

- Informed Consent: Obtaining informed consent is a fundamental requirement for any research involving human subjects. Researchers must ensure that participants understand the purpose, risks, benefits, and confidentiality of the survey. For underrepresented populations, such as people with disabilities, or refugees, additional safeguards may be necessary to ensure that their consent is fully informed and voluntary. The consent forms will be distributed and explained at the beginning of the survey to make sure the participants are aware of the project, its objectives and the rights they have over their data.

- Privacy and Confidentiality: Protecting the privacy and confidentiality of survey participants is crucial. Data collected from underrepresented populations should be treated with extra care and stored securely. Researchers must take measures to de-identify data whenever possible and ensure that participants cannot be identified or harmed as a result of their participation. In this regard, the team will apply an anonymisation system to de-associate the questionnaires from the responder.
- Non-discrimination: Surveys should be designed and conducted in a manner that avoids discrimination and ensures equal treatment of all participants. It is important to be sensitive to cultural, ethnic, religious, or other factors that could influence survey responses or create biases.
- Protection from Harm: Researchers should take steps to minimise any potential harm or distress to survey participants. This is particularly important when working with underrepresented populations who may be more susceptible to emotional or psychological harm. Adequate support and referral mechanisms should be in place to address any negative consequences that may arise. In order to achieve this point, along with the previous one, the team will pursue close collaboration with people working with the selected underrepresented groups, in order to better accommodate their needs.
- Special Considerations for Underrepresented Groups: Underrepresented populations require additional protections due to their unique circumstances. These groups may include refugees, people with disabilities, and marginalised communities. The AI4Gov team will be aware of any specific regulations or guidelines that apply to working with these populations and ensure their inclusion and meaningful participation in the research process.

After considering these general provisions, it is essential to consult relevant local laws, guidelines, and ethical frameworks specific to the region where the survey will be conducted to ensure compliance with all applicable regulations and human rights provisions. This will be done upon the selection of the specific underrepresented groups the survey will engage.

## 4.4. Identifying Areas of Overlooked Discrimination

"Overlooked discrimination areas" refer to instances of discrimination that often go unnoticed or unrecognized by individuals, even if they are personally experiencing them. These are situations where discriminatory opinions, biases, or behaviours are not easily identified or acknowledged by the affected individuals or the broader society.

In the context of the AI4Gov survey on problems with using online government services, identifying, and including underrepresented groups is crucial. However, it is equally important to go beyond this identification and explore the areas where discrimination may occur but remain overlooked. These areas may not be immediately apparent or recognised by those experiencing them, highlighting the need for a comprehensive investigation to uncover subtle forms of discrimination.

The survey provides an opportunity to map these overlooked discrimination areas. Through targeted questions, participants can share their experiences, perspectives, and observations regarding instances of discrimination that might not be readily evident. The data collected during the survey can reveal patterns and highlight areas where discriminatory practices or biases are present but often disregarded.

Once the survey and focus groups are conducted, the findings related to overlooked discrimination areas can be presented, along with reflections and conclusions. This presentation aims to raise awareness and shed light on subtle forms of discrimination that may have been previously unacknowledged. It allows for a deeper understanding of the complexities and nuances of discrimination faced by underrepresented groups when using online government services.

By highlighting these overlooked discrimination areas, policymakers, service providers, and society can work towards addressing systemic biases, improving accessibility, and implementing measures to ensure equal treatment and opportunities for all individuals, regardless of their vulnerability. The reflections and conclusions drawn from the survey and focus groups can inform policy changes, awareness campaigns, and interventions aimed at mitigating discrimination and promoting inclusivity within online government services.

## 4.5. Methodology of the Interviews and Surveys

This section provides an overview of the planning and consideration to ensure ethical practices, data privacy, and inclusivity. It includes the outline of the survey plan to guide the process:

1. Define the Objectives: First, the goals of the survey will be identified and the specific information that needs to be gathered from the target groups will be established. These objectives are already presented in section 3.3.
2. Identify the Target Population: The second step is to clarify the specific underrepresented population which needs to be included in the survey and the focus groups. This will help to understand their characteristics, needs, and potential barriers to participation. In the case of AI4Gov, this will be inspired by the pilot use-cases, as it has been described in section 3.2.
3. Ethical Considerations: A crucial step is ethical considerations and the project needs to ensure that the survey adheres to ethical guidelines and protects the rights and welfare of participants. This can be done through obtaining the necessary approvals from relevant authorities, such as institutional review boards or ethics committees. This will be defined based on the groups that will be selected.

4. Survey Design: After taking under consideration the above steps, the survey methodology that best suits the target population and their accessibility will be selected (e.g., online, phone, in-person). The goal is to create clear and concise survey questions that are relevant and unbiased. An important factor to this is to use validated scales or measures to ensure the reliability and validity of data.
5. Pilot Testing: At first, a small-scale pilot test of the survey will be conducted with a sample of the target population. This will evaluate the clarity, comprehensibility, and cultural appropriateness of the survey questions. The necessary revision and refinement will be based on the feedback received. This will minimise the potential issues that might appear during conducting the survey.
6. Recruitment Strategy: An important step is to develop a recruitment plan that considers the unique characteristics and challenges of the target population. This will be done through collaboration with community organisations, advocacy groups, or trusted individuals to help reach and engage participants. A starting point will be the pilot uae-cases of AI4Gov and the connections and networks they can provide. People working with underrepresented groups will be included in the recruitment strategy to facilitate communication and cooperation.
7. Informed Consent: The survey will also develop a consent process that is culturally sensitive, easy to understand, and respects participants' autonomy. This will clearly explain the purpose of the survey, potential risks, confidentiality measures, and voluntary participation. The informed consent will be obtained from participants before they begin the survey.
8. Data Collection: After defining the methodology and implementing the survey, the data will be collected ensuring accessibility and transparency to the target population (e.g., providing language support, accommodating disabilities). This process will be also guided by facilitators who are familiar with the target population.
9. Data Analysis and Reporting: The analysis and collection of the data will be conducted using appropriate statistical methods and qualitative analysis techniques. This will ensure data confidentiality and privacy throughout the analysis process, and summarise the findings, identifying key insights and trends.
10. Dissemination of Results: After analysing the data, the results will be shared in an accessible and understandable format with the target population. The most prominent purpose of the survey is to feed the HRF of AI4Gov. In addition to this, the results can be used in the form of reports or policy briefs that can be circulated to engage with relevant stakeholders in the sector of AI and biases in e-governance. The project will use the findings to inform policy, program development, or interventions that address the needs of the underrepresented population. This way, the impact of the survey will be maximised.
11. Evaluation and Follow-up: After producing the results, the final step is to reflect on the survey process and identify strengths, weaknesses, and areas for improvement. As this is

an ongoing process, the plan will be adapted accordingly to ensure it meets the unique needs and considerations of the target population.

In conclusion, this chapter describes the general aims and objectives of the survey that will be conducted to support the HRF of AI4Gov project, while addressing the general provisions that should be taken under consideration while conducting the survey and analysing the data.

# 5. Initial analysis for the AI-based Democracy Holistic Regulatory Framework (HRF).

## 5.1. Research Design

The research design employed in this study involved the utilization of the Delphi method to develop a comprehensive SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis for the AI-based Democracy Holistic Regulatory Framework (HRF). The Delphi method, a well-established research technique, facilitated the collection of expert opinions and insights.

A panel of experts with expertise in AI governance and policy-making participated in multiple iterative rounds of data collection and analysis. In each round, the experts ranked a set of points within the SWOT categories, reflecting a wide range of factors related to the integration and regulation of AI and the use/handling of Big Data in policy management.

The responses were compiled, analysed, and shared with the experts for further refinement. The final SWOT analysis serves as a foundation for the development of the HRF, providing valuable recommendations and guidelines to address identified weaknesses and threats while leveraging strengths and opportunities.

The research design allowed for systematic exploration and aggregation of expert insights, ensuring the relevance and effectiveness of the HRF in fostering transparent and democratic governance.

## 5.2. Methodology

The Delphi method, originally introduced by Dalkey and Helmer (1963), is a widely recognized research technique used to gather expert opinions and insights. In this study, the Delphi method was employed to develop a comprehensive SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis within the context of the AI-based Democracy Holistic Regulatory Framework (HRF). The aim was to gain a deep understanding of the strengths, weaknesses, opportunities, and threats associated with AI systems in governance and policy-making organizations.

The Delphi process consisted of multiple iterative rounds, allowing for systematic data collection and analysis. A panel of experts, selected based on their expertise and experience in the field of AI governance and policy-making, participated in the study. In each round, the experts were provided with a set of ten points for each of the SWOT categories, representing a broad range of factors relevant to the integration and regulation of AI and the use/handling of Big Data in policy management.

During the first round, the experts ranked the points within each category in descending order of preference, using a numerical scale from 1 to 10. This ranking process helped determine the relative importance of each point. Additionally, the experts evaluated the significance of each point on a scale of 1 to 4, where 1 indicated "very important" and 4 indicated "not important at all". This evaluation process provided an assessment of the overall importance of each point within the SWOT analysis.

The responses from the initial round were compiled, analysed, and summarized. The anonymity of the Delphi method ensured independent and unbiased contributions from the experts. The summarized results were then shared with the expert panel in subsequent rounds, including areas of agreement and divergence, while maintaining anonymity.

The subsequent rounds provided the experts with an opportunity to reconsider their rankings and evaluations based on the collective input from the panel. Feedback was provided on the degree of consensus or disagreement among the experts, allowing for refinement and adjustment of responses. This iterative process continued until a level of convergence and stability was achieved among the expert opinions.

The final SWOT analysis, derived from the Delphi process, serves as a comprehensive assessment of the strengths, weaknesses, opportunities, and threats associated with AI systems in governance and policy-making organizations. Building upon this analysis, the AI-based Democracy Holistic Regulatory Framework (HRF) will be developed, aiming to address the identified threats and weaknesses while capitalizing on the strengths and opportunities. The HRF will provide valuable recommendations and guidelines for the integration and regulation of AI, as well as the handling of Big Data in policy management, with a focus on fostering transparent, democratic governance and mitigating potential risks.

In conclusion, the Delphi method, with its iterative nature and emphasis on expert consensus, was employed to generate a rigorous and comprehensive SWOT analysis for the AI-based Democracy Holistic Regulatory Framework (HRF). The application of this method allowed for the collection of diverse perspectives while ensuring anonymity and minimizing the impact of group dynamics. The insights obtained from the expert panel will inform the development of the HRF, providing actionable recommendations and guidelines for the integration and regulation of AI and the use/handling of Big Data in policy management within a democratic framework.

## 5.3. Results

### 5.3.1. Strengths

Artificial Intelligence (AI) has emerged as a transformative force across sectors, from businesses and SMEs to governmental organizations like the Greek Ministry of Tourism and the Diputación de Badajoz, as well as manufacturing companies such as Siemens. Its benefits are manifold, multifaceted, and span various levels of operations, processes, and strategic decision-making.

Firstly, AI is hailed for enhancing efficiency and productivity, given its ability to automate repetitive tasks at a pace and scale unmatched by human capabilities. By handling routine operations such as data entry, answering frequently asked questions, and monitoring the time and spatial evolution of data, AI allows employees to move to higher-value tasks, fostering creativity and innovation within the workforce. This operational optimization extends to both internal and external operations, enabling organizations to streamline processes, reduce costs, and ultimately, increase overall productivity.

Secondly, AI supports informed decision-making by providing robust data analytics capabilities. It can swiftly analyse vast amounts of complex data, identifying trends, patterns, and anomalies that would otherwise go unnoticed. AI's predictive analytics further empowers organizations to anticipate future trends and outcomes, offering valuable insights that guide strategic decisions.

The importance of AI extends to customer or visitor experience as well. By leveraging AI-powered tools such as chatbots and virtual assistants, organizations can provide round-the-clock support, significantly improving customer satisfaction. In the context of the tourism sector, AI can help gather and analyse customer data to better understand visitors' preferences and sentiments, facilitating the development of tailored experiences and offerings.

Another noteworthy advantage of AI lies in its ability to foster innovation and disruption. AI technologies have the potential to drive transformative change, enabling organizations to deliver unique services, rapidly adapt to market dynamics, and maintain a competitive edge in their respective sectors. In manufacturing, AI's role is pivotal in optimizing production, detecting faults early, and improving overall equipment effectiveness (OEE).

Moreover, AI plays a crucial role in bolstering security and managing risks. AI-driven cybersecurity systems can detect and respond to threats in real-time, safeguarding sensitive data and assets. In the financial domain, machine learning is instrumental in fraud detection and risk classification, ensuring the financial stability of an organization.

Finally, AI can expedite research and development processes by automating complex experiments, simulations, and data analyses. By enabling the efficient analysis of extensive scientific data, AI aids in the discovery of valuable insights and the optimization of research procedures.

In summary, AI offers a plethora of benefits ranging from operational efficiency, cost savings, and improved decision-making, to enhanced customer experience, advanced data analytics, and security strengthening. As the application of AI continues to expand, organizations that harness its power stand to gain a competitive advantage, driving innovation, disruption, and transformation in their respective fields.

AI, as a burgeoning field within computer science, encompasses a broad range of subdomains including machine learning, deep learning, natural language processing, and computer vision. These technologies are being increasingly used across a multitude of sectors, from healthcare and governance to business and environmental management, where they bring about profound benefits.

The key advantage of AI lies in its capacity for rapid analysis of large volumes of data - often referred to as Big Data - in real-time. It is equipped to perform tasks typically undertaken by humans, such as learning, decision-making, and problem-solving, but with significantly more efficiency. This contributes to improved productivity and reduces human error, as AI can handle the bulk of data processing and analysis, freeing up human workers to concentrate on tasks that necessitate unique human skills.

In specific sectors like healthcare, AI can provide invaluable support to professionals by offering advanced tools for real-time analysis of healthcare data. This supports personalized healthcare and public health initiatives while minimizing the possibility of human error. In the realm of governance, AI can simplify complex procedures, provide comprehensive insights into relevant data, and facilitate citizen engagement.

Furthermore, AI can facilitate better decision-making by analysing vast amounts of data and identifying patterns and insights that might not be easily detectable by humans. This can lead to optimized outcomes for both individuals and organizations. For societal and environmental issues, AI can help researchers and policymakers develop a deeper understanding. It can create models that simulate policy impacts on societal and environmental well-being and predict natural disasters or other environmental risks, thus helping to mitigate their impact.

AI's ability to reduce human error, uncover biases, and process large amounts of data effectively makes it a critical tool in contemporary decision-making processes. While the threat to privacy and personal data is a legitimate concern, AI also holds the potential to enhance data security by preventing fraudulent activities, cyberattacks, and security breaches. Privacy-focused AI tools can enhance data protection and minimize risks of data misuse and unauthorized access.

However, AI should be employed with careful human oversight and accountability to ensure that its use aligns with ethical standards and societal values. Transparency, particularly in explaining AI processes (also known as explainability), is vital for engendering trust and credibility in AI systems. Transparency enables us to identify and address biases and other problems within AI systems. Thus, while AI can dramatically enhance operational efficiency and productivity, it should always be deployed with an awareness of its ethical implications and a commitment to responsible usage.

Key benefits and advantages of using AI in an organization:

- Efficient Data Analysis: AI's capability to quickly analyse large volumes of data (Big Data) in real-time offers organizations a significant advantage by providing actionable insights.
- Optimized Decision Making: The data analysis capability of AI can also improve the decision-making process within an organization, helping identify patterns and insights that may not be easily detectable by humans, which leads to more optimized outcomes.
- Reduction in Human Error: AI can perform complex tasks with precision, which reduces the risk of human error, contributing to overall operational efficiency and reliability.
- Increased Productivity: AI's ability to automate tasks typically requiring human intervention enables employees to focus on tasks that require human creativity and critical thinking, resulting in increased productivity.
- Cost and Time Efficiency: The automation of tasks by AI also results in significant time and cost savings due to increased efficiency and speed.
- Improved Healthcare Tools: AI can support healthcare organizations by providing advanced tools for real-time analysis of healthcare data, enhancing personalized healthcare and public health initiatives.
- Simplified Governance Procedures: AI can simplify complex procedures for organizations involved in governance, provide comprehensive insights, and foster increased citizen engagement.
- Enhanced Data Security: AI can help enhance data security by preventing fraudulent activities, cyberattacks, and security breaches. Privacy-focused AI tools can further improve data protection and minimize risks of data misuse and unauthorized access.
- Transparency and Accountability: AI systems can be made transparent and accountable with proper human oversight. This feature is important for identifying and addressing biases and other issues within AI systems.
- Bias Detection and Mitigation: AI algorithms can detect and mitigate biases in decision-making processes, promoting more objective and fair outcomes.
- Environmental Monitoring and Prediction: AI can be used to monitor and predict natural disasters and other environmental risks, helping to mitigate their impact on communities.
- Understanding Societal and Environmental Issues: AI can help understand societal and environmental issues by creating predictive models based on data analysis.
- Personalized Services: In various sectors like healthcare, AI can analyse individual data in real-time to provide personalized services.

- Public Engagement: AI can facilitate citizen engagement in sectors like governance by providing comprehensive insights into relevant data and simplifying complex procedures.

### 5.3.2. Weaknesses

The use of AI in organizations, although beneficial, comes with several challenges and limitations. Notably, there is a widespread concern about the lack of training and experience among the employees in some sectors, particularly in the public sector. Despite this, any AI tools deployed by organizations have to comply with legal and ethical regulations, including laws about bias avoidance.

AI development and implementation can be costly and time-consuming, and may also lack the ability to make decisions based on emotion and creativity, which are inherently human traits. Over time, the performance of AI systems can degrade without proper maintenance, and there may be job losses as AI systems take over human tasks. Ethical concerns like privacy and potential bias, as well as the need for data quality and the quality of AI-generated outputs, are also significant challenges.

AI relies on large volumes of high-quality data for optimal performance. Limited access to relevant data, poor data quality, and existing biases can impact the performance and effectiveness of AI systems. A lack of interpretability, or the ability to understand how AI makes decisions, is another significant challenge that needs addressing. This lack of transparency can hinder trust, particularly in sensitive domains like healthcare or finance where explanations are crucial.

In addition to the technical challenges, organizations may struggle to find AI professionals with the necessary skills and knowledge. AI models may also be underrepresented to adversarial attacks where inputs are manipulated to deceive the model's predictions, requiring constant vigilance to ensure system security.

Addressing these challenges requires a focus on several areas. The data collection and preparation processes could be optimized to ensure clean, representative, and unbiased data. Improving interpretability and explainability, as well as ethical considerations, are also critical. This requires collaboration between academia, industry, and policymakers to share advancements in AI research, development, and best practices.

Other areas to improve include human-AI collaboration and trust, robustness and security, continuous learning and adaptability, and generalization and transfer learning. Adequate legal and regulatory frameworks need to be developed to govern the use of AI technology responsibly.

AI adoption also raises concerns about its societal impact and potential job displacement. Therefore, efforts should focus on reskilling and upskilling the workforce and promoting equitable distribution of benefits. At the same time, there are limitations related to data privacy, bias avoidance, and the need for AI models to provide explainable outputs.

In summary, while AI can offer many advantages, it also presents a variety of challenges that need to be addressed. These include the cost and time of implementation, lack of human emotion and creativity, degradation over time, potential job loss, and ethical issues. Overcoming these challenges requires improvements in data quality, interpretability, and ethical considerations, as well as collaboration between various sectors, maintaining security, and continuous learning and adaptability. As the field of AI continues to expand, addressing these challenges will be key to unlocking its full potential.

**Challenges and Limitations:**

- Data Quality and Availability: AI systems rely heavily on high-quality data. Incomplete, poor quality, or biased data can lead to inaccurate results and biased outcomes.
- Lack of Emotion and Creativity: AI lacks the human ability to use emotion and creativity in decision-making.
- Degradation over Time: AI systems can degrade over time if they are not properly maintained.
- Reduction in Human Jobs: Advanced AI systems may replace humans in certain tasks, leading to potential job losses.
- Ethical Issues: Ethical concerns surround AI, including privacy, bias, and potential societal impacts.
- Technical Expertise: Implementing and maintaining AI systems require specialized skills and resources, which are not always readily available.
- Interpretability and Explainability: Many AI systems, especially deep learning models, lack transparency in their decision-making processes.
- Vulnerability to Attacks: AI models can be susceptible to adversarial attacks that intentionally manipulate inputs to deceive the model.
- Legal and Regulatory Compliance: AI methods must comply with existing legal and ethical regulations, which can pose a challenge in their implementation.

**Areas for Improvement or Optimization:**

- Data Collection and Preparation: Improving data collection processes and ensuring data quality can greatly enhance the performance of AI systems.
- Interpretability and Explainability: Techniques to enhance the interpretability and explainability of AI models need to be developed and implemented.

- Ethical Frameworks and Guidelines: Developing ethical frameworks, guidelines, and regulations can ensure responsible AI development and use.
- Collaboration and Best Practice Sharing: Collaboration between academia, industry, and policymakers can facilitate the sharing of advancements in AI research and best practices.
- User-centric Design: Involving end-users in the development process can lead to more effective and user-friendly AI applications.
- Continual Learning and Adaptability: AI systems need to be continually updated and retrained to adapt to new data, changing environments, and evolving user needs.
- Robustness and Security: Developing robust AI models that can resist attacks and maintaining system security is an ongoing challenge.
- Scalability and Resource Efficiency: Exploring efficient AI algorithms and architectures that minimize resource requirements is crucial for scalability and sustainability.
- Legal and Regulatory Frameworks: Active contributions to policy discussions and compliance with legal and regulatory frameworks can ensure responsible AI use.
- Societal Impact and Workforce Preparedness: Efforts to understand and mitigate societal consequences of AI, and to prepare the workforce for AI, are essential.

### 5.3.3. Opportunities

Emerging trends in AI predominantly focus around security and privacy, aiming to tackle potential breaches as the utilization of AI methods expands across organizations, including public sectors such as the Ministry of Tourism. Techniques such as federated learning enable training of machine learning models across decentralized devices, ensuring privacy preservation while allowing for efficient learning from shared data. This has significant potential for areas dealing with sensitive data, such as healthcare.

Another critical trend is the deployment of Explainable AI (XAI) that adds transparency and interpretability to AI-driven outcomes. The capability of AI to explain the reasoning behind its decisions facilitates acceptance and trust among users and policy makers alike. This trend is essential for enhancing AI ethics and fairness, which in turn, fosters societal trust and responsible AI adoption. This emphasis on ethical practices also necessitates the development of methodologies to counter bias and discrimination in AI systems.

The advancements in Natural Language Processing (NLP), marked by techniques such as transformer models and pre-training methods, has led to sophisticated language understanding and generation, paving the way for services like chatbots, sentiment analysis, and machine translation. The combination of edge computing and AI has also led to real-time data processing, reducing latency, and improving privacy.

Artificial Intelligence finds applications across diverse domains, including predictive maintenance, design, emergency services, and security. AI's capability to process and analyse large, complex datasets significantly boosts efficiency across industries. It also has profound implications for healthcare, where AI can streamline workflows and enhance patient outcomes.

AI also supports continual learning and lifelong adaptation, addressing the limitations of traditional systems. Reinforcement Learning (RL) equips AI agents to learn from dynamic environments and adapt to evolving situations. Similarly, transfer learning allows models to apply knowledge from one task to another, augmenting their performance. AI's ability to summarize vast amounts of information and handle complex systems has made it an invaluable tool in various fields, including education and healthcare.

The development of AI should focus on establishing robust, scalable, and safe technologies. One significant aspect is the generation of interpretable models while maintaining high learning performance. This gives rise to "Interpretable AI," which can explain their operational principles, strengths, and weaknesses.

The interplay between AI and societal challenges presents opportunities for AI to address issues like climate change, poverty, and healthcare accessibility. A focus on AI research aiming to create a positive social impact can contribute to sustainable development. This interdisciplinary collaboration between AI and other fields offers a diverse perspective, bringing forth complex real-world solutions.

Finally, AI systems are expected to support, rather than replace, human decision-making. This calls for a balance between the strengths of AI and the unique capabilities of human judgment. With the growing awareness of eliminating biases in AI systems, the opportunity to create more inclusive and equitable systems presents itself. AI's role in privacy protection and resilience to cyberattacks also offers prospects for robust system development. This implies the emergence of new roles and careers in AI governance, ethics, and regulation. Consequently, AI's transparent and innovative nature fosters collaboration and alignment with human values.

**Benefits of Using AI in an Organization:**

- Improved Decision Making: AI systems can analyse vast amounts of data and provide actionable insights, enhancing the decision-making process.
- Increased Efficiency: AI can automate routine tasks, allowing staff to focus on more complex and creative duties, thereby boosting overall productivity.
- Enhanced Customer Experience: Through AI-powered chatbots and personalized recommendations, organizations can provide a more tailored and responsive customer service.

- Reduced Operational Costs: By automating tasks and improving efficiency, AI can significantly reduce operational costs.
- Risk Management: AI can predict and mitigate potential risks by analysing patterns in historical data, proving vital for sectors like finance and healthcare.
- Innovation: AI facilitates innovation by identifying trends and insights that might be overlooked by humans.
- Transparency and Trust: Explainable AI (XAI) makes the decision-making process of AI systems transparent, enabling users to understand and trust AI outcomes. This can foster trust and acceptance among stakeholders and regulators.

**Advantages to Operations and Services:**

- 24/7 Availability: AI-powered services such as chatbots can provide round-the-clock support, significantly enhancing customer service.
- Data Analysis: AI's ability to analyse large datasets can help organizations draw insights to improve their services, products, and overall business strategy.
- Accuracy and Consistency: AI systems can perform tasks with high accuracy and consistency, minimizing errors that might occur with human involvement.
- Predictive Capabilities: AI can predict customer behaviour and market trends, helping organizations tailor their offerings and stay competitive.
- Personalization: AI can analyse customer data to deliver personalized experiences, leading to higher customer satisfaction and loyalty.
- Scalability: AI systems can handle an increasing amount of work or its potential to be enlarged to accommodate growth, helping businesses scale their operations effectively.
- Security Enhancement: AI can strengthen cybersecurity efforts, detecting and preventing breaches before they occur.
- Transparency in Operations: Explainable AI can make the operations of an AI system transparent, enhancing understanding of its functioning and improving accountability and regulatory compliance.

### 5.3.4. Threats

The benefits of Artificial Intelligence (AI) come with certain risks and threats that need to be thoroughly understood and mitigated. One of the key issues is bias and discrimination. AI systems can inadvertently perpetuate the biases present in their training data, leading to unfair or discriminatory results. To ensure fairness, it's crucial to carefully curate training datasets, conduct bias audits, and employ fairness-aware algorithms. Transparency and openness can be achieved by providing interactive visualizations to policymakers and citizens.

Another significant issue revolves around privacy and data security. AI processes vast amounts of data, sometimes personal or sensitive in nature. Therefore, institutions need to establish robust data privacy protocols and stringent cybersecurity measures to prevent unauthorized access and data breaches. Sensitive data should be protected through methods like data anonymization, encryption, and secure storage. Adherence to privacy regulations and obtaining explicit consent for data usage are also key aspects to consider.

AI systems, particularly complex deep learning architectures, can also lead to unintended consequences and unpredictability. Thorough testing, validation, and risk assessment, as well as rigorous quality assurance processes, are essential to identify and handle potential issues before deployment.

The opacity of some AI models can cause a lack of trust. To combat this, promoting transparency and explainability is critical. Developing interpretable AI techniques and providing understandable explanations for AI decisions can enhance transparency, enable users to comprehend the reasoning behind AI outcomes, and build trust.

Job displacement and impact on the workforce is another concern. The potential for AI to automate tasks can lead to job displacement and changes in the workforce. Therefore, it's essential to develop strategies for reskilling and upskilling employees and to consider the socio-economic impacts of AI adoption.

AI systems can also be susceptible to adversarial attacks, which involve the intentional manipulation of input data to deceive the AI system. Therefore, robustness against such attacks should be a priority during AI development. Employing adversarial training, input validation techniques, and model robustness assessments can help bolster the resilience of AI systems.

Ethical considerations and accountability are vital in the context of AI. AI raises complex ethical questions, such as the potential for autonomous weapons, privacy infringements, and decision-making transparency. Adherence to ethical guidelines, advocacy for responsible AI practices, and implementation of accountability and auditability mechanisms are crucial in this regard.

AI can exacerbate existing social disparities if not designed with equity in mind. Therefore, it's crucial to promote diversity in AI research and development, ensure fair representation in training datasets, and consider the broader societal impact of AI systems. Engaging with affected communities, conducting impact assessments, and actively working towards reducing bias and inequalities can help address these concerns.

Adherence to evolving legal and regulatory requirements, including data protection, privacy, and safety regulations, is necessary. Research institutions should actively monitor and comply with these requirements. Collaborating with policymakers, industry stakeholders, and legal experts to contribute to the development of robust AI governance frameworks is essential.

Lastly, AI systems need continuous monitoring and iterative improvement. Establishing mechanisms for feedback collection, user engagement, and ongoing evaluation is crucial. Regular audits, algorithmic impact assessments, and version control can help ensure that AI systems remain effective, unbiased, and aligned with stakeholder needs.

In summary, managing AI responsibly and ethically requires a multidisciplinary approach, a commitment to ethical considerations, robust governance frameworks, and active participation in policy discussions.

**Potential Risks or Threats when Using AI:**

- Bias and Discrimination: AI can perpetuate biases present in the training data, leading to unfair results.
- Privacy and Data Security: AI processes large amounts of data, which can include personal or sensitive information.
- Unintended Consequences and Unpredictability: Complex AI systems can sometimes lead to unpredictable outcomes.
- Lack of Transparency and Explainability: The reasoning behind AI decisions can be opaque, leading to a lack of trust.
- Job Displacement and Workforce Impact: AI can automate tasks, potentially leading to job displacement.
- Susceptibility to Adversarial Attacks: AI systems can be deceived by intentionally manipulated input data.
- Ethical Considerations and Accountability: AI can raise complex ethical questions, such as potential for autonomous weapons, privacy infringements, and decision-making transparency.
- Social Impact and Equity: AI can exacerbate social disparities if not designed with equity in mind.
- Regulatory Compliance: AI technologies need to adhere to evolving legal and regulatory requirements.
- Continuous Monitoring and Iterative Improvement: AI systems need ongoing evaluation and adjustment to remain effective and unbiased.

**How to Minimize These Risks and Use AI Responsibly and Ethically:**

- Address Bias: Carefully curate training datasets, conduct bias audits, and use fairness-aware algorithms.
- Ensure Privacy and Data Security: Establish robust data privacy protocols, implement stringent cybersecurity measures, and adhere to privacy regulations.
- Test for Unintended Consequences: Conduct thorough testing, validation, and risk assessment, and implement rigorous quality assurance processes.
- Enhance Transparency and Explainability: Develop interpretable AI techniques and provide understandable explanations for AI decisions.
- Consider Workforce Impact: Develop strategies for reskilling and upskilling employees and consider the socio-economic impacts of AI adoption.
- Ensure Robustness Against Adversarial Attacks: Employ adversarial training, input validation techniques, and model robustness assessments.
- Uphold Ethical Considerations and Accountability: Adhere to ethical guidelines, advocate for responsible AI practices, and implement mechanisms for accountability and auditability.
- Consider Social Impact and Equity: Promote diversity in AI research, ensure fair representation in training datasets, and consider the broader societal impact of AI systems.
- Comply with Regulatory Requirements: Actively monitor and comply with legal and regulatory requirements and contribute to the development of robust AI governance frameworks.
- Implement Continuous Monitoring and Iterative Improvement: Establish feedback collection, user engagement, and ongoing evaluation mechanisms, conduct regular audits, and maintain version control.

# 6. Conclusions

Discrimination and bias in societal institutions have far-reaching impacts on marginalized groups, perpetuating social and economic inequalities. Interventions to address these issues must be multi-faceted and tailored to specific contexts. Intergroup contact, careful AI development, and ongoing monitoring and evaluation are all important components of effective interventions. By addressing discrimination and bias in societal institutions, a more just and equitable society can be developed for all. This deliverable report consists only the introduction and approach to be followed for and to the activities of WP2. The upcoming iteration on project month 18 will provide further information along with the results of the activities foreseen under WP2 (T2.1 and T2.2) as well as of those described in the very current document, providing the project's Holistic Regulatory Framework which will ensure that the proposed framework protects citizens from potential abuse enabled by the use of Big Data and AI. The HRF will be in-line with applicable laws, protocols, and regulations (i.e., the GDPR), but also with ethical recommendations for AI (e.g., the recommendations of the HLEG). This framework will be integrated into different architecture blueprints acquiring/ensuring a holistic view on intersectional bias and ethics. A final comprehensive analysis of multiple types of bias (built upon the current ongoing work and) based on different grounds such as age, disability, gender, race, sexual orientation, and gender identity will be provided to establish the grounds for the realisation of the project's outcomes providing support for "regulatory compliance by design".

# 7. References

- AI4Gov. (n.d.). About Us. Retrieved from https://ai4gov-project.eu/
- Alexander, M. (2010). The new Jim Crow: Mass incarceration in the age of colorblindness. The New Press.
- Banaji, M. R., & Greenwald, A. G. (2013). Blindspot: Hidden biases of good people. Delacorte Press.
- Baron, J. (2000). Thinking and deciding. Cambridge University Press.
- Capturing value from Artificial Intelligence | Academy of Management ... Available at: https://journals.aom.org/doi/abs/10.5465/amd.2023.0106 (Accessed: 26 June 2023).
- Bielby, W. T., & Baron, J. N. (2003). Men and women at work: Sex segregation and statistical discrimination. In F. D. Blau, M. C. Brinton, & D. B. Grusky (Eds.), The declining significance of gender? (pp. 233-262). Russell Sage Foundation.
- Cook, C., & Spray, C. (2012). Ecosystem services and integrated water resource management: Different paths to the same end?. Journal of environmental management, 109, 93-100.
- Dalkey, N. C., & Helmer, O. (1963). An Experimental Application of the Delphi Method to the Use of Experts. Management Science, 9(3), 458-467. doi: 10.1287/mnsc.9.3.458
- Dovidio, J. F., Gaertner, S. L., & Kawakami, K. (2003). Intergroup contact: The past, present, and the future. Group Processes & Intergroup Relations, 6(1), 5-21.
- Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping and discrimination: theoretical and empirical overview. The SAGE handbook of prejudice, stereotyping and discrimination, 3-28.
- Georgina (2022) Shadow reports on racism in Europe, European Network Against Racism. Available at: https://www.enar-eu.org/Shadow-Reports-on-racism-in-Europe-203/ (Accessed: 26 June 2023).
- Gilovich, T., Griffin, D., & Kahneman, D. (2002). Heuristics and biases: The psychology of intuitive judgment. Cambridge University Press.
- Office, U.S.G.A. (no date) Federal workforce: Strengthening diversity, equity, inclusion, and accessibility, Federal Workforce: Strengthening Diversity, Equity, Inclusion, and Accessibility | U.S. GAO. Available at: https://www.gao.gov/products/gao-23-106254 (Accessed: 26 June 2023).
- Greenwald, A. G., & Banaji, M. R. (2017). Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychology Press.
- Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. California Law Review, 945-967.

- Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-16).
- Jost, J. T., & Kay, A. C. (2010). Social justice: History, theory, and research. Handbook of social psychology, 2, 1122-1165.
- Kahneman, D. (2011). Thinking, fast and slow. Macmillan.
- National Center for Education Statistics. (2018). Status and trends in the education of racial and ethnic groups 2018. US Department of Education.
- National Conference on State Legislatures. (2020). Women and minorities in state legislatures. Retrieved from https://www.ncsl.org/womens-legislative-network/women-in-state-legislatures-for-2020
- National Fair Housing Alliance. (2019). Fair housing laws. Retrieved from https://nationalfairhousing.org/fair-housing-laws/
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175-220.
- Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. Annual Review of Sociology, 34, 181-209.
- Pager, D., Western, B., & Bonikowski, B. (2009). Discrimination in a low-wage labor market: A field experiment. American Sociological Review, 74(5), 777-799.
- Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. Annual review of psychology, 60, 339-367.
- Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. Journal of personality and social psychology, 90(5), 751-783.
- Russell, S. T., & McGuire, J. K. (2008). The experiences of lesbian, gay, bisexual and transgender students in rural and small town schools. Journal of LGBT Youth, 5(1-2), 15-36.
- Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the implicit association test. Group Processes & Intergroup Relations, 10(3), 359-372.
- Smedley, B. D., Stith, A. Y., & Nelson, A. R. (2003). Unequal treatment: Confronting racial and ethnic disparities in health care. National Academies Press.
- Steele, C. M. (2010). Whistling Vivaldi: And other clues to how stereotypes affect us (issues of our time). WW Norton & Company.
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5(2), 207-232.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131.

- UNDP (2021). Qualitative assessment of digital access and skills of underrepresented groups. Retrieved from: https://www.undp.org/sites/g/files/zskgke326/files/migration/mn/UNDP-A-Lab-Report-Eng-20211004.pdf
- United Nations. (2020). Disability inclusion. Retrieved from https://www.un.org/development/desa/disabilities/
- US Department of Housing and Urban Development. (2019). Fair housing: It's your right. Retrieved from https://www.hud.gov/program_offices/fair_housing_equal_opp/fair_housing_act_overview
- Williams, D. R. (2012). Miles to go before we sleep: Racial inequities in health. Journal of Health and Social Behavior, 53(3), 279-295.
- Zamfir, A., & Corbos, R. A. (2015). Towards sustainable tourism development in urban areas: Case study on Bucharest as tourist destination. Sustainability, 7(9), 12709-12722.
- Greenwald, A., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. Journal of Personality and Social Psychology, 74(6), pp. 1464-1480.
- Hamilton, D., & Gifford, R. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. Journal of Experimental Social Psychology, 12(4), pp. 392-407.
- Jost, J., & Banaji, M. (1994). The Role of Stereotyping in System-Justification and the Production of False Consciousness. British Journal of Social Psychology, 33, pp. 1-27.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. (T. Gilovich, D. Griffin, & D. Kahneman, Eds.) Heuristics and biases: The psychology of intuitive judgment, pp. 49-81.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4), pp. 237–251. doi:https://doi.org/10.1037/h0034747
- Markus, H., & Kitayama, S. (2010). Cultures and selves: A cycle of mutual constitution. Perspectives on Psychological Science, 5(4), pp. 420-430.
- Nieto, S. (2000). Affirming diversity: The sociopolitical context of multicultural education. Allyn & Bacon.
- Nosek,, B., Greenwald, A., & Banaji, M. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. ersonality and Social Psychology Bulletin, 31(2), pp. 166-180.
- Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. Annual Review of Sociology, 34, pp. 181-209.
- Pager, D., Western, B., & Bonikowski, B. (2009). Discrimination in a low-wage labor market: A field experiment. American Sociological Review, 74(5), pp. 777-799.
- Schug, J., Alt, N., & Klauer, K. (2016). The psychological bases of discrimination and exclusion: A meta-analytic review. Psychological Bulletin, 142(4), pp. 380-421.

- Smedley, B., Stith, A., & Nelson, A. (2003). Unequal treatment: Confronting racial and ethnic disparities in health care. National Academies Press.
- Sue, D., Rasheed, M., & Rasheed, J. (2016). Multicultural social work practice: A competency-based approach to diversity and social justice. John Wiley & Sons.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), pp. 1124-1131.
- Ragnedda, M., & Ruiu, M. (2019). Digital exclusion: An introduction. Routledge.
- Van Dijk, J. A. (Ed.). (2013). The digital divide (Vol. 14). Routledge.
- Selwyn, N. (2004). Reconsidering political and popular understandings of the digital divide. New media & society, 6(3), 341-362.
- Kim, M. A. (2017). The impact of Internet use on social isolation among older adults: An empirical analysis using panel data. Cyberpsychology, Behavior, and Social Networking, 20(7), 411-416.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

# ANNEX 1. Ai4GOV Vocabulary

1. Accountability - The responsibility of individuals or organizations to answer for their actions and decisions.
2. Algorithm - A set of instructions that a machine can follow to perform a specific task, such as identifying patterns in data or making decisions based on data.
3. Artificial Intelligence (AI) - A field of computer science that focuses on creating intelligent machines that can perform tasks that typically require human intelligence.
4. Bias - A tendency to favour one group or individual over another, often based on preconceived notions or stereotypes.
5. Blockchain-based Information Exchange (BIE) solution - A tool to regulate data access and facilitate secure data exchange using blockchain technology.
6. Data - Information that is collected and analysed to inform decision-making processes.
7. Data Governance Framework (DGF) - A framework to govern data flows, providing technical solutions regulated by organizational rules or legal recommendations.
8. Discrimination - The unjust or prejudicial treatment of individuals or groups based on factors such as race, gender, age, religion, or disability.
9. Equality - The state of being equal, or having the same rights, opportunities, and treatment as others.
10. Inclusivity - The quality of being inclusive, or including individuals from diverse backgrounds and perspectives.
11. Machine Learning (ML) - A subset of AI that involves training machines to learn from data and make predictions or decisions based on that data.
12. Multilingual Sentiment Analysis/Topic Modelling - Techniques used to analyse and evaluate citizens' feedback on policies, taking into account multiple languages and sentiments.
13. Self-Explained Visualisation Workbench - A tool that enables a holistic understanding of data and policies through advanced visual analytics techniques.
14. Situation-Aware Explainability (SAX) - A method to foster citizen participation and provide enhanced data provisioning by offering explanations for multi-domain policy adaptation.
15. Social Exclusion: This term refers to the processes of "disqualification from social relations" (Sen, 2000) wherein certain groups are systematically marginalized from social, economic, and political systems of a society due to factors such as socio-economic status, race, gender, age, or disability. Sen's analysis of social exclusion focuses on the inability of individuals to participate fully in the cultural, economic, political, and social life of their society.
16. Virtualized Unbiasing Framework (VUF) - A framework that uses AI-based analytics to extract policies from datasets, mitigate bias, and preserve data privacy.
17. XAI (eXplainable AI) - A library of tools and techniques used to explain the rationale behind AI systems' decisions, enhancing transparency and trust.

18. Bias Detector Toolkit - A toolkit to identify and mitigate potential discrimination cases based on various factors such as age, gender, or underrepresented groups.
19. Interactive Visualization Workbench - A tool that facilitates transparency and openness by visualizing policies and their adaptations based on emerging situations.