AutoFair
(Horizon Europe Grant agreement ID: 101070568)

HUMANCOMPATIBLE.ORG
HUMAN-COMPATIBLE AI WITH GUARANTEES

D3.1 A study on Fair AI policies and regimes (EU and selected third countries)

ARC

June 6, 2024

Status: Live Document

Scheduled Delivery Date: 30/09/2023

Document History

- (November 30th, 2023) Version 1.0 Submitted to the EC and uploaded to AutoFair website.

Executive summary

Artificial intelligence powers a growing number of systems today. This is why it's important to make sure the AI algorithms work correctly. In this context, the EU-funded AutoFair project will focus on the issue of fairness. Specifically, it will deal with the design of explainable and transparent AI algorithms. AutoFair aims to improve the algorithms themselves while educating end users. It draws on knowledge from computer and data sciences, control theory, optimisation and other scientific disciplines, including ethics and law. Three case studies will be carried out to test the findings on the automation of fair evaluation in recruitment, the emergence and mitigation of bias amplification in online advertising and the elimination of discrimination against bank clients.

D3.1 reports on existing AI and fairness policies and presents suggestions and best practices for developing fair AI systems. First, an extensive review of the AI legislation worldwide is presented, along with a focused presentation of current EU and US law on discrimination. Then, an overview of algorithmic fairness definitions, methods and assessment criteria is provided. These two first sections try to provide a thorough context on the main concepts, issues and challenges on pursuing AI fairness, taking into account both the legal and the algorithmic viewpoint. Next, a series of EU law, precedents and use cases are presented, which emphasize some key challenges to be handled on the intersection of law and algorithmic, followed by an extensive review of the policies, tools and framework for ethical AI. Finally, a set of suggestions are providing comprising (i) recommendations identified within prominent studies of the literature and (ii) a proposed template for the development of AI fairness policies.

Please check the website of the project (`https://humancompatible.org/`) or CORDIS (`https://cordis.europa.eu/project/id/101070568`) under the deliverables section for additional deliverables and updates.

Document Information

Contract Number	101070568	**Acronym**	AutoFair
Name	Human-Compatible Artificial Intelligence with Guarantees		
Project URL	https://humancompatible.org/		
EU Project Officer	David Zunel-Ballester		

Deliverable	**D3.1**	A study on Fair AI policies and regimes (EU and selected third countries)	
Work Package	**3**	Detection and Quantification of Multiple Types of Bias	
Date of Delivery	30/09/2023	**Actual**	30/11/2023
Status	Live Document		
Nature	Report		
Distribution Type	Public		
Authoring Partner	ARC		
QA Partner	CTU		
Contact Person	Giorgos Giannopoulos	giann@athenarc.gr	
	Loukas Kavouras	kavouras@athenarc.gr	
	Maria Psalla	mpsalla@yahoo.gr	
	Prodromos Tsiavos	ptsiavos@athenarc.gr	

List of Contributors: German Martinez Matilla (CTU), Giorgos Giannopoulos (ARC), Loukas Kavouras (ARC), Jakub Marecek (CTU), Maria Psalla (ARC), Eleni Psaroudaki (ARC), Prodromos Tsiavos (ARC)

Project Information

This document is part of a research project funded by Horizon EUrope programme of the Commission of the European Communities under Grant agreement ID 101070568. The Beneficiaries in this project are:

No.	Name	Short Name	Country
1	Czech Technical University in Prague	CTU	Czech Republic
2	Athena Research Center	ARC	Greece
3	IBM Ireland Limited	IBM	Ireland
4	Technion – Israel Institute of Technology	Technion	Israel
5	Workable Software Single Member Private Company	WOR	Greece
6	National and Kapodistrian University of Athens	UoA	Greece
7	Dateio s.r.o.	DAT	Czech Republic

Table of Contents

List of Terms and Abbreviations

Abbreviation	**Definition**
AI	Artificial Intelligence
EU	European Union
GDPR	General Data Protection Regulation
ML	Machine Learning
US	United States

1 Introduction

In this deliverable, we perform a review of the relevant literature and regulation on discrimination and AI. We examine existing policies, as well as we briefly present a set of prominent fairness definitions and methods to detect and quantify algorithmic bias, drafting a set of criteria for the assessment of different bias detection methods. We identify existing gaps and challenges on the intersection of law and algorithms and discuss possible routes for resolution Finally, we present suggestions on best practices towards implementing AI fairness.

In our work, we identify the considerable gap between the legal and the algorithmic viewpoint of discrimination and fairness, as also done by prominent works on the intersection of the two disciplines [12]. Consequently, we make an effort to bridge this gap, by incorporating literature material from both disciplines and presenting a discussion on their intersection.

To this end, the deliverable is organized as follows. Section 2 serves as an introduction to the main concepts and notions discussed in the deliverable. This includes a brief introduction to the legal framework on discrimination in the European Union (EU) and the United States (US), an introduction to the concept of fairness in AI and the AI Act, a discussion on the Ethics Guidelines for Trustworthy AI[1] and an enumeration of the current AI legislation in a series of countries worldwide.

Section 3 proceeds in presenting a series of prominent algorithmic fairness definitions along with a series of methods for bias detection. Additionally, it enumerates a set of criteria to use for the assessment of fairness definition and measures, which are drawn both from the literature and from the writers' experience on the field of fairness and explainability in AI. A particular effort is put in providing simplified descriptions for the presented fairness definitions, so as to popularize them to non-technical/algorithmic audience.

Section 4 reviews existing policies, frameworks and tools on ethical AI, and brings out some exemplary cases regarding either decisions the European Court of Justice, or examples of real world application of AI where fairness/discrimination was on the epicenter.

Section 5 first summarizes a set of suggested policies and best practices towards achieving AI fairness drawn from the recent relevant literature. We focus on three studies, which more or less summarize the findings or a broader set of manuscripts, discussion existing gaps and challenges, but also means for resolution and best practices towards bridging the gap between law and algorithms and implementing fair by design AI systems. Additionally, we propose a general template for developing AI fairness policies, which takes into account knowledge and insights reported in the previous sections. Section 6 concludes the document.

[1] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

2 The Legal framework on discrimination and fairness in AI

In this section, we focus on the legal and ethical aspects of discrimination and fairness. First, we provide an overview of the relevant legislation on discrimination examining EU and US law. Then, we discuss the ethical aspects of fairness in AI, followed by a brief presentation of the AI Act and the Ethics Guidelines for Trustworthy AI. Finally, we broaden the scope and examine international initiatives to regulate ethical AI.

2.1 Discrimination in the EU law

The aim of non-discrimination law is to allow all individuals an equal and fair prospect to access opportunities available in a society [28]. In the EU, the legal framework against discrimination has been shaped by both the laws of the Council of Europe and EU legislation. This framework is enshrined in both primary and secondary legal instruments.

2.1.1 The law of the Council of Europe

The Council of Europe is an intergovernmental organization with the aim of, among other things, promoting human rights and social equality. The European Convention for the Protection of Human Rights and Fundamental Freedoms, also known as the European Convention on Human Rights (ECHR) [29], which was adopted by the member states of the Council of Europe in 1950, enshrines in Article 14 the prohibition of discrimination, stating specifically: "*The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status*". With the Twelfth Protocol of the ECHR, signed in 2000, member states "having regard to the fundamental principle according to which all persons are equal before the law and are entitled to the equal protection of the law" and considering that they should "*take further steps to promote the equality of all persons through the collective enforcement of a general prohibition of discrimination*", adopted a more recent provision, introducing a general prohibition of discrimination. This provision specifically states that "*the enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status*".

The European Social Charter (ESC) (revised) of the Council of Europe[2] is another fundamental treaty of the Council of Europe for human rights. In Part V, Article E specifically refers to the non-discrimination clause during the enjoyment of the rights of the Charter. It states: "*The enjoyment of the rights set forth in this Charter shall be secured without discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national extraction or social origin, health, association with a national minority, birth or other status*".

In addition to the above, there are many other Acts and specific Conventions that contain and have, as their fundamental principle, the prohibition of discrimination, which underpins all

[2] https://www.coe.int/en/web/european-social-charter

legislation of the Council of Europe.

2.1.2 The law of the European Union

In the initial Treaties of the European Community, there is no mention of fundamental rights or their protection. The early regulations against discrimination were limited to a provision that prohibited gender discrimination in employment, and subsequently, other areas were regulated, such as pensions, pregnancy, and mandatory social security systems. In 2000, the EU and its member states, recognizing that their policies could impact human rights, made the proclamation of the Charter of Fundamental Rights of the European Union[3]. Article 21 of the Charter addresses the prohibition of discrimination, stating: "*Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited*". Additionally, Article 20 guarantees equality before the law, Article 22 declares respect for cultural, religious, and linguistic diversity, and Article 23 ensures gender equality.

In the Treaty on European Union[4], which is based on the Maastricht Treaty [which was subsequently amended with the following treaties: Amsterdam Treaty (1997), Nice Treaty (2001), Lisbon Treaty (2007)], provisions regarding the prohibition of discrimination exist. Specifically, in Articles 2 and 3, it is mentioned that the Union "*is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail*" and that "*shall combat social exclusion and discrimination, and shall promote social justice and protection, equality between women and men, solidarity between generations and protection of the rights of the child*".

Beyond the aforementioned EU constitutional texts, there are four Directives that address non-discrimination law and provide further specific regulations:

- The Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.
- The Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation.
- The Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services.
- The Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation.

[3] https://www.europarl.europa.eu/charter/pdf/text_en.pdf
[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12012M%2FTXT

2.1.3 Discrimination categories

Non-discrimination law aims to allow all individuals an equal and fair prospect to access opportunities available in a society [28]. European non-discrimination law addresses two general types of discrimination: direct and indirect discrimination. Direct discrimination means that a person is treated less favorably based on a protected attribute (e.g. race and ethnicity, gender, religion and belief, age, disability, or sexual orientation) that they possess in matters of a protected sector (e.g. the workplace, provision of goods and services). Different groups receive different levels of protection. Direct discrimination is grounded in the Aristotelian postulate of treating 'like cases alike' and treating 'different cases differently' unless there is an objective reason not to do so. Equality achieved on these terms is also called "formal equality," or the "merit principle" [42].

Indirect discrimination occurs when ostensibly neutral provisions or practices, universally applied, disproportionately disadvantage individuals with specific protected characteristics, such as religion, disability, age or sexual orientation. This form of discrimination is evident when ostensibly fair rules unrelated to protected attributes affect a particular group disproportionately. The concept acknowledges that addressing existing inequalities may require differential treatment among groups, as justified by the principle of justified indirect discrimination. In such cases, pursuing a legitimate aim is acceptable if the mechanisms pass the "proportionality test", ensuring legal necessity and proportionate nature. Distinguishing itself from direct discrimination, this approach recognizes and addresses social challenges faced by protected groups, emphasizing the need for tailored strategies. Importantly, the specter of indirect discrimination is pronounced in Artificial Intelligence, machine learning, and automated decision-making, as these systems often rely on inferences that may inadvertently perpetuate biases and disparities among different demographic groups.

2.2 Discrimination in the US law

In this subsection, we discuss the antidiscrimination law of the United States. We begin with a brief description of the legal system of the United States, and then delve into the specifics of antidiscrimination laws.

2.2.1 The Legal Framework of the United States: Navigating American Jurisprudence

The American legal system is a dynamic framework designed for justice and fairness. Operating under the U.S. Constitution, it consists of federal and state components, each with three branches—legislative, executive, and judicial—to ensure checks and balances. Led by the Supreme Court, the judiciary interprets laws and relies on common law tradition and precedent. Committed to upholding justice, protecting individual rights, and maintaining order, the system navigates complexity through constitutional principles, statutes, and established judicial precedents.

The Congress (legislature)

Congress, one of the three coequal branches of government, holds significant powers according to the Constitution, being the sole entity authorized to create new laws or modify existing ones. These laws, termed statutory laws, play a crucial role in the legal landscape. Given the evolving nature of society, these laws are often broadly formulated, allowing flexibility in interpretation. Congress delegates authority to federal agencies for implementation, with the courts providing interpretive functions and acting as a check on Congress's power.

Congress has utilized its authority to enact antidiscrimination statutes covering various activities, but there remain gaps and limitations in federal antidiscrimination law. Constitutional constraints and congressional inaction contribute to these shortcomings. Consequently, state laws sometimes step in to address these gaps. The interplay between Congress, federal agencies, and the courts underscores the dynamic nature of lawmaking and enforcement in the United States.

The Courts (judiciary)

In the United States, the court system plays a pivotal role within the common-law framework, allowing courts to establish legal precedents that guide future decisions. This precedent, formed through past cases, obligates judges to follow the reasoning applied in similar situations. Courts are not only responsible for interpreting statutory laws and the Constitution but also for shaping a body of case law, which holds comparable authority to other legal sources. This stands in contrast to the civil law system prevalent in most of Europe, where legislation takes precedence, and judicial decisions carry less weight as precedent.

Federal courts possess the exclusive authority to interpret, assess the constitutionality of laws, and apply them to individual cases. Similar to Congress, courts can compel the production of evidence and testimony using subpoenas, underscoring their crucial role in the legal system.

Federal Agencies (executive governance)

Federal agencies, specialized government bodies, are created through legislative or presidential action for purposes like resource management and national security. Directed by presidential appointees, they regulate industries requiring oversight and expertise. A key role is anti-discrimination through rulemaking, guidance, and law enforcement. Rulemaking involves drafting regulations, forming administrative laws alongside statutory and case law.

While court-referenced, guidelines are non-binding. Federal agencies are pivotal for statute effectiveness. Varying in political independence, some act within the executive, while others, with enforcement powers, are quasi-judicial. Collaboration with courts can face inefficiencies due to diverse legal sources and enforcement methods.

2.2.2 Anti-discrimination Laws

In this chapter, we reference anti-discrimination laws that may safeguard various rights and groups based on sex, age, race, disability, sexual orientation, gender, sex characteristics, religion and other characteristics. Specifically:

1. Title VII of the Civil Rights Act of 1964, which prohibits employment discrimination based on race, color, religion, national origin, or sex. It also forbids retaliation against those who report discrimination. The law mandates that employers reasonably accommodate sincerely held religious practices unless it imposes an undue hardship. Title VII addresses disparate treatment and disparate impact, aiming to ensure equal employment opportunities and eliminate workplace discrimination. Employment discrimination cases usually fall into two primary categories: disparate treatment and disparate impact. We will provide a brief discussion of these two cases in the next chapter.

2. The Equal Protection Clause of the Constitution, found in the 14th Amendment to the U.S. Constitution, prohibits states from making or enforcing laws that deprive citizens of their privileges or immunities, or deny any person within its jurisdiction the equal protection of the laws. Ratified in 1868 after the Civil War to prevent racial discrimination, the clause has evolved to cover a broader range of protections beyond its original intent. Its primary purpose is to ensure fairness and prevent unjust treatment under the law.

3. The Equal Credit Opportunity Act (ECOA) is a federal law preventing discrimination in credit transactions. It applies to all credit extensions, including those to businesses, and prohibits discrimination based on factors like race, religion, sex, and more. ECOA ensures that lenders assess loan applicants solely on their ability to repay, protecting consumers from unfair credit discrimination.

4. Title VIII of the Civil Rights Act of 1968, the Fair Housing Act, prohibits discrimination in housing based on race, color, religion, sex, familial status, national origin, and disability. It applies to the sale, rental, financing, and advertising of dwellings, as well as other housing-related transactions. The Act also mandates the affirmative promotion of fair housing in federal housing and urban development programs.

5. Title VI of the Civil Rights Act of 1964 prohibits discrimination based on race, color, or national origin in any program or activity receiving federal financial assistance. It ensures that no person in the United States is excluded from participation, denied benefits, or subjected to discrimination under such programs. Title VI specifically applies to programs or activities receiving federal financial assistance from the Department of Housing and Urban Development (HUD).

6. The Pregnancy Discrimination Act of 1978 ("PDA"), an amendment to Title VII, prohibits discrimination based on pregnancy, childbirth, or related medical conditions. It also protects individuals from retaliation for reporting discrimination. The Act applies to employers with 15 or more employees and mandates disability and sick leave for women recovering from abortions, excluding elective abortions unless the mother's life is at risk.

7. The Equal Pay Act of 1963 (EPA), an amendment to the Fair Labor Standards Act, prohibits sex-based wage discrimination between men and women performing equal work in the same establishment. Administered by the EEOC, the law aims to eliminate gender pay disparities and prohibits retaliation against those reporting discrimination. It ensures equal pay for substantially equal work, considering factors like skill and responsibility under similar working conditions. The Act was enacted to address the adverse effects of sex discrimination on wages, labor resources, and commerce.

8. The Age Discrimination in Employment Act of 1967 (ADEA) is a U.S. labor law prohibiting employment discrimination against individuals aged 40 or older, ensuring equal employment opportunities. ADEA covers age discrimination, pensions, and benefits standards, requiring employers to share information about older workers' needs publicly. The Older Workers Benefit Protection Act amends ADEA, setting conditions for waiving its protections.

9. Title I of the Americans with Disabilities Act of 1990 (ADA) prohibits discrimination against qualified individuals with disabilities in the private sector and government. It also prevents retaliation against those reporting discrimination. Employers must reasonably accommodate known disabilities unless it causes undue hardship. Enacted in 1990, the ADA ensures equal opportunities for individuals with disabilities in various aspects of public life, such as employment and public spaces, mirroring protections for race, color, sex, national origin, age, and religion under five distinct titles.

10. Sections 102 and 103 of the Civil Rights Act of 1991 amend Title VII and the Americans with Disabilities Act (ADA) to allow jury trials and awards for compensatory and punitive damages in cases of intentional discrimination.

11. Sections 501 and 505 of the Rehabilitation Act of 1973 prohibit disability discrimination in the federal government. They require federal agencies to reasonably accommodate qualified employees or applicants with disabilities and make it illegal to retaliate against those reporting discrimination or participating in related actions. Employers must provide accommodations unless it imposes undue hardship on their business.

12. The Genetic Information Nondiscrimination Act of 2008 (GINA) is U.S. federal legislation protecting individuals from discrimination based on their genetic information in health insurance and employment. Enacted in 2008, GINA makes it illegal to discriminate against employees or applicants because of their genetic information, encompassing genetic tests and family medical history. The law also prohibits retaliation against those who report discrimination or participate in related investigations or lawsuits.

13. The Pregnant Workers Fairness Act of 2022 (PWFA) mandates covered entities to provide reasonable accommodations to qualified workers with known limitations related

to pregnancy, childbirth, or related medical conditions, unless it causes undue hardship. The law also prohibits retaliation against those who report discrimination, file charges, participate in employment discrimination proceedings, or reasonably oppose discrimination. The PWFA aims to ensure fair treatment and accommodations for employees or applicants affected by pregnancy-related conditions.

14. The Immigration and Nationality Act of 1965 ("INA") established a preference system prioritizing relatives, skilled professionals, and refugees. It marked a shift by setting numerical limits on immigration for the first time, particularly from the Western Hemisphere. Abolishing quotas, the law aimed to attract those contributing significantly to the country's growth. The preference system focused on immigrants' skills and family ties with U.S. citizens or residents, leading to increased immigration from Asia, Africa, and Eastern/Southern Europe.

As we observe, the management of discrimination and prejudice is anchored in a sector-specific approach—with distinct characteristics.

The initial observation underscores that selecting legislative safeguards for a specific and targeted right or group facing discrimination leads to enhanced and more comprehensive protection for that particular right or group. It becomes apparent that a broad legislative framework striving to shield all groups from discriminatory behaviors or uphold all rights may lack clarity and fail to incorporate the necessary safeguards for effective implementation of anti-discrimination measures.

Delving into anti-discrimination laws yields a second crucial insight: "*even though laws are sector-specific, understanding discrimination is challenging when examining any one set of institutions in isolation.*" Discriminatory behavior is intricately molded by numerous factors each time, potentially impacting a broader array of rights or groups. Regarding the factors influencing the shaping of discriminatory behavior, it is evident that they encompass the historical evolution and treatment of the rights of specific groups or even the groups themselves.

Finally, when reflecting on the entire legal framework concerning non-discrimination, it is essential to acknowledge that "*legal change is not the end of the road but, in some ways, the beginning.*" This statement carries a dual implication. On one hand, legal equality or protection against discrimination is insufficient in itself to fully redress societal and everyday equality. On the other hand, no legislative provision alone can comprehensively overcome all the enduring and historically ingrained discriminatory behaviors directed at specific groups and individual characteristics.

2.2.3 Decoding Discrimination: Disparate Treatment vs. Disparate Impact

Disparate treatment, a manifestation of intentional discrimination, denotes the intentional differential treatment of individuals based on specific characteristics, notably within the purview of employment law, where it stands as substantiated evidence of illegal discrimination, notably under Title VII of the United States Civil Rights Act. This discriminative paradigm is characterized by the bestowal of unequal treatment upon an employee predicated upon a protected characteristic, such as race or gender. To substantiate a disparate treatment claim, litigants must articulate intentional discrimination, thereby demonstrating that they experienced less favorable treatment stemming from motivation attributable to a protected characteristic. The

establishment of causation assumes paramount significance in the evidentiary framework of disparate treatment, with two methodological avenues for substantiating it: either by manifesting that the protected characteristic served as a "motivating factor" in the genesis of the adverse decision or by elucidating that it constituted a "but-for cause."

The conceptual underpinnings of disparate treatment resonate with the societal understanding of discriminatory behavior, necessitating meticulous consideration of the potential alterations in actions if an individual's protected characteristic underwent modification, all the while maintaining the constancy of other case facts. In the United States legal context, disparate treatment finds its antithesis in disparate impact, where ostensibly neutral rules inflict prejudicial consequences upon specific protected groups. Title VII, a legislative bastion against discrimination, explicitly prohibits differential treatment of applicants or employees predicated upon their membership in a protected class. A disparate treatment transgression materializes when an individual is singled out and subjected to less favorable treatment due to an impermissible criterion, thereby raising probing inquiries into the discriminatory intent underpinning the actions of the employer.

Conversely, disparate impact represents unintentional discrimination that disparately affects a specified group. In contradistinction to disparate treatment claims, the onus of proving intent is obviated within disparate impact cases, where the analytical focus pivots toward the discernment of discriminatory effects resultant from ostensibly neutral practices.

The applicability of this theory extends to predictive algorithms, whose deployment may unwittingly result in the disproportionate exclusion of racial minorities from employment opportunities, irrespective of the absence of overtly intentional discriminatory practices.

Diverging from disparate treatment, disparate impact centers on practices that, though unintentional, engender a disproportionate impact on a protected class. This necessitates a stringent criterion of justification and avoidability, implemented through a methodological framework reliant on burden-shifting. Beyond its diagnostic function in uncovering latent intentional discrimination, disparate impact is grounded in a motivation for distributive justice. It seeks to mitigate unjustified inequalities in outcomes, thereby aligning with the paradigm of equality of opportunity. In essence, disparate impact mandates decision-makers to accord similar treatment to ostensibly dissimilar individuals, seeking redress for existing dissimilarities stemming from historical injustices and compensating for disadvantages incurred due to unjust causes.

In contrast, disparate treatment is emblematic of intentional employment discrimination, exemplified by discriminatory practices such as the exclusive testing of specific skills for certain minority applicants. Disparate impact materializes when ostensibly neutral policies, like uniform testing for all applicants, inadvertently result in a disproportionate adverse impact on a protected group, thus exemplifying unintentional discrimination wherein procedures ostensibly identical for all adversely affect individuals within a protected class.

2.3 Fairness as a principle

Fairness is a concept that transcends cultural, societal, and individual boundaries. It's a fundamental principle deeply ingrained in human consciousness, reflecting our innate sense of justice and equity. While the specific interpretation of fairness may vary across cultures and contexts, its essence remains constant – the equitable and just treatment of all individuals. At its core, fairness represents the commitment to treat people impartially and without prejudice. It embodies the idea that everyone should have an equal opportunity to succeed, that rewards should be commensurate with efforts and contributions, and that discrimination and bias should be eliminated from all aspects of life.

Fairness manifests in diverse ways, from the distribution of resources and opportunities to the enforcement of laws and social norms. In economics, it's seen as an essential aspect of market systems, ensuring that transactions are conducted honestly and that wealth is distributed in a way that minimizes inequalities. In governance, it is the foundation of democratic principles, guaranteeing that laws are applied equally to all citizens.

The notion of fairness extends to personal interactions as well, guiding our moral compass in everyday life. It underlies the concepts of trust and reciprocity, promoting harmonious relationships in families, communities, and workplaces. Fairness is also at the heart of resolving conflicts and disputes, as it provides a framework for reaching equitable solutions. Despite its universality, fairness can be challenging to define precisely, as it often involves balancing conflicting interests and values. What may seem fair to one person might not be seen the same way by another, depending on their perspective and circumstances. Yet, the pursuit of fairness remains an enduring and aspirational ideal, inspiring individuals and societies to continually strive for a more just and equitable world.

Fairness is a fundamental principle that guides human behavior and societal structures. It embodies the ideals of justice, equity and impartiality, serving as a cornerstone for creating harmonious, ethical and prosperous societies across the globe. Nevertheless, the concept of fairness is inherently complex. It is traditionally defined as the quality of being fair, particularly emphasizing impartial treatment. However, this simplicity belies the intricate nature of fairness, as what is considered "fair" can be highly contingent on the specific context and the perspectives of different stakeholders.

Fairness in the context of artificial intelligence (AI) represents a multifaceted and evolving objective. Its core purpose is to establish AI systems that consistently deliver unbiased, equitable decisions while avoiding the perpetuation or exacerbation of societal inequalities. This goal arises from the recognition of AI's transformative potential and its potential impact on individuals and society. Within the domain of AI, "fairness" serves as a linchpin term. It is prominently featured in nearly all frameworks and principles governing responsible and ethical AI development. These principles advocate for the meticulous design of AI systems that adhere to legal frameworks, respect human rights, uphold democratic values, and promote diversity. This approach seeks to ensure that AI actively contributes to the cultivation of a fair and just society, minimizing the risk of algorithmic biases and discriminatory outcomes.

Yet, achieving fairness in AI is far from straightforward. It necessitates a granular understanding of the multifaceted facets of fairness, acknowledging its dynamic and context-dependent nature. Furthermore, the pursuit of fairness in AI is a continual process, requiring ongoing vigilance and adaptation to evolving societal norms and ethical standards. Consequently, it

prompts the need for agile AI development, robust testing, and continuous refinement to ensure that AI technologies operate in harmony with the ever-evolving definitions and expectations of fairness in our rapidly changing world.

2.4 The Artificial Intelligence Act

The Proposal for the Regulation on the Establishment of Harmonized Rules regarding Artificial Intelligence (Artificial Intelligence Act)[5] [6] of the European Parliament states, among other things, in point 3.5 of its Explanatory Report, regarding fundamental rights:

> "*The use of AI with its particular characteristics (e.g., opacity, complexity, data dependence, autonomous behavior) can negatively affect certain fundamental rights protected by the EU Charter of Fundamental Rights. This proposal aims to ensure a high level of protection for these fundamental rights and seeks to address various sources of risks through a clearly defined risk-based approach. With a set of requirements for trustworthy AI and proportional obligations for all participants in the value chain, the proposal will strengthen and promote the defense of rights protected by the Charter, such as the right to human dignity (Article 1), prohibition of discrimination (Article 21), and gender equality (Article 23), among others. Additionally, as applied in certain areas, the proposal will positively impact the rights of specific groups, including the rights of workers to fair and conducive working conditions (Article 31), high-level consumer protection (Article 28), children's rights (Article 24), and the inclusion of persons with disabilities (Article 26). Requirements for prior testing, risk management, and human oversight will also facilitate the respect of other fundamental rights, minimizing the risk of incorrect or biased decisions made with the assistance of AI in critical areas. In cases where violations of fundamental rights persist, effective remedies for affected individuals will be possible by ensuring transparency and traceability of AI systems, coupled with strict post-market controls*"".

Thus, from the outset, reference is made to the individual's fundamental rights, and there is explicit mention of fundamental rights that constitute the concept of fairness, such as equality and the prohibition of discrimination. Furthermore, the necessity to safeguard these rights in the use of AI is emphasized, along with outlining methods to ensure their protection. Furthermore, in the recitals of the Regulation, it is explained:

The need for a harmonized institutional framework to ensure the protection of fundamental rights, as stated in recital 5:

> "*A Union legal framework laying down harmonised rules on artificial intelligence is therefore needed to foster the development, use and uptake of artificial intelligence in the internal market that at the same time meets a high level of protection of public interests .... and the protection of fundamental rights, as recognised and protected by Union law..... By laying down those rules, this Regulation supports the objective of the Union of being a global leader in the development of secure, trustworthy and ethical artificial intelligence, ...*".

The necessity of non-discrimination in relation to the promotion of fundamental rights through common rules, as stated in recital 13:

[5] https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf
[6] https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

> "*In order to ensure a consistent and high level of protection of public interests as regards health, safety and fundamental rights, common normative standards for all high-risk AI systems should be established. Those standards should be consistent with the Charter of fundamental rights of the European Union (the Charter) and should be non-discriminatory and in line with the Union's international trade commitments*".

Specific use cases of AI that may lead to discrimination and the management of these cases, such as:

Recital 17:

> "*AI systems providing social scoring of natural persons for general purpose by public authorities or on their behalf may lead to discriminatory outcomes and the exclusion of certain groups. They may violate the right to dignity and non discrimination and the values of equality and justice. Such AI systems evaluate or classify the trustworthiness of natural persons based on their social behaviour in multiple contexts or known or predicted personal or personality characteristics. The social score obtained from such AI systems may lead to the detrimental or unfavourable treatment of natural persons or whole groups thereof in social contexts, which are unrelated to the context in which the data was originally generated or collected or to a detrimental treatment that is disproportionate or unjustified to the gravity of their social behaviour. Such AI systems should be therefore prohibited*".

Recital 36:

> "*AI systems used in employment, workers management and access to self-employment, ... should also be classified as high-risk, ... Throughout the recruitment process and in the evaluation, promotion, or retention of persons in work-related contractual relationships, such systems may perpetuate historical patterns of discrimination, for example against women, certain age groups, persons with disabilities, or persons of certain racial or ethnic origins or sexual orientation. AI systems used to monitor the performance and behaviour of these persons may also impact their rights to data protection and privacy*".

With the Proposal for the Regulation placing equality, non-discrimination, equal treatment, and the protection of the rights of all individuals as its guiding principles, Article 5 of this proposal introduces prohibitions on certain practices in AI. These practices include placing on the market, putting into operation, or using AI systems that exploit any vulnerabilities of specific groups of individuals, as well as placing on the market, putting into operation, or using AI systems to assess or classify the credibility of natural persons for a certain period based on their social behavior or known or predicted personal characteristics or personality traits. These practices are deemed to lead to harmful or prejudicial treatment of certain natural persons or entire groups of natural persons and are considered unjustified or disproportionate.

Upon a comprehensive evaluation of the text of the Regulation Proposal for the AI Act, which primarily introduces methods for assessing and managing AI tools, we can observe that even though it does not make specific or detailed references to the concept of fairness, it

consistently refers to the respect for fundamental rights. Furthermore, through its methodology and proposed controls, it introduces rules related to the promotion of fair, unbiased, and equal treatment of individuals.

2.5 The Ethics Guidelines for Trustworthy AI

In the context of promoting the reliability of AI, the European Commission has issued Ethical Guidelines for Trustworthy Artificial Intelligence [7]. Trustworthy AI has three components, which should be met throughout the system's entire life cycle:

1. It should be lawful, complying with all applicable laws and regulations;
2. It should be ethical, ensuring adherence to ethical principles and values; and
3. It should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

Each of these three components is necessary but not sufficient in itself to achieve Trustworthy AI.

2.5.1 Lawful AI

Regarding lawful AI, it is worth noting that AI is already developed within an environment where legal binding rules apply. Notable among these are the conditions set forth by the European Union Charter of Fundamental Rights, EU derivative laws (such as the General Data Protection Regulation and anti-discrimination directives), as well as United Nations human rights conditions and conventions of the Council of Europe (such as the European Convention on Human Rights), all of which include provisions for equality, individual freedom, and the prohibition of discrimination. Furthermore, the prerequisite for ethical AI involves the existence of ethical and moral rules, even if they are not explicitly or specifically included in legally binding regulations. It appears, therefore, that a necessary condition for AI to be considered trustworthy is compliance with legal and ethical rules, which, among other things, ensure that its use is conducted with legality, justice, equality, and impartiality.

In the Ethical Guidelines for Trustworthy AI, it is noted that the European Union's approach to AI ethics is based on fundamental rights, which are enshrined in the EU Treaties, the EU Charter of Fundamental Rights, and international human rights law. These rights are explicitly described and legally protected. As specifically mentioned in the text of the Ethical Guidelines for Trustworthy AI, "*The EU Treaties and the EU Charter prescribe a series of fundamental rights that EU member states and EU institutions are legally obliged to respect when implementing EU law. These rights are described in the EU Charter by reference to dignity, freedoms, equality and solidarity, citizens' rights and justice. The common foundation that unites these rights can be understood as rooted in respect for human dignity – thereby reflecting what we describe as a "human-centric approach" in which the human being enjoys a unique and inalienable moral status of primacy in the civil, political, economic and social fields.*"

The Ethical Guidelines also mention the fundamental rights categories considered relevant for ensuring trustworthy AI, among which are equality, the prohibition of discrimination, solidarity, and the respect for human dignity. The legal protection of these rights is, in many cases, mandatory, but the ethical and moral considerations surrounding them can contribute significantly to a proper understanding of what constitutes trustworthy and, therefore, fair AI.

[7] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

2.5.2 Ethical AI

The Ethical Guidelines also provide four ethical principles rooted in fundamental rights that should be followed to ensure that AI systems are developed, deployed, and used in a trustworthy manner. These principles are:

1. Respect for human autonomy
2. Prevention of harm
3. Justice
4. Explainability

Specifically regarding the principle of justice, the Ethical Guidelines state: "*The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity in terms of access to education, goods, services and technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable*".

The above text provides a clear description of what justice is, what role justice can play in the development, deployment, and use of AI systems, and how justice can be achieved in AI. Summarizing what was mentioned above, both in the Chapter on the Proposal for the AI Act and in the Ethics Guidelines for Trustworthy AI, we recognize the foundational role of the principle of justice in the development of AI systems, as well as how all fundamental rights are interlinked and interact to ensure justice and concepts included in it, such as human dignity and equality

2.6 AI Legislation: An international overview

The following subsection offers an overview of international initiatives to regulate ethical AI. A separate chapter is devoted to ongoing legislative procedures in the EU.

2.6.1 United States

To date, there is no comprehensive federal legislation regulating the use of AI in the United States (US). However, on June 20, 2023, the US lawmakers introduced a bill, the National AI Commission Act, to create a blue-ribbon commission that will review the United States' current approach to AI regulation, make recommendations on any new office or governmental structure that may be necessary, and develop a comprehensive framework for AI regulation.

Following are a few AI regulations that are in force at the state level in the US:

- **Connecticut's Artificial Intelligence Law** regulates the state's use of AI and has established a task force to develop an AI bill of rights and make recommendations for the adoption of other AI legislations. Along with establishing the Office of Artificial Intelligence and the Connecticut Artificial Intelligence Advisory Board, the law also establishes a task force to: (a) study artificial intelligence, and (b) develop an artificial intelligence bill of rights.

- **Illinois' Artificial Intelligence Video Interview Act** requires all employers using AI technologies to analyze candidates interviewing for employment positions to appropriately inform all applicants and gain their consent before subjecting them to this automated processing.

- **New York City's Law on Automated Employment Decision Tools** expressly prohibits employers from using an automated employment decision tool (AEDT) to make an employment decision unless the tool is audited for bias annually, the employer publishes a public summary of the audit, and the employer provides certain notices to applicants and employees who are subject to screening by the tool. Pursuant to the adoption of final implementing regulations on April 5, 2023, law enforcement began from July 5, 2023.

In 2020, the White House issued the **Guidance for Regulation of Artificial Intelligence Applications,** the purpose of which was to establish an appropriate framework for all relevant federal agencies that may have to regulate various emerging AI technologies, in addition to the ethical and legal issues that would arise in tandem.

The aforementioned Guidance has helped various US agencies formulate, from time to time, different guidelines, recommendations, and plans of their own. These include:

- The US Department of Defence's **AI Principles: Recommendations on the Ethical Use of Artificial Intelligence**;

- The US Food and Drug Administration's **Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan**;

- The Federal Trade Commission's:
  - **Using Artificial Intelligence and Algorithms guidelines**,
  - **Aiming for truth, fairness, and equity in your company's use of AI**,
  - **Keep your AI claims in check**,
  - **AI applications, deepfakes, and voiceclones: AI deception for sale**,
  - **The Luring Test: AI and the engineering of consumer trust**,
- The Department of Health & Human Services' **Trustworthy AI Playbook**,
- Consumer Financial Protection Bureau's **Rules to Implement the Dodd-Frank Act** governing the use of "automated valuation models" in the housing market.

In October 2022, the White House, per current US President Biden's direct instructions, issued a Blueprint for an AI Bill of Rights that laid down critical protections all US citizens must have as AI continues to expand in capabilities and functionalities. These include:

- **Data privacy:** A consumer should be protected from abusive data practices via built-in protections and the consumer should have an agency over how data about the consumer is used.
- **Notice & explanation:** A consumer should know that an automated system is being used and understand how and why it contributes to the outcomes that impact the consumer.
- **Algorithmic discrimination protection:** A consumer should not face discrimination by algorithms and systems should be used and designed in an equitable way.
- **Safe & effective systems:** A consumer should be protected from unsafe and ineffective systems.
- **Human alternatives, consideration, and fallback options:** A consumer should be able to opt-out, where appropriate, and have access to a person who can quickly consider and remedy problems the consumer encounters.

In January 2023, the National Institute of Standards & Technology issued its AI Risk Management Framework (AI RMF), which is aimed at offering a resource to the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems. The AI RMF is intended to be voluntary, rights-preserving, non-sector-specific, and use-case agnostic.

Most recently, in May 2023, the U.S. Congressional Research Service published its Generative Artificial Intelligence and Data Privacy: A Primer focusing on privacy issues and policy considerations for the U.S. Congress. The report sheds light on the collection and use of data by AI developers and the role data privacy legislation can play in regulating such use. The

report proposes the following three requirements/mechanisms that may be considered in privacy regulations to govern the use of data by AI developers:

Notice and disclosure requirements: Companies developing or deploying AI may be required to acquire consent from the individuals before collecting or using their data or notifying them that their data will be collected and used for certain purposes.

Opt-out requirements: Companies developing or deploying AI may be required to provide the data subjects an option to opt-out of data collection.

Deletion and minimization requirements: Companies developing or deploying AI may be required to provide mechanisms for data subjects to delete their data from existing datasets.

2.6.2 China

New Generation Artificial Intelligence Development Plan: While not a legislation per se, this comprehensive strategic plan acknowledges the need for AI to be developed ethically and calls for the establishment of ethical norms and policies. Given China's significant advancements in AI, steps taken within this plan have global implications, such as setting standards for how AI should respect user privacy and prevent data bias, which could influence AI fairness in applications like facial recognition technology.

The Administration of Deep Synthesis of Internet-based Information Services contains provisions that strictly punish deep synthesis technology such as deepfakes and other forms of AI-generated media.

The new Regulations also required all AI-generated content to be appropriately labeled as such.

The Regulations offer detailed guidance for the application of deep synthesis technology in providing Internet information services within China. They specify the responsibilities of national and local departments, highlighting the importance of information security, robust management systems, user authentication, content oversight, and effective measures against spreading rumors.

Furthermore, the regulations address the management of deep synthesis data and technology, emphasizing data security, regular evaluation of algorithms, and clear labeling of generated content. Adhering to these regulations is essential to prevent misuse, maintain transparency, and ensure responsible use of deep synthesis technology.

Moreover, the national network information department is responsible for coordinating the governance and related supervision and management of national in-depth synthesis services.

These Regulations also come with Frequently Asked Questions (FAQs). The FAQs clarify that deep synthesis service providers have responsibilities such as establishing management systems for user registration, algorithm review, data security, and personal information protection.

Regulation at the provincial level has been similarly proactive, with the Shanghai Regulations on Promoting the Development of the AI Industry and Shenzhen Special Economic Zone Artificial Intelligence Industry Promotion Regulations placing distinct obligations on subject organizations.

Shanghai Regulations apply to activities such as AI Science and Technology (S&T) innovation, industrial development, application empowerment, and industrial governance within the administrative region of Shanghai. These Regulations apply to all organizations in Shanghai involved in the AI industry. The regulatory authority is the local municipal economic and

information departments and is responsible for planning, implementing, coordinating, and promoting the development of the AI industry.

Shanghai Regulations are formulated in accordance with relevant laws and administrative regulations and based on the actual situation of the Shanghai area in order to promote the high-quality development of the AI industry. Additionally, these Regulations aim to strengthen the functions of new-generation AI S&T innovation sources, promote the deep integration of AI with the economy, everyday life, urban governance, and other fields, and create a world-class AI industrial cluster.

One of the major aims of The Shanghai AI Regulations is to facilitate the responsible and sustainable development of AI technology. It introduces grading management and "sandbox" supervision, which provide companies with opportunities to explore and test their technologies in a regulated environment. This approach encourages innovation while ensuring adherence to guidelines and standards.

Shenzhen AI Regulations have been formulated to promote the high-quality development of the AI industry in the Shenzhen Special Economic Zone, encourage AI integration in the economy and society, and ensure orderly and standardized industry growth in accordance with relevant laws and Shenzhen area situation. As per these Regulations, the local government will establish a working mechanism to coordinate and promote the development of the artificial intelligence industry in the city.

This includes ensuring the industry's security, fostering its healthy and orderly growth, and harnessing the potential of AI for sustainable development in the economy, society, and ecology.

The regulatory authority under the Shenzhen AI Regulations is the municipal industrial and information technology department which will serve as the competent authority responsible for implementing, coordinating, and supervising its development within the city's jurisdiction.

Shenzhen AI Regulations categorize activities and applications on three levels. High-risk AI applications require pre-assessment and risk early warning, while medium- and low-risk applications need pre-disclosure and post-tracking regulation. The Shenzhen area government will develop separate measures for classifying and supervising AI applications.

Additionally, AI services and products based in Shenzhen that are deemed to pose "low risk" can undergo testing and trials, even in the absence of local and national norms. However, adherence to international standards is a prerequisite for such testing and trials.

China has arguably been the most proactive country regarding regulating AI technologies and engaging various stakeholders to ensure the best ethical standards are adopted.

In 2017, the State Council of the People's Republic of China published A Next Generation Artificial Intelligence Development Plan. The guide contained a detailed roadmap on how various state and private institutions can help in the development, deployment, and oversight of AI technologies in a responsible manner.

Then, in 2021, the National Special Committee of New Generation Artificial Intelligence, a body established by the aforementioned guide, issued a Code of Ethics for New-Generation Artificial Intelligence to ensure any future development of AI technologies is in line with appropriate ethics and regulatory requirements. It also established six critical ethical standards that must be considered in developing such AI technologies. These include:

- Improving human well-being;
- Promoting fairness and justice;

- Protecting privacy and security;
- Ensuring controllability and credibility;
- Strengthening responsibility;
- Improving ethical literacy.

The following year in March 2022, the Internet Information Service Algorithmic Recommendation Management Provisions came into effect that required all organizations that develop, promote, and facilitate the development of AI-based personalized recommendations on mobile devices to allow users to delete any tags about their personal characteristics that the internal AI-recommendation model may have developed based on their browsing patterns.

It also required the organizations to let users disable such recommendations on their devices.

In November 2022, the Ministry of Public Security, the Cyberspace Administration of China (CAC), and the Ministry of Industry and Information Technology released a new set of regulations.

Most recently, CAC released a set of draft measures for managing generative AI services. These include requiring all organizations using such services to submit to an independent security assessment before such tools can be commercially deployed.

2.6.3 Australia – Human Rights Commission

The Australian Human Rights Commission has been proactive in assessing the impact of technology and AI on human rights. They have proposed a focus on AI decision-making, particularly emphasizing the use of AI in important decisions about individuals, such as their eligibility for disability benefits. By suggesting that legal accountability should be ensured for such AI-assisted decisions, they are pushing for legislation that would require, for example, that any AI system used in such contexts be transparent and contestable by the individuals affected.

Australia does not yet have a dedicated AI regulation. Most of its regulatory actions related to AI either come through existing laws or regular government policy papers guiding various regulatory bodies on how to approach different challenges posed by AI.

The New South Wales (NSW) Government came up with the first AI strategy, recognizing the challenges that come with the use of AI and charting a course for AI to be used safely across the government with the right safeguards in place.

For this purpose, the NSW Government published the AI Assurance Framework to assist agencies in designing, building, and using AI-enabled products and solutions. It is mandatory for all projects that incorporate an AI component or utilize AI-driven tools. This encompasses the utilization of large language models and generative AI, explicitly falling within the framework's application scope. The framework is intended to be used by:

- project teams who are using AI systems in their solutions,
- operational teams who are managing AI systems,
- Senior Officers who are accountable for the design and use of AI systems,
- internal assessors conducting agency self-assessments, and

- the AI review body (TBC).

However, a project is not expected to use the framework if it meets the following criteria:

- It uses an AI system that is a widely available commercial application.
- The solution is not customized or used in any way other than intended.

The AI Assurance Framework became effective in March 2022. The State Government also established the NSW AI Review Committee to provide expert guidance and oversight on using AI within the government. As the first of its kind in Australia, this committee plays a vital role in fostering community trust and ensuring transparency in our AI initiatives.

In March 2022, the government issued a call for papers on the regulation of AI, calling on various stakeholders on how the government should approach AI regulation in a manner that enables the creation of a harmonic legislative framework without jeopardizing the use of AI to its maximum potential.

In the paper, the government referred to several of its own reports and guides as examples of what kind of ideas it hoped to receive. These include the following:

- **The Artificial Intelligence Ethics Framework**, published in 2019 as a guide for government and private bodies on how to responsibly design, develop, and implement AI in Australia;
- **The Review of the Privacy Act**, which contained several recommendations related to automated decision-making systems and mechanisms;
- **The Australian Human Rights Commission Human Rights and Technology Final Report**, published in 2021, recommended the establishment of an independent AI safety commissioner to oversee various aspects related to commercial AI use in Australia;
- **The AI Action Plan**, published in 2021, laid down Australia's strategic vision to become a global leader in developing reliable AI technologies and systems;
- **The Blueprint for Critical Technologies**, published in 2021 as "framework for capitalizing on critical technologies to drive a technologically-advanced, future-ready nation."

Various other regulatory bodies in Australia have also undertaken steps on their own to promote the responsible use of AI under their jurisdiction. For example, since the Online Safety Act of 2021 has come into effect, the National eSafety Commissioner requires all organizations to appropriately inform their users of the use of automated recommendation systems.

Similarly, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) launched its own independent Responsible AI Network to promote collaboration between various Australian firms and create ethically safe and viable AI technologies.

2.6.4 India – Draft Personal Data Protection Bill

While India's focus with the Draft Personal Data Protection Bill is primarily on privacy, it has implications for AI fairness. The bill proposes restrictions and conditions on data processing, which would require companies to ensure that the AI systems they employ for processing personal data do so in a manner that respects individual privacy and does not propagate bias or discrimination. For instance, e-commerce companies using AI for personalized marketing would need to ensure that their systems do not inadvertently discriminate by showing certain products only to specific ethnic groups or genders.

2.6.5 Brazil

To date, there is no comprehensive federal legislation regulating the use of AI in Brazil.

However, in May 2023, the Bill of Law 2338/2023, which provides for the use of Artificial Intelligence (AI) in Brazil, was introduced in the Brazilian Federal Senate. The bill replaces three bills, Bill of Law 5.051/2019, Bill of Law 21/2020, and Bill of Law 872/2021, which were pending before the legislature over the past four years.

Along with imposing various obligations on businesses using AI systems, the law provides the following rights to the consumers:

- Right to prior information regarding their interactions with artificial intelligence systems.
- Right to an explanation of the decision, recommendation, or prediction made by artificial intelligence systems.
- Right to challenge decisions or predictions of artificial intelligence systems that produce legal effects or significantly impact the interests of the affected party.
- Right to human determination and human participation in decisions of artificial intelligence systems, taking into account the context and the state of the art of technological development.
- Right to non-discrimination and the correction of direct, indirect, illegal, or abusive discriminatory biases.
- Right to privacy and to the protection of personal data, in accordance with the relevant legislation.

To date, there are no comprehensive state legislations regulating the use of AI.

2.6.6 United Kingdom

In March 2023, the Department for Science, Innovation and Technology announced the introduction of the Data Protection and Digital Information (No. 2) Bill ('the Bill'). The Bill, amongst other objectives, aims to address the risks associated with AI-powered automated decision-making and determine the data protection controls required for such processes.

The Bill is expected to provide clarity on how the right to not be subjected to automated decision-making, as granted under Article 22 of the UK GDPR, can be invoked and exercised.

The UK Information Commissioner's Office, the body primarily responsible for overseeing all data privacy-related affairs in the UK, has released guidelines on how organizations can responsibly explain the use of AI to both their own employees and customers, titled Guidance on AI and Data Protection and AI and Data Protection Risk Toolkit. The ICO has also recently warned organizations using emotional analysis technologies irresponsibly.

The Guidance provides a roadmap to data protection compliance for developers and users of generative AI. The Risk Toolkit enables organizations to identify and mitigate data protection risks and contains eight significant questions that organizations developing or using generative AI should consider.

Moreover, other guidances in relation to the use of artificial intelligence have also been issued by different public entities in the UK. These include the following:

- **Data Ethics Framework** issued by the Department for Digital, Culture, Media, and Sport,
- **Intelligent Security Tools** guidance by the National Cyber Security Center,
- **Roadmap** issued by the UK Medicines and Healthcare Products Regulatory Agency.

The UK government issued a white paper in March 2023 titled "A pro-innovation approach to AI regulation." Within the whitepaper, the UK government highlighted the role of its existing legal framework in regulating the use of AI and underscored its "reputation for high-quality regulators and [UK's] robust approach to the rule of law, supported by [its] technology-neutral legislation and regulations. UK laws, regulators, and courts already address some of the emerging risks AI technologies pose."

The white paper further stated that "this strong legal foundation encourages investment in new technologies, enabling AI innovation to thrive and high-quality jobs to flourish."

Additionally, the whitepaper outlined five essential principles that all regulatory bodies must consider when evaluating the use of AI within their scope. These include:

- Safety, security, and robustness;
- Transparency and explainability;
- Fairness;
- Accountability and governance;
- Contestability and redress.

In the same 2023 whitepaper, the UK government also laid down its own plan on how it aims to curate both the national development and regulation of AI technologies. Partly inspired by its 2021 10-Year National AI Strategy, the UK government categorically ruled out establishing a new regulatory body or commission to oversee AI-related regulation.

Instead, existing regulatory bodies such as the Health & Safety Executive, the Equality & Human Rights Commission, and the Competition & Markets Authority will expand their powers and jurisdictions to ensure effective oversight of AI-related technologies within their sectors.

In March 2023, the UK government's Department of Education released another whitepaper titled "Generative Artificial Intelligence in Education" in response to the alarming growth in the use of ChatGPT by students nationwide.

The Centre for Data Ethics and Innovation (CDEI): The UK has established the CDEI to analyze and anticipate gaps in governance and regulation in AI and data-driven technologies. While it doesn't create laws, its recommendations can shape legislative and regulatory approaches in the UK, focusing on ensuring that AI decision-making is transparent, and systems are built and used in the public interest.

Established in 2018, the CDEI was created in response to the growing need for ethical standards in AI and data-driven technologies. It acknowledges that these technologies, while beneficial, pose risks such as privacy intrusion, biases, and decision opacity. The CDEI's role is to research these issues, engage with stakeholders, and propose measures to promote ethical AI development and deployment.

Objectives and Responsibilities: The CDEI has several key objectives:

- **Research and Analysis:** It conducts in-depth studies on emerging issues in AI and data use, assessing risks and benefits. For instance, it might explore how AI impacts employment sectors, identifying potential biases in AI recruitment tools.
- **Promoting Best Practices:** The CDEI advises on best practices for responsible AI and data usage. This isn't just about compliance with laws but also promoting fairness, transparency, and accountability in AI systems. For example, it might recommend methods for auditing AI systems for biases.
- **Advisory Reports:** It produces reports on various AI topics, providing evidence-based recommendations to the government. These reports might cover diverse areas like online targeting, predictive policing, or facial recognition technologies.
- **Dialogue and Engagement:** The CDEI facilitates dialogue between the public sector, private sector, and the general public to inform its recommendations. It might hold public consultations or workshops to understand different perspectives on AI-related issues.

Notable Work and Recommendations: The CDEI has undertaken various initiatives reflecting its broad mandate. Some of its notable work includes:

1. **Online Targeting:** The CDEI has reviewed online targeting practices (like personalized advertising) and their societal impacts. It recommended greater transparency and user control over online targeting, suggesting that companies disclose how they target users with content or ads.
2. **Algorithmic Bias:** Recognizing the risks of biases in decision-making AI, the CDEI has suggested developing tools and frameworks to detect and mitigate such biases. This includes recommendations for cross-sector standards for algorithmic transparency and accountability.
3. **Facial Recognition:** With the controversy surrounding facial recognition, the CDEI has called for clarity on legal issues, suggesting a need for explicit legislation on where, when, and how this technology should be used, ensuring it's deployed ethically and transparently.

Impact on Legislation and Policy: While the CDEI's recommendations are advisory and not legally binding, they are influential. They inform policymakers, potentially shaping future legislation on AI and data use. By highlighting ethical pitfalls and proposing balanced approaches, the CDEI helps steer the conversation on AI regulation towards more responsible, fair, and transparent practices, ensuring that the UK's legal framework remains robust in the face of rapidly evolving technologies.

2.6.7 Singapore – Model Governance Framework

Singapore has established itself as a leader in AI governance by introducing the Model AI Governance Framework. This is an accountability-based framework that helps companies ensure consumer trust in their AI solutions. For example, a company providing financial services through AI might use the framework to demonstrate how its systems are designed to avoid biased decision-making in offering loans, thereby ensuring that applicants from various demographics are treated fairly.

Singapore is one of the few countries in the world with a dedicated government body tasked with curating Singapore's digitalization journey. It aims to nurture a vibrant digital economy fuelled by technological innovation. The Advisory Council on the Ethical Use of AI and Data was established in 2018 to help Singapore identify and address all ethical questions and dilemmas that may arise. The body's primary responsibility is to advise the government on ethical, policy, and governance issues related to the use of AI technologies.

As a result of the body's recommendation, the country's first National Artificial Intelligence Strategy was published in 2019. Not only did it aim to identify strategic areas where resources need to be deployed most urgently while also addressing the emerging risk of AI becoming expansive beyond control.

Additionally, various existing regulations have been amended to include AI systems and technologies, such as:

- **The Road Traffic (Autonomous Vehicles) Rules 2017**[8] – The new amendments regulate the trial of all AI-driven autonomous vehicles;
- **The Cybersecurity Act of 2018**[9] – The new amendment requires all AI security methodologies and mechanisms adopted by organizations to be appropriately revealed to the users;
- **The Protection From Online Falsehoods and Manipulation Act 2019**[10] – The new amendments require all organizations to ensure their users' personal data is not used to create any fake avatars or personas online.

The Infocomm Media Development Authority (IMDA) has identified four distinct "pillar" technologies to drive Singapore's digitalization journey. These include:

- Cybersecurity;

[8] https://sso.agc.gov.sg/SL/RTA1961-S464-2017?DocDate=20170823
[9] https://sso.agc.gov.sg/Acts-Supp/9-2018/#: :text=An%20Act%20to%20require%20or,or%20related%20amendments%20to
[10] https://sso.agc.gov.sg/Acts-Supp/18-2019

- Immersive Media;
- The Internet of Things;
- Artificial Intelligence.

Each pillar has its own dedicated development program, with the AI section driven by AI Singapore[11], launched to build and grow Singapore's AI ecosystem, including research institutions, startups, and tech.

Two distinct AI programs[12], the National AI Program in Government and National AI Program in Finance, were established to guide various regulatory agencies via guidance and policy papers. Additionally, other government bodies have issued various other guides related to the use of AI within their sectors. These include the following:

- The Monetary Authority of Singapore: **Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector**;
- The Monetary Authority of Singapore: **Veritas Initiative**;
- The Personal Data Protection Commission (PDCP): **Model AI Governance Framework**;
- The Personal Data Protection Commission (PDCP): **Implementation and Self-Assessment Guide for Organisations** and a **Compendium of Use Cases Volume 1** and **Volume 2**;
- The Infocomm Media Development Authority (IMDA): **A Guide to Job Redesign in the Age of AI**;

2.6.8 Canada

To date, there is no comprehensive federal legislation regulating the use of AI in Canada.

However, in June 2022, the Government of Canada tabled the landmark Artificial Intelligence and Data Act (AIDA) as part of the omnibus Bill C-27, Digital Charter Implementation Act 2022. The AIDA aims to set out new measures to regulate international and inter-provincial trade and commerce in AI systems and establish common requirements for the design, development, and use of AI systems.

The law establishes common requirements for the design, development, and use of artificial intelligence systems and would also prohibit specific practices with data and artificial intelligence systems that may result in serious harm to individuals or their interests.

In March 2023, the Government of Canada issued the AIDA Companion document aimed at highlighting Canada's approach towards the regulation of AI and how AIDA shall contribute to that approach once enacted. The Companion document also identified a number of existing frameworks for consumer protection, human rights, and criminal law that apply to the use of AI, including the following:

[11] https://aisingapore.org/
[12] https://www.smartnation.gov.sg/media-hub/press-releases/new-ai-programmes-2021

- The Canada Consumer Product Safety Act;
- The Food and Drugs Act;
- The Motor Vehicle Safety Act;
- The Bank Act;
- The Canadian Human Rights Act and provincial human rights laws; and
- The Criminal Code.

As per the consultation timeline provided in the Companion document, AIDA would come into force no sooner than 2025.

To date, no province in Canada has enacted comprehensive legislation regulating the use of AI. However, the provincial human rights laws apply to the use of AI and afford some protections to consumers.

In November 2020, the Office of the Privacy Commissioner of Canada (OPC) issued A Regulatory Framework for AI: Recommendations for PIPEDA Reform containing OPC's final recommendations after the public consultation on proposals for ensuring the appropriate regulation of AI in the Personal Information Protection and Electronic Documents Act (PIPEDA). The recommendations, among others, included the recognition of privacy as a human right, specific provisions of automated decision-making, and demonstrable accountability of the business community.

In May 2021, the Government of Ontario published its report on Consultation: Ontario's Trustworthy Artificial Intelligence (AI) Framework. The report provided an overview of the potential actions the government could take to ensure the responsible and safe use of AI and the feedback from the consumers on those actions.

In April 2023, the Government of Canada issued a report on the Responsible use of artificial intelligence (AI), which describes how the government makes sure that the use of AI by its department and agencies is responsible and accountable. Among other things, the document outlines the following actions that need to be taken and monitored so governments may use AI responsibly:

- **Understand and measure** the impact of using AI by developing and sharing tools and approaches.
- **Be transparent** about how and when we are using AI, starting with a clear user need and public benefit.
- **Provide meaningful explanations** about AI decision-making, while also offering opportunities to review results and challenge these decisions.
- **Be as open as we can** by sharing source code, training data, and other relevant information, all while protecting personal information, system integration, and national security and defense.

- **Provide sufficient training** so that government employees developing and using AI solutions have the responsible design, function, and implementation skills needed to make AI-based public services better.

One practical effect of this approach can be seen in the launch of the Directive on Automated Decision Making – which is a policy directive by the Federal Government of Canada on how to responsibly incorporate AI decision-making within the public sphere.

2.6.9 The European Union

As with data privacy regulations in the form of the General Data Protection Regulation, the European Union (EU) looks increasingly likely to provide the rest of the world with an appropriate blueprint on how to proceed with AI regulation.

The AI Act The AI Act's categorization of AI systems according to risk is central to its regulatory approach. This stratification dictates the compliance requirements for AI systems and is intended to prevent consumer harm while encouraging innovation.

- **Unacceptable risk**: These AI applications are banned due to their potential to manipulate individuals through subliminal techniques or exploit vulnerabilities of specific groups that could cause significant material, physical, or psychological harm. For example, an AI system that manipulates public opinion by subtly promoting addictive behavior or harmful content, especially targeting minors or vulnerable individuals, falls under this category. The Act prohibits such systems outright, recognizing the severe negative impact they could have on society's fabric.

- **High-risk**: AI systems, such as those involved in traffic management, recruitment processes, or determination of eligibility for social security benefits, are considered high-risk because their failure or inaccuracy could directly affect human safety or fundamental rights. For instance, an AI application used in job recruitment must be transparent, with clear explanations of its decision-making process, to prevent discriminatory hiring. It should avoid biases that could disadvantage applicants based on gender, ethnicity, or disability. These systems require thorough testing, certification, and documentation before deployment.

- **Limited risk**: These are AI systems that require transparency measures to ensure users are aware they are interacting with AI. For example, a AI applications used for online shopping assistance should disclose that it's a bot, not a human. This disclosure helps users make informed decisions, knowing the interaction limitations and the potential lack of human-like judgment in responses.

- **Minimal risk**: The category for AI applications like those in entertainment (e.g., recommendation features for online streaming platforms) or non-essential services. These AIs have wide leeway for development and use, as their direct

impact on fundamental rights or safety is minimal. The Act encourages best practices but doesn't impose legal obligations, recognizing the low likelihood of significant adverse effects.

- **Legal Obligations for High-risk AI systems**: This aspect ensures that high-risk AI systems are held to stringent standards to safeguard individuals' rights and prevent harm.
  - For instance, if AI systems used in predictive policing are not banned, but classified as high-risk, they must adhere to strict data quality standards, ensuring data used in forecasting crime trends or identifying potential suspects is accurate and non-discriminatory. The system's algorithms would need to be transparent, with accessible records for review or audit, ensuring accountability and the possibility to challenge or appeal decisions made by the AI.
- **Transparency Requirements**: This provision mandates clarity about AI's role in decision-making processes, particularly when individuals might assume they are dealing with a human.
  - For example, in the case of AI-generated news articles, the outlet must clearly label the content as created by AI. This requirement prevents misinformation, as readers may critique or interpret AI-generated content differently, knowing it has not undergone human editorial processes reflecting judgment, ethics, or empathy.
- **Governance and Enforcement**: The Act proposes robust administrative structures for monitoring compliance and enforcing rules, critical for the Act's effectiveness.
  - If a company deploys a high-risk AI in public spaces, like facial recognition technology, it must comply with strict standards of data processing, consent, and privacy. The national supervisory authorities could perform random checks, review documentation, or demand demonstrations of the system's functionality to ensure compliance. Non-compliance could result in sanctions, including fines, a mandated modification of the AI system, or, in severe cases, a ban on the use of the AI system within the EU.

2.6.10 EU AI Act and Bias

The EU AI Act is particularly attentive to issues of fairness and combating bias in AI systems, recognizing that these technologies can inadvertently perpetuate discrimination and inequality if not properly managed. The Act addresses these concerns through several provisions aimed at ensuring AI systems operate fairly and without prejudice:

- **Data and Design Requirements for High-risk AI Systems**: The Act imposes rigorous demands on the quality of data used to train high-risk AI systems. For instance, developers of AI used in employment or finance sectors (e.g., AI for job applicant screening, loan eligibility assessments) must use datasets that are representative of all demographic segments, ensuring that the system's predictions or decisions are not biased against certain groups based on race, gender, age, or other characteristics.

  Additionally, these AI systems must undergo careful design and implementation processes to identify and mitigate any potential bias. For example, an AI system used in judicial sentencing must be meticulously evaluated to prevent biases that could lead to unfair sentence lengths based on a person's background rather than the facts of the case.

- **Transparency and Explainability**: The Act mandates that high-risk AI systems provide explanations of their decisions in a detailed and transparent manner. For example, if an AI system denies a loan application, it must provide the applicant with a comprehensible explanation of the decision-making criteria, ensuring individuals understand the reasoning and can identify potential grounds for unfair treatment.

  This level of transparency is crucial for sectors like recruitment, healthcare, or public services, where AI decisions significantly impact individuals' lives. It allows for the scrutiny of AI decisions, fostering fairness, and accountability.

- **Testing and Reporting Requirements**: Developers and operators of high-risk AI systems are required to conduct thorough testing and ongoing monitoring to check for and eliminate biases continuously. For instance, a company using AI for talent acquisition must regularly report on their system's decision-making patterns to a regulatory authority, demonstrating active measures taken to prevent discriminatory hiring practices.

  These entities must also keep detailed documentation and records on the system's development, functionality, and decision-making processes. This requirement ensures that any instance of unfairness or discrimination can be traced, analyzed, and corrected.

- **Redress Mechanisms**: The Act ensures that individuals affected by decisions made by high-risk AI systems have access to effective redress. For example, if a person suspects they were unfairly denied a job due to biased AI used in the selection process, they have the right to challenge the decision and seek human review. This aspect is crucial for maintaining fairness, as it empowers individuals to contest unjust outcomes and seek correction.

- **Enforcement and Penalties**: The Act empowers national authorities to enforce its provisions, including the imposition of penalties for non-compliance. For example, if an AI system used in housing allocation discriminates against applicants based on ethnicity or socioeconomic status, the responsible entities could face significant fines, forcing them to address and correct the unfair bias in their systems.

The Act elaborates at great length on how different technologies are categorized.

An AI system is deemed as having an Unacceptable Risk if it clearly endangers people's safety, livelihood, and fundamental rights. Such AI systems are completely prohibited.

Since the mechanism that will be used to categorize systems as either High Risk or Low Risk is still being debated, the aforementioned criteria is likely to be adjusted in the future.

The AI Act is expected to be adopted by the end of 2023 after due consideration, discussion, and necessary adjustments due to dynamic AI developments.

In June 2023, the Confederation of European Data Protection Organisations (CEDPO) published an AI and Personal Data guidance for Data Protection Officers. In the guidance, the CEDPO answered some fundamental questions that arise in relation to the intersection of the data protection legislative framework and the use of artificial intelligence and machine learning.

The guidance delves into matters such as the need for the AI Act, whether the GDPR regulates artificial intelligence and machine learning and which core data protection principles apply thereto, and the role of DPOs in the ever-evolving digital and technological landscape. The European Commission has also released guidelines on the Ethical Use of Artificial Intelligence in educational settings.

EU member countries have been in the headlines for their regulatory actions against emerging AI technologies. Italy became the first European country to temporarily ban the use of ChatGPT after its data protection authority, Garante, raised serious suspicions about ChatGPT's collection, use, and maintenance of users' personal data. The ban led other regulatory bodies in EU countries, such as France and Spain, to review the use of the famous AI applications in their own jurisdictions.

The Italian DPA also issued a provisional limitation on further processing of data by Replika – an AI application with a written and vocal interface based on AI that generates a "virtual friend" - highlighting violations of the GDPR.

Recently, the French DPA issued a 20 million Euro fine on Clearview AI for processing biometric data without an appropriate legal basis and the failure to exercise data subjects' rights and requests to erase their data. The Austrian DPA has also ruled that Clearview AI cannot process biometric data and must delete complainants' existing personal data.

The Finnish DPA has warned healthcare providers that automated decisions for detecting patients' healthcare needs can fail to meet data protection requirements.

Finally, in April 2023, the European Data Protection Board set up a taskforce dedicated to cooperation and exchange of information on possible enforcement actions by various data protection agencies across the EU. The EU Advocate General has also issued an opinion on the lawfulness of processing and automated decision-making under the GDPR and noted that an appropriate legal basis of data processing for AI systems to ensure compliance with the requirements of the GDPR.

2.7 Main social and ethical issues of AI

AI applications raise several potential ethical issues. An AI application is trained with data, usually derived from previous user interactions, and can continue to collect and learn from data provided during its development. This introduces potential issues around biasing training data that could lead to embedded information, biased responses, or offensive language. It also

raises security and privacy issues, depending on what the data includes and how it is stored and how or if permission is sought.

Designers of AI applications must also address questions about how to minimize the built-in bias in the models, as well as questions about the procedures in place to detect and prevent incorrect behavior, whether it is discriminatory, offensive language, or simply providing incorrect answers or advice.

Although many large companies, research institutions and public sector organisations have issued guidelines on ethical AI, recent research[13] has shown substantial divergences in the way they are interpreted, highlighting the difficulty of designing guidelines for systems with complex social impact.

The integration of AI systems into various social spheres brings about a number of often unpredictable and harmful effects. In addition, users from disadvantaged backgrounds, such as people with disabilities or those facing racial, gender or other prejudices, may face disproportionate harm.

Several studies demonstrate this, as bias can be detected in: pedestrian skin colour detection[14] , police and justice prediction systems[15] , job advertisement display in technology fields[16] , search engine policy[17] , medical applications[18] , automatic speech recognition[19] , and job recruitment algorithms.[20]

However, there are a number of ethical considerations that are unique to machine-human conversation that have not yet passed as de facto considerations in the design and development stages of building AI applications or other AI assistants.

2.7.1 Responsible artificial intelligence systems

According to Dignum (2019), "responsible AI means being responsible for the power that AI brings". Realizing the significant impact of responsible AI systems, the academic community, international and other organizations are trying to first understand and then act to mitigate the potential negative impacts of AI systems. This takes several forms, such as launching important ethical guidelines, principles and recommendations (HLEG, 2019; Floridi et al., 2018; Google, 2021; Microsoft, 2021), formulating theoretical and practical approaches, developing tools (PwC, 2019) and considering the possibility of a common language of AI ethics in the development and implementation of AI-based products and services (Morley et al., 2020).

[13]Jobin, A., Ienca, M., Vayena, E.: The global landscape of ai ethics guidelines.Nature Machine Intelligence pp. 1-11 (2019)

[14]Wilson, B., Hoffman, J., Morgenstern, J.: Predictive inequality in object detection. arXiv preprint arXiv:1902.11097 (2019).

[15]Richardson, R., Schultz, J., Crawford, K.: Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice New York University Law Review Online, Forthcoming (2019).

[16]Lambrecht, A., Tucker, C.: Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. management science (2019).

[17]Introna, L., Nissenbaum, H.: The politics of search engines, IEEE Spectrum 37(6), 26-27 (2000)

[18]Ferryman, K., Pitcan, M.: Fairness in precision medicine.Data & Society (2018).

[19]Tatman, R.: Gender and dialect bias in youtube's automatic captions. in Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. pp.53-59 (2017).

[20]Ajunwa, I., Friedler, S., Scheidegger, C.E., Venkatasubramanian, S.: Hiring by algorithm: predicting and preventing disparate impact. available at SSRN (2016).

There is a high degree of overlap and repetition among the ethical principles proposed in the literature (Ryan & Stahl, 2021). In order to address this duplication, Floridi (2013) introduced a framework of five principles (beneficence, nonmaleficence, autonomy, justice and explicability) adopting principles used in bioethics. Since then, this framework has been adopted by several projects, such as the Ethical Guidelines for Trustworthy AI published by the European Commission's High Level Expert Group on AI (HLEG, 2019) and the OECD Council Recommendation on AI (Floridi & Cowls, 2019).

However, there remain principles that are underrepresented in current academic debates, such as Transparency, Accountability, Responsibility and Fair Participation (Dignum, 2019; Vakkuri et al, 2019a) which do not take part in the synthesis to the core of AI accountability frameworks. The Berkman Klein Center counters by proposing eight key themes towards the adoption of more detailed principles: privacy, accountability, safety and security, transparency and accountability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values (Healey, 2020).

Although literature suggests an emerging convergence around the importance of certain ethical principles, a significant amount of research points to significant divergence between areas of application of the principles (Jobin et al., 2019). The literature review reveals specific challenges related to the use of AI systems in/by public sector organisations that can be classified into the following categories:

Ambiguity. Ambiguity is defined as the lack of understanding of how the system works (Vakkuri et al., 2020). It includes the algorithm, the data, as well as the technical aspects and the development process of the system. A major challenge is to establish transparency and explanatory power of AI systems. Many algorithms are incomprehensible to end users (Cath et al., 2017). The user should not only be able to understand how the system works, but should also be able to gather information about who built the system as it ultimately is and why (Vakkuri et al., 2020).

An important challenge with regard to e-government and the public sector is to ensure the explanatory power and transparency of the solutions used (David et al, 2019; Rahwan, 2017). Most AI systems used are incomprehensible to users, and although some techniques may focus on providing explanations, these are usually not fully accurate (Cysneiros & do Prado Leite, 2020). For this reason, there is a lack of public participation and engagement, meaning that service providers cannot easily verify that an AI solution is aligned with the interests of the public (Aitken et al, 2020; Marri et al., 2019). There are also difficulties related to low data quality and decision making, which can be a significant challenge in the human-centric public sector environment (Cath et al., 2017).

Vulnerability. Responsible AI systems offer benefits and improvements, but they also expose private data to risks and uncertainty. E-government leverages information and communication technologies to improve relationships between citizens, businesses and government, with the goal of better service, communication and government efficiency (Cath et al., 2017; Montes & Goertzel, 2019; Rahwan, 2017).

Particularly with the adoption of AI-based eGovernment services, threats to privacy and security remain major concerns. Previous findings have indicated that AI-based software solutions will require enhanced security due to the frequency of interaction and the volume of

information exchanged. E-government AI services reveal private data and may trigger cyber attacks and exploitation of personal information. This can create tangible threats not only to privacy but also to the security of citizens (Cysneiros & do Prado Leite, 2020; Gill, 2019).

Privacy. For the past 30 years, data protection has been a major issue in AI ethics in both the private and public sectors (Mason, 1986). Privacy is an ongoing concern, as demonstrated by various AI ethics guidelines and studies. Data protection has become a topic that needs to be reviewed (Juho, 2019). Although these positions seem to be mainly compliance-oriented, the need for a more proactive approach to privacy issues has been recognized (Kleindienst et al., 2017).

But human beliefs and artificial intelligence are constantly evolving. As a result, technological developments will drastically alter what culture considers acceptable and will examine how privacy positions have shifted due to the convenience offered by mobile phones, the internet and smart cities (Rahwan, 2017).

Smart cities offer municipalities new and unparalleled economic opportunities, but new developments are often followed by issues of protection and privacy. On top of that, a smart city needs a higher level of network access to personal data to support a wide range of different devices with different software and hardware capabilities. For example, a smart city with healthcare devices makes it easier for people to live by providing quick access to medical services. Smart cities offer connections through AI applications to citizens and medical service providers and assist public health experts by integrating medical systems and patient data records (Yang et al., 2019).

Accountability. AI applications are also being used to improve both e-government systems and citizen interactions (Cath et al., 2017; Rahwan, 2017). However, there are challenges associated with AI accountability in automated e-government applications and services (Marri et al., 2019).

Accountability is about the consequences imposed on an entity for certain actions and decisions. The absence of accountability in the case of autonomous systems and their possible misuse is one of the biggest challenges for organisations and end users/citizens. To maintain clarity, decisions must be made by and explained by the decision making algorithms used. In AI, accountability includes both the task of directing behaviour (forming values and making decisions) and the purpose of explanation by placing decisions in a broader context and aligning them with moral values (Dignum, 2019).

There are examples of e-government where both the public sector and the private sector are engaged in the responsibility and accountability of AI. In the case of the United Arab Emirates (UAE), both the government and the private sector are responsible for the implementation and adoption of AI policy in the e-services sector (Ghandour & Wood- ford, 2019). This has been achieved through the Ministry of AI, which was created primarily to enable the government to implement AI in its various sectors. The private sector will contribute to the development of AI by participating in research and integrating AI into various aspects of life.

One approach to address the problems arising from the adaptation of AI is to integrate AI technology into a citizen-centred, citizen-focused, citizen-first, inclusive policy. Consequently, some attention must be paid to generational, educational, income and language differences. Bias can also be avoided by involving multidisciplinary and representative groups in the

implementation of AI. AI cannot be used to make critical decisions that would have a huge impact on people's lives before this is achieved (Marri et al., 2019).

According to Rahwan (2017), developments in AI have raised several concerns regarding regulatory and governance mechanisms for autonomous machines and complex AI systems. It is suggested that algorithmic processes cannot be held accountable because they are 'black boxes' whose inner workings are not open to all stakeholders.

AI systems cannot be held accountable in the same way that humans can, as they are not capable of having intent or consciousness. However, the actions of AI systems can have a significant impact on people's lives, so there is a need to ensure that they are held accountable in some way. This can be done by making developers or companies responsible for the actions of their AI systems or by requiring them to implement systems of oversight and accountability. For example, this may include creating ethical guidelines and principles for the development of AI systems, conducting regular audits and testing to identify and address biases, and establishing clear procedures for reporting and addressing any unintended consequences of AI systems. Ultimately, ensuring accountability for AI systems will require a collaborative effort between the private sector, governments and civil society to develop and enforce clear

Bias. Other concerns include data-driven decision-making processes can exacerbate inequality as they can often be biased either by their nature or by the existence of biases in their training data. In addition, algorithms can create feedback loops that perpetuate inequalities, such as the use of AI in predictive policing or credit prediction, making it impossible for individuals to avoid the vicious cycle of poverty or exclusion.

Issues of bias are fundamentally related to the human evaluation of results obtained using AI, thus linking the creation of AI systems to ethical principles (Kuleshov et al., 2020). Algorithms are not intelligent on their own, but are trained by humans and their explicit biases are observable, with gender and race biases being more prevalent in machine learning algorithms at present. Making the right decision in the face of a moral conflict is extremely difficult. Laws are often based on fuzzy concepts, which can prove difficult to quantify. As a result, in recent years there has been increasing recognition of the need to ensure that AI systems are aligned with human values.

In the same way, as regards the area of AI use in the context of governance of public organisations, it is questionable whether a machine learning algorithm can be successfully aligned with ethical theories. If an ethical principle is chosen for the system, it should be applied before its development. Consequently, the programmer must be impartial in this case, which is also a prerequisite for responsible AI (Bjørgen et al., 2018). The US Federal Trade Commission (FTC) reported that algorithms based on big data sets could replicate existing patterns of discrimination, inherit previous biases of decision makers, or simply represent the widespread biases that remain in society (Howard & Borenstein, 2017).

Transparency. An important prerequisite for an algorithm to be considered fair and ethical is transparency, and the concept of transparency is multifaceted. When we refer to transparency in AI systems, we are interested, for example, in why the algorithm reaches a decision, how it did so, what steps can be attributed to the decision making process, what were the determinants that led to the choice, what alternatives could have emerged (McAleenan, 2020).

Many authors agree that a required characteristic for an AI system in all disciplines is that it should be transparent to avoid bias, be ethical and ensure autonomy and freedom (Rességuier & Rodrigues, 2020). This is particularly true in the public sector, where many researchers stress the importance of transparency in e-government services and practices where the public is directly involved (Marri et al., 2019). In particular, the European Union is a strong advocate of transparency in AI systems (Cath et al., 2017). For example, it proposes the proper documentation, coding and labelling of phases so that one can immediately identify why a mistake was made during the process and thus help prevent future

3 Methods for bias detection and assessment criteria

In this section, we focus on the algorithmic aspect of fairness/bias. First, we present an overview of the most prominent fairness definitions for classification, making an effort to intuitively explain them with examples to better communicate them to non-technical audiences. Since classification is mostly examined in the literature, we emphasize this ML task while providing briefer overviews for different settings, such as fairness in ranking and in spatial fairness. Then, we present some prominent techniques for bias detection, explaining the core principles behind the techniques and providing a critical analysis of their effectiveness and shortcomings. Finally, we present a set of criteria for the assessment of fairness definitions and bias detection methods, drawn both from the literature and the writers' background work on algorithmic fairness and explainability. These criteria are not drafted to rank existing definitions and methods but rather to help assess their suitability in different application fields and settings.

3.1 Fairness definitions for classification

We begin by presenting some of the attempts that have been made since 2016 to define measures of fairness. Since then, defining and quantifying fairness has become a prominent research topic in machine learning, given the absence of a universal definition. Different definitions can even be in contradiction with each other, making it challenging to assess machine learning models.

An important distinction among fairness definitions is between group and individual notions. While group fairness criteria compare quantities at a group level, typically identified by sensitive attributes, individual criteria compare individuals. We will start by defining some prominent group fairness definitions and translating the requirements of each definition into our running example of auditing hiring decisions of a model when the sensitive attribute is gender. This means that we will audit, with the help of each definition, if the model is biased against females or males, assuming that the positive outcome is *hire* and the negative outcome is *no-hire*.

3.1.1 Group fairness definitions

Group fairness definitions formalize the idea that an ML system should treat certain groups of individuals similarly, and therefore a fair outcome demands the existence of parity between different protected groups, such as those defined by gender or race.

In the following definitions, we will use the notation: R represents the final classification predicted by the binary classifier, Y represents the actual outcome, that is, the real classification of the individual, P represents the probability score of the classifier, S represents any subset of the individuals, and A denotes the sensitive attributes of the subjects.

Demographic parity. Demographic Parity [11] states that the proportion of each segment of a protected class (e.g. gender) should receive the positive outcome at equal rates. Therefore, a classifier is fair if the following formula is satisfied:

$$Pr(R = + \mid A = a) = Pr(R = + \mid A = b) \;\; \forall a, b \in A$$

Example: Suppose that we have 10 female and 20 male applicants for a job. If 10 males receive the outcome *hire*, then we have a 50% probability of males being hired. The model is considered fair if the probability of females receiving the outcome *hire* is also 50%, meaning that 5 females should be hired. If fewer than 5 females are hired, the model is biased against females, and if more than 5 females are hired, the model is biased against males.

Conditional statistical parity. Conditional statistical parity [9] also demands that subjects with both protected and unprotected characteristics should be equally likely to receive a positive classification prediction, but only when other legitimate factors are taken into account. This implies that demographic parity should hold, but only for specific subsets of the instances:

$$Pr(R = + \mid S = s, A = a) = Pr(R = + \mid S = s, A = b) \;\; \forall a, b \in A \; \forall s \in S$$

Example: Suppose that we have 10 female and 20 male applicants for a job, and we know that 10 male applicants are young, while 6 female applicants are young. If 5 young males receive the outcome *hire*, then we have a 50% probability of young males being hired. The model is considered fair if the probability of young females to receive the outcome *hire* is also 50% meaning that 3 young females should be hired. If fewer than 3 young females are hired, the model is biased against females, and if more than 3 young females are hired, the model is biased against males.

Equal opportunity. The definition of equal opportunity [14] requires the positive outcome to be independent of the protected class A, conditional on Y being an actual positive. Therefore, equal opportunity is based on the predicted outcome and the actual outcome whereas demographic parity and conditional statistical parity are only based on the predicted outcome.

$$Pr(R = + \mid Y = +, A = a) = Pr(R = + \mid Y = +, A = b) \;\; \forall a, b \in A$$

Example: Suppose that we have 10 female and 20 male applicants for a job and we know that 10 male applicants are good matches for the job and 6 female applicants are good matches for the job. If 5 males that are good matches get the outcome *hire*, then we have a 50% probability of males being hired conditioned they are good matches. The model is fair if the probability of females that are good matches to get the outcome *hire* is also 50% meaning that 3 females should be hired conditioned that they are good matches. If less than 3 females who are good matches are hired, then the model is biased against females and if more than 3 females who are good matches are hired, then the model is biased against males.

Equalized odds. This definition [14] is the most restrictive since it demands that individuals in protected and unprotected groups should have equal true positive rate and equal false positive rate, satisfying the formula:

$$Pr(R = + \mid Y = y, A = a) = Pr(R = + \mid Y = y, A = b) \;\; y \in \{+, -\} \;\; \forall a, b \in A$$

Example: Suppose that we have 6 female and 12 male applicants for a job and we know that 6 male applicants and 3 female applicants are good matches for the job, while other applicants are bad matches for the job. Furthermore, suppose that the model gives the outcome *hire*

to 9 applicants and the outcome *no-hire* to the other 9 applicants. If the 6 males that are good matches get the outcome *hire* and the other 6 males that are not good matches get the outcome *no-hire*, then we have a 100% probability of males being hired conditioned they are good matches and 100% probability of males not being hired conditioned they are not good matches. The model is considered fair if the probability of females getting outcome *hire* is 100% conditioned that they are good matches and the probability of getting the outcome *no-hire* is also 100%. This means that the model should hire all the 3 females who are good matches and reject all the 3 females who are bad matches.

Demographic Disparity. Next, we present a group fairness definition that uses a different way to assess group fairness of protected groups. More specifically, the definition applies to each protected group independently and checks if the fraction of accepted outcomes is larger than the fraction of the rejected outcomes:

$$Pr(R = + \,|A = a) \geq Pr(R = - \,|A = a) \;\; \forall a \in A$$

Example: Suppose that we have 10 female applicants. The model is fair towards females if it gives the outcome *hire* to more females than it gives the outcome *not-hire*. This means that if more than 5 females are rejected, then the model is unfair towards females.

Conditional Demographic Disparity. The Demographic Disparity definition [16] can be further refined to arrive at the Conditional Demographic disparity definition that conditions specific subgroups of the examined protected group.

$$Pr(R = + \,|S = s, A = a) \geq Pr(R = - \,|S = s, A = a) \;\; \forall a \in A \; \forall s \in S$$

Example: Suppose that we have 100 female applicants that apply to 5 different jobs and suppose that the model gives the outcome *hire* to 40 females and the outcome *no-hire* to 60 females. The definition of demographic disparity will conclude that the model is unfair. However, it may be the case that all females are accepted in the first 4 jobs and all females are rejected in the fifth job. The conditional demographic disparity will conclude that the model is fair towards females conditioned they apply for one of the first 4 jobs and unfairly conditioned they apply for the fifth job.

The last two group definitions are based on the actual outcome Y and the predicted probability score P of the classifier.

Calibration. A classifier satisfies this definition [8] if individuals with the same predicted probability score P have the same probability of being classified in the positive class when they belong to any of the protected groups:

$$Pr(Y = + \,|P = p, A = a) = Pr(Y = + \,|P = p, A = b) \;\; \forall p \in P \; a, b \in A$$

*Example:*Suppose that we have 10 female with probability score p to be hired for a job and 10 male applicants with probability score p to be hired for a job. If the probability of males to receive the outcome *hire* is p, then the model is considered fair if the probability of females to receive the outcome *hire* is also p, meaning that equal number of males and females with the same probability score should receive the outcome *hire*.

Well-calibration. This definition [24] is an extension of the previous definition. It states that when individuals of protected groups have the same predicted probability score P they must have the same probability of being classified in the positive class, and this probability must be equal to P:

$$Pr(Y = + \,|P = p, A = a) = Pr(Y = + \,|P = p, A = b) = P \;\; \forall p \in P \;\; a, b \in A$$

Example: Suppose again that we have 10 female with probability score p to be hired for a job and 10 male applicants with probability score p to be hired for a job. Again, well-calibration asks for an equal number of males and females with the same probability score to receive the outcome *hire*, but in this case the number of individuals from each protected group should match the predicted probability, meaning that $p * 10$ males and $p * 10$ females should receive the outcome *hire*.

3.1.2 Individual fairness definitions

Individual fairness definitions follow the principle that "similar individuals should receive similar treatments". The following definitions are used to audit if a model is fair towards individuals as opposed to group fairness definitions that audit fairness towards groups.

Fairness through Unawareness. This definition [27] requires to not explicitly employ sensitive features when making decisions. This is effectively a notion of individual fairness since two individuals differing only in the values of their sensitive attributes should receive the same outcome.
Example: Suppose that we have one male and one female individual who are identical in their job skills, education, and work experience and only these attributes are used from a model to predict the outcome *hire* or *no-hire*. Then, the model is considered fair, if it predicts the same outcome for both individuals, unfair otherwise.

Fairness through Awareness. This definition [11] seems to be similar to the previous one, although some technical details differentiate them. The definition requires that similar individuals should be mapped to the same outcome. However, in this context, similarity does not mean identical features for both individuals, as in the Fairness through Unawareness definition. Instead, it means that the individual should be mapped to a similar individual with different values in the sensitive attribute by adjusting all features that are correlated with the sensitive attribute.

Example: Suppose that we have a male individual and we know his job skills, education, work experience, weight, and height, and only these attributes are used from a model to predict the outcome *hire* or *no-hire*. Then, the model is fair if it predicts the same outcome for the corresponding female individual, which is not identical but has the proper adjustments in her weight and height.

Counterfactual Fairness. The last individual fairness definition [27] again appears to be very close to the Fairness through Awareness definition. However, it employs a different

technique to capture the notion of the similarity of individuals. More specifically, the definition states that if the value of a sensitive attribute of an individual changes, then the outcome predicted by the model should remain the same.

Example: A male individual received the outcome *hire*. We change the gender of the male individual to female (adjusting other features to this change) and let the model predict again as if the individual was female from the beginning. If the result is again *hire*, then the model is fair towards the individual, unfair otherwise.

3.2 Fairness definitions for other tasks

3.2.1 Fairness definitions for rankings

Research in fair machine learning mainly focuses on classification and prediction tasks, but there is also an extensive body of literature dedicated to models that produce ranked outputs [31, 46]. Fundamental machine learning problems with ranked output include, given a set of n items (representing individuals, products, or web pages), ordering the items in terms of the relevance of a given query or selecting and ordering a subset of k items that are "most" relevant to the given query. Much like in the classification setting, fairness in rankings can be categorized into definitions of individual fairness (treating similar individuals similarly) or group fairness (treating different groups of individuals similarly).

The focus of fairness in rankings primarily lies in group fairness, whether it pertains to the entire ranking or only the top-k positions. Fairness principles may require a minimum number or proportion of items from a protected group to be evenly distributed throughout the ranking, also referred to as proportional fairness [6]. However, the main motivation behind the development of fairness metrics for rankings is rooted in the concept of exposure, which pertains to the visibility or attention that an item can receive based on its position (can be also found as click probability).

This concept led to measures of statistical parity in ranking schemes. These metrics aim to assess whether belonging to a protected group has an impact on an item's position in the ranked output, by quantifying the relative representation of a protected group S^+ within a ranking $\boldsymbol{\tau}$ [45], e.g., normalized discounted difference (rND), normalized discounted ratio (rRD), normalized discounted KL-divergence(rKL), etc. These are normalized to the $[0, 1]$ range for ease of interpretation, with 0 as the most fair value and 1 as the worst.

Normalized discounted difference (rND) It computes the difference in the proportions of members of the protected group S^+ at the top-k and in the overall population. Normalizer Z is computed as the highest possible value of rND for a given number of items n and protected group size $|S^+|$.

$$\text{rND}(\boldsymbol{\tau}) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{N} \frac{1}{log_2 k} \left| \frac{|S^+_{1\ldots k}|}{i} - \frac{|S^+|}{N} \right|$$

Normalized discounted ratio (rRD) It is formulated similarly to rND, with the difference in the denominator of the fractions: the size of $S^+_{1\ldots i}$ is compared to the size of $S^-_{1\ldots k}$, not to k (and similarly for the second term, S^+). When either the numerator or the denominator of a fraction is 0, we set the value of the fraction to 0.

$$\text{rRD}(\boldsymbol{\tau}) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{log_2 k} \left| \frac{|S^+_{1\ldots k}|}{|S^-_{1\ldots k}|} - \frac{|S^+|}{|S^-|} \right|$$

rRD is only applicable when the protected group is the minority group, i.e., when S^+ corresponds to at most 50% of the underlying population, and when the fairness probability is below 0.5.

Normalized discounted KL-divergence(rKL) It uses the Kullback-Leibler (KL) divergence to quantify the expectation of the logarithmic difference between two discrete probability distributions P_k, which quantifies the proportions in which groups are represented in the top-k, and Q, which quantifies the proportions in which groups are represented in the overall ranking.

$$P_k = \left(\frac{|S^+_{1...k}|}{k}, \frac{|S^-_{1...k}|}{k} \right), Q = \left(\frac{|S^+|}{n}, \frac{|S^-|}{n} \right)$$

Specifically, KL-divergence between P_k and Q

$$D_{KL}(P||Q) = \sum_k P(k) log \frac{P(k)}{Q(k)}$$

is computed at every cut-off point k, and position-based discounting is applied as the values are compounded, with normalizer Z defined analogously as for rND, producing the rKL definition

$$\text{rKL}(\boldsymbol{\tau}) = \frac{1}{Z} \sum_{k=10,20,...}^{n} \frac{D_{KL}(P_k||Q)}{log_2 k}$$

Unlike rND and rRD, which are limited to a binary sensitive attribute, rKL can handle a multinary sensitive attribute and so is more flexible.

3.2.2 Spatial Fairness

In many cases, it is important to ensure that the algorithm does not discriminate against individuals based on their location (place of origin, home address, etc.).In this context, we consider location as the protected attribute and we want the algorithm to exhibit *spatial fairness*. Unlike typical protected attributes, such as race and gender, location is a continuous attribute. The group-based definition of fairness does not apply straightforwardly to continuous protected attributes. Instead, the standard approach involves first discretizing the continuous domain to create groups, such as age or income groups, and then comparing the outcomes for each group. The same concept can be applied to location by defining non-overlapping spatial partitions (e.g., city blocks, zip codes, districts) and computing the measure in each. This leads to a partitioning-based definition, wherein an algorithm is spatially fair if the measures per partition are equal

The problem of spatial fairness has not received much attention in the literature despite its potential applicability in a plethora of real-world applications. Consider, for example, an algorithm that decides whether mortgage loan applications are accepted or not. It is desirable that its decisions do not discriminate based on the home address of the applicant. This could be to avoid redlining, i.e., indirectly discriminating based on ethnicity/race due to strong correlations between the home address and certain ethnic/racial groups, or to avoid gentrification, e.g., when applications in a poor urban area are systematically rejected to attract wealthier people. As another example, consider crime forecasting, where an algorithm predicts how likely a crime is to occur in a particular area. It is desirable that the algorithm is spatially fair in terms of its accuracy. That is, we require the predicted crime rate to not differ greatly from the observed crime rate in all areas. This could be to avoid under- and over-policing and the sense of injustice they are typically associated with.

In particular, there currently exist two definitions for spatial fairness in the literature. The first one, hereafter denoted as **MeanVar**, derived by [44], is based on the idea of superimposing a high-resolution grid over the space and considering all possible rectangular, grid-aligned partitionings of the space. In each partitioning, the variance of the measure in the partitions is computed. Then, the mean-variance across all partitionings is computed. Lower values of MeanVar suggest lower variance across the partitions in all partitionings and hence more fairness. Note that MeanVar is designed to assess the spatial fairness of outcomes that are regularly distributed in space: in each cell of the superimposed grid (or equivalently in each partition of the partitioning with the highest resolution), there is roughly the same number of outcomes.

To handle cases where outcomes are arbitrarily distributed in space, in [34] we introduce the second definition of spatial fairness, **ScanFair**. Since location is the protected attribute, so we naturally consider an algorithm to be spatially fair, if the measure is independent of location. This implies that for any region of the space, the distribution of the measure inside and outside the region should be the same. Operationalizing this definition poses several challenges, with the most complicated being how to determine the distribution of the measure within a region. If the region covers many observations, the observed (empirical) distribution of the measure is a good proxy for the actual distribution. Otherwise, what we observe in a sparse or small region might differ dramatically from what we observe outside it. Note that this issue does not manifest itself in categorical protected attributes (e.g., gender), simply because the number of observations per protected group (e.g., number of women and men) is typically very large. To address this issue, instead of looking at the observed distribution, we want to express how likely it is to observe such a distribution if the algorithm is fair. Intuitively, we should expect to find a region with only four negative points clustered together even when the algorithm is fair. Conversely, we should not expect to find a region that contains thirty negative points alone - observing such a region should be a stronger indication that the algorithm is not fair. Inspired by the work in spatial scan statistics [25], we form two hypotheses and seek to quantify which one is a better fit for the data observed. The null hypothesis is that of spatial fairness, i.e., there is a single distribution that controls how the measure is distributed in the space, or mnemonically $inside = outside$. The alternate hypothesis states that there is a difference in the distribution inside and outside a region, or $inside! = outside$. Given the observed data, we can determine the maximum likelihood for each hypothesis, compute their ratio, and test whether this likelihood ratio is statistically significant at a desired level. We note that this spatial fairness definition can be utilized to represent various notions of statistical fairness, in the spatial context, as defined in the previous subsection, e.g. *statistical parity*, *equal opportunity*.

3.3 Prominent methods for bias detection

In the previous section, we explored the most common approach for auditing the fairness of the model: applying fairness criteria that aggregate statistics of the model's behavior on protected subpopulations. While aggregate statistics can reveal broad patterns of potential discrimination, they do not provide additional information that sheds light on the underlying discriminatory mechanism at play, which is crucial when assessing whether the behavior is truly problematic. A growing set of strategies has emerged for testing and detecting such discriminatory behaviors, and this section is dedicated to presenting some of them that are considered state-of-the-art.

3.3.1 Fairtest

Fairtest is an instantiation of the methodology introduced in [37] that implements the framework of unwarranted associations (UA). UA is a principled methodology for discovering unfair, discriminatory, or offensive user treatment in data-driven applications, unifying and rationalizing prior attempts at formalizing algorithmic fairness.

While a general way to indicate an unwarranted association would be the presence of a strong statistical dependency on an algorithm's output concerning the protected class, the authors find such a definition to be fuzzy. It lacks a wide-scope applicability, a method to provide scalable assessment, and the inclusion of any natural explanatory factors to justify the perceived bias. Therefore, they informally define unwarranted associations as any strong associations between the algorithm's output and the attributes of a protected user group. These associations arise in a meaningful subset of users, have no explanatory factors, and can be used in a testing toolkit for a wide variety of tasks and datasets.

The proposed UA framework is packaged in Fairtest, a testing toolkit that enables scalable and statistically rigorous investigation of associations between application outcomes and sensitive user attributes. Fairtest uniquely combines multiple investigative primitives and fairness metrics with broad applicability, granular exploration of unfair treatment in user subgroups, and the incorporation of natural notions of utility that may account for observed disparities. It can investigate disparate impact, offensive labeling, and uneven rates of algorithmic error.

Approach. Consider an algorithm that uses the data collected on the users, including sensitive features like location or age. The output of the algorithm that needs to be inspected is denoted as O. The accuracy of O could have different user utility depending on the task. The attributes of the dataset can be characterized into three categories.

1. **Protected attributes**, denoted as S, are the primary attributes along which discrimination can occur. Typically, a group of individuals with certain S values constitutes a sensitive group and is protected by the law and policies.

2. **Contextual attributes**, denoted as X, are the attributes along which a population can be split to highlight hidden unwarranted associations. X is usually a proxy attribute for S that can be directly used by the algorithm and unwittingly reveal S values to the algorithm.

3. **Explanatory attributes**, denoted as E, are the attributes whose values can justify a seemingly discriminatory behavior by the algorithm.

The subjectiveness of the attributes can be explained by considering a hiring decision as an example. A company would want to hire candidates with more experience, even though more experience can be a proxy for a candidate's age or gender. Based on the values that O and S can take, the authors classify the choice of metrics that can be used to assess the strength of the association between O and S below:

- **Frequency Distribution Metrics:** They are often useful when examining the algorithm output for unwarranted associations.
- **Mutual Information**: They are used when testing for associations between O and S.
- **Correlation:** Calculates Pearson's correlation to quantify the relationship between O and S, when they are linearly dependent. It is used when the users want to profile the algorithm for errors.
- **Regression:** Regression is employed when the outputs of the algorithm are not known a priori or when the domain of output values is very large. The regression coefficient for each output value can provide evidence for the strength of association.
- **Conditional Metric**: It is used when looking for explanatory factors for a possible unwarranted association.

The authors point out that looking for associations across the full user population is not useful, as discrimination tends to occur in specific user groups. Therefore, the need is to search for a smaller but meaningful subpopulation that exhibits higher association. The framework uses association-guided tree construction to accomplish this. Since the associations can be justified in the presence of explanatory factors, discovering the association bugs is not a one-shot process. The framework is designed for multiple subsequent inspections, supported by statistical validity.

The final output of FairTest is an association report that an auditor can read to gain insights into possible discrimination. In Figure 1 and Figure 2, we can see two examples of association reports produced from FairTest, in the Adult dataset [1], which reports census data and income (under or over 50K $) for 48,842 U.S. citizens.

3.3.2 Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness

Statistical fairness definitions fix a small collection of high-level, predefined groups (such as race or gender), and then ask for approximate parity of some statistics of the classifier like positive classification rate or false-positive rate across these groups. Constraints of this form are susceptible to fairness gerrymandering, in which a classifier appears to be fair to each individual group, but badly violates the fairness constraint on one or more structured subgroups defined over the protected attributes (such as certain combinations of protected attribute values).

In [21] the authors propose a methodology that demands statistical notions of fairness across exponentially (or infinitely) many subgroups, defined by a structured class of functions over the protected attributes to prevent gerrymandering. Their method faces several computational challenges since it interpolates between statistical definitions of fairness, and recently proposed individual notions of fairness. They can prove that the computational problem of auditing

Report of associations of O=Income on S_i=Gender:
Global Population of size 24,421
p-value = 1.44e-178 ; NMI = [0.0381, 0.0540]

Income	Female	Male	Total
<=50K	7218(89%)	11422(70%)	18640 (76%)
>50K	876(11%)	4905(30%)	5781 (24%)
Total	8094(33%)	16327(67%)	24421(100%)

1. Subpopulation of size 1,371
Context = 9 <= Education <= 11, Age >= 47
p-value = 2.23e-35 ; NMI = [0.0529, 0.1442]

Income	Female	Male	Total
<=50K	423(88%)	497(56%)	920 (67%)
>50K	57(12%)	394(44%)	451 (33%)
Total	480(35%)	891(65%)	1371(100%)

2. Subpopulation of size 6,791
Context = Education >= 12
p-value = 3.71e-124 ; NMI = [0.0517, 0.0883]

Income	Female	Male	Total
<=50K	1594(76%)	2156(46%)	3750 (55%)
>50K	492(24%)	2549(54%)	3041 (45%)
Total	2086(31%)	4705(69%)	6791(100%)

Figure 1: The full population and two subpopulations with higher disparate effects when gender is the protected attribute

Report of associations of O=Income on S_i=Race:
Global Population of size 24,421
p-value = 1.39e-53 ; NMI = [0.0063, 0.0139]

Income	Asian	Black	...	White	Total
<=50K	556(73%)	2061(88%)		15647(75%)	18640 (76%)
>50K	206(27%)	287(12%)		5238(25%)	5781 (24%)
Total	762 (3%)	2348(10%)	...	20885(86%)	24421(100%)

1. Subpopulation of size 341
Context = Age <= 42, Hours <= 55, Job: Fed-gov
p-value = 3.24e-03 ; NMI = [0.0085, 0.1310]

Income	Asian	Black	...	White	Total
<=50K	10(71%)	62(91%)		153(63%)	239 (70%)
>50K	4(29%)	6 (9%)		91(37%)	102 (30%)
Total	14 (4%)	68(20%)	...	244(72%)	341(100%)

2. Subpopulation of size 14,477
Context = Age <= 42, Hours <= 55
p-value = 7.50e-31 ; NMI = [0.0070, 0.0187]

Income	Asian	Black	...	White	Total
<=50K	362(79%)	1408(93%)		10113(83%)	12157 (84%)
>50K	97(21%)	101 (7%)		2098(17%)	2320 (16%)
Total	459 (3%)	1509(10%)	...	12211(84%)	14477(100%)

Figure 2: The full population and two subpopulations with higher disparate effects when race is the protected attribute. Additional races were omitted for clarity.

subgroup fairness for both equality of false positive rates and statistical parity is equivalent to the problem of weak agnostic learning — which means it is computationally hard in the worst case, even for simple structured subclasses.

The paper contributes by deriving two algorithms that converge to the best fair distribution over classifiers within a given class, provided access to oracles capable of optimally solving the agnostic learning problem. These algorithms are formulated based on subgroup fairness as a two-player zero-sum game between a Learner (primal player) and an Auditor (dual player). Both algorithms compute an equilibrium of this game.

The first algorithm simulates the game by having the Learner play an instance of the no-regret Follow the Perturbed Leader algorithm, while the Auditor plays the best response. This algorithm provably converges to an approximate Nash equilibrium and, consequently, to an approximately optimal subgroup-fair distribution over classifiers in a polynomial number of steps.

The second algorithm simulates the play of the game by having both players play Fictitious Play which only has asymptotic convergence but has the merit of simplicity and faster per-step computation. The Fictitious Play version is implemented using linear regression as a heuristic oracle and shows that we can effectively both audit and learn fair classifiers on real datasets.

We should emphasize that the underlying assumption for deriving both algorithms is the existence of an efficient oracle capable of solving the agnostic learning problem. Then, it is proved that the optimal fair classifier can be found as the equilibrium of a two-player, zero-sum game, in which the strategy space of the "Learner" player corresponds to classifiers, and the strategy space of the "Auditor" player corresponds to subgroups. The best response problems for the two players correspond to agnostic learning and auditing, respectively. Then, it is proved that both problems can be solved with a single call to a cost-sensitive classification oracle, which is equivalent to an agnostic learning oracle.

In the first algorithm, the Learner employs the Follow the Perturbed Leader (FTPL) algorithm on an appropriate linearization of its best-response problem, while the Auditor approximately best-responds to the distribution over classifiers of the Learner at each step. Since FTPL has a no-regret guarantee, the first algorithm provably converges in a polynomial number of steps. The first algorithm is randomized and the best-response step for the Auditor is polynomial time but computationally expensive.

Therefore a second algorithm is proposed that is deterministic, simpler, and faster but has weaker theoretical guarantees. The second algorithm is based on both players adopting the Fictitious Play learning dynamic and is more practical due to its simplicity and its fast convergence in practice.

3.3.3 Optimal Transport of Classifiers to Fairness

Fairness criteria and methods usually force the prediction of classifiers to have similar statistical properties for people of different demographics. The violation of these properties is reduced by simply rescaling the classifier scores, ignoring similarities and dissimilarities between members of different groups. However, this information is relevant in quantifying the unfairness of a given classifier and can be exploited by using the transportation theory and more specifically optimal transport.

Transportation theory studies the optimal transportation and allocation of resources. We

will focus on optimal transport which is the general problem of moving one distribution of mass to another as efficiently as possible. Therefore, optimal transport can be used to calculate the distance between distributions of protected groups (e.g., males and females). Moreover, it provides a very efficient way to correct the unfairness since it can suggest efficient ways to make the distribution of one protected group identical to the distribution of the other protected group, therefore removing the corresponding biases.

One of the most prominent works that have considered optimal transport to audit unfairness can be found in [5]. They introduced the method of Optimal Transport to Fairness (OTF) that quantifies the violation of fairness constraints as the smallest Optimal Transport cost between a probabilistic classifier and any score function that satisfies these constraints.

Approach. Assume that a set F is available that denotes all score functions that satisfy the required notion of fairness. The unfairness of a classifier's score function h is quantified as the amount of 'work' minimally required to make it a member of this set by measuring how far h is from its fair projection onto F, i.e. the $f \in F$ that is closest to h. The proximity is measured as the optimal transport cost between h and f and then a higher unfairness cost to h is assigned if scores need to be transported between highly dissimilar individuals to reach a fair function f.

Thus, the methodology proposes to quantify unfairness as the Optimal Transport to Fairness (OTF) cost, i.e., the cost of the optimal transport (OT) based projection of h onto F:

$$OTF(h) = \min_{f \in F} OT(h, f)$$

The method computes OTF as a differentiable fairness regularizer that can be added to the training loss of any probabilistic classifier and is efficiently computed for popular notions of fairness such as demographic parity and equalized odds. This approach is capable of handling multiple sensitive variables, that can be categorical or continuous.

3.3.4 FlipTest

FlipTest [4] is a comprehensive and interpretable technique motivated by the question: had an individual been of different protected status, would the model have treated them differently? FlipTest leverages optimal transport to transform the distribution between protected class labels and observe the shift in the model outcome. A change in model outcome indicates discrimination based on the protected attribute. The authors also highlight that the optimal transport mapping does not depend on causal relationships to capture the discrimination caused at the outcome stage without considering any assumptions about the underlying data.

Computing an optimal transport map is computationally expensive, and the cost grows with the increase in the size and dimensions of the dataset. To that end, the authors also introduce and validate a faster and more efficient approximation method to compute optimal transport maps using Generative Adversarial Networks (GANs). The model output is assumed to be binary, positive, and negative, and the results of the optimal transport map from one class to the other are termed flipsets. Finally, a transparency report, that highlights the differences between the flipsets, is created, which provides an insight into what features might be responsible for the discrimination.

Approach Consider the two distributions S and S' for two classes over the feature space. Let n be the number of samples drawn from these two distributions, such that the set $S = \{x_1, \ldots, x_n\}$ and set $S' = \{x'_1, \ldots, x'_n\}$, where $n = |S| = |S'|$. Although the two sets are of equal size here, it is not a hard requirement for the proposed approximation method. The cost function $c(x, x')$ denotes the cost of moving a point from S to S'. An optimal transport map accomplishes moving between two points by minimizing the cost function:

$$\mathbb{E}[c(x, f(x))] = \frac{1}{n} \sum_{i=1}^{n} c(x_i, f(x_i))$$

Robust approximations for computing the optimal transport map use the implementation of Wasserstein GAN (WGAN) to train a generator G to learn the optimal transport mapping. The generator's loss function is modified to include the cost function specified above. The sets of individuals whose outcome changes for their mapped counterparts are called the *flipsets*, which we will define through the following example.

Example: Let S denote the female candidates, S' denote the male candidates, and h be the binary decision function whose output determines if the candidate should be hired or not. Then $F^+(h, G)$ would be the set of female candidates who are hired, but their mapped male counterparts are not. Similarly, $F^-(h, G)$ would be the female candidates who are not hired, but their male counterparts are hired. Conversely, the mapping could be reversed to become $G' : S' \to S$ to generate male flipsets. In an ideal scenario, when the input data is independent of any information about the protected attributes, the distributions S and S' would be equal, G would become an identity function, and the positive and negative flipsets would be empty. On the other hand, if the two flipsets are nonempty but of equal size, it can indicate demographic parity but does not necessarily mean individual fairness.

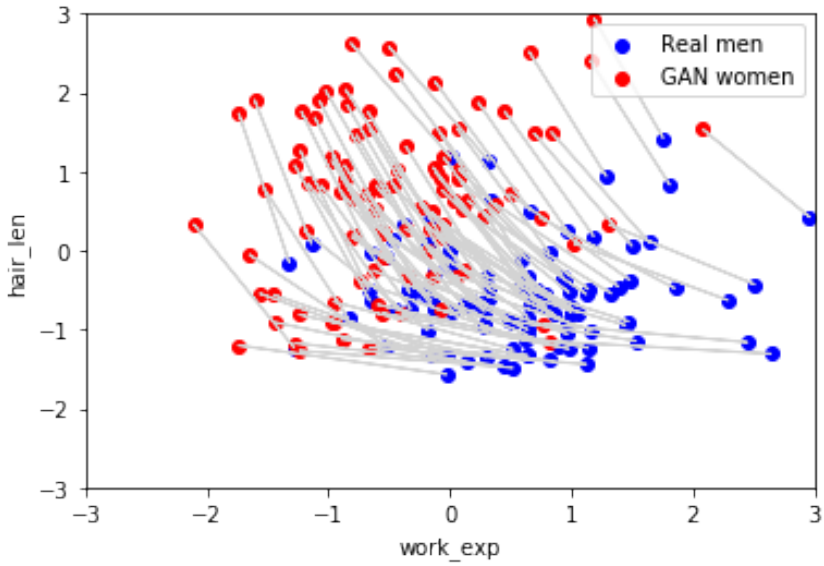

Figure 3: Optimal transport mapping of men to women with GAN

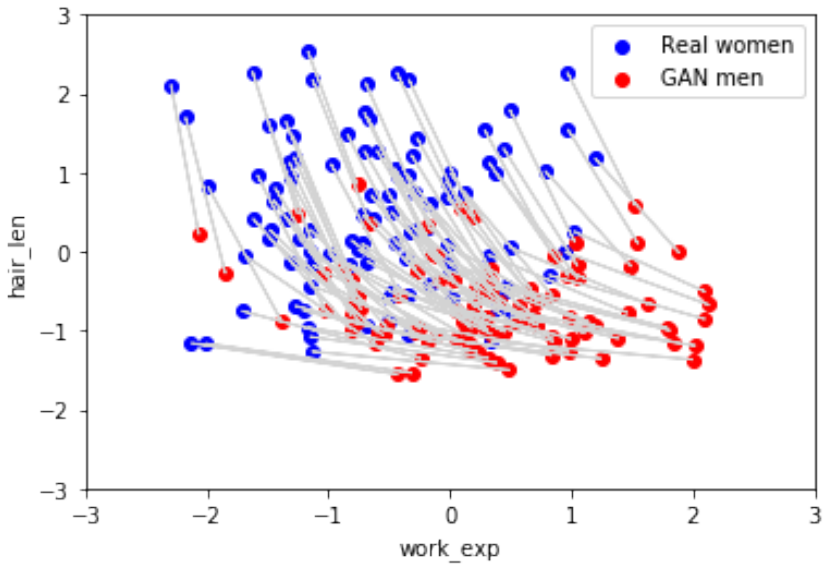

Figure 4: Optimal transport mapping of women to men with GAN

3.3.5 Fairness of Recourse

Bias towards protected subgroups is captured by the previous methods exploiting notions of *fairness of prediction*, where all subgroups defined by a protected attribute should have the same probability of being assigned the positive (favorable) predicted class. These definitions capture the *explicit* bias reflected in the model's predictions. Nevertheless, an *implicit* form of bias is the difficulty for, or the *burden* [36, 26] of, an individual (or a group thereof) to achieve *recourse*, i.e., perform the necessary *actions* to change their features to obtain the favorable outcome [13, 38]. Recourse provides explainability (i.e., a counterfactual explanation [41]) and actionability to an affected individual, and is a legal necessity in various domains, e.g., the Equal Credit Opportunity Act mandates that an individual can demand to learn the reasons for a loan denial. *Fairness of recourse* captures the notion that the protected subgroups should bear equal burden [13, 18, 39, 26].

Fairness of recourse is a more recent notion, related to *counterfactual explanations* [41], which explains a prediction for an individual (the factual) by presenting the "best" counterfactual that would result in the opposite prediction, offering thus *recourse* to the individual. Best, typically means the *nearest counterfactual* in terms of a distance metric in the feature space. Another perspective is to consider the *action* that transforms a factual into a counterfactual and specify a *cost function* to quantify the effort required by an individual to perform an action. In the simplest case, the cost function can be the distance between factual and counterfactual, but it can also encode the *feasibility* of an action (e.g., it is impossible to decrease age) and the *plausibility* of a counterfactual (e.g., it is out-of-distribution). It is also possible to view actions as interventions that act on a structural causal model capturing cause-effect relationships among attributes [19]. Hereafter, we adopt the most general definition, where a cost function is available, and assume that the best counterfactual explanation is the one that comes from the minimum cost action. Counterfactual explanations have been suggested as a mechanism to detect possible bias against protected subgroups, e.g., when they require changes in protected

attributes [17].

Fairness of recourse, first introduced in [38] and formalized in [13], is defined at the group level as the disparity of the mean cost to achieve recourse (called burden in subsequent works) among the protected subgroups. Fairness of recourse for an individual is when they require the same cost to achieve recourse in the actual world and in an imaginary world where they would have a different value in the protected attribute [39]. This definition however only applies when a structural causal model of the world is available. In [20] the authors introduce FACTS that expands on these ideas and proposes alternate definitions that capture fairness of recourse.

FACTS. Fairness Aware Counterfactuals for Subgroups (FACTS) is a framework for auditing subgroup fairness through counterfactual explanations. FACTS formulates different aspects of the difficulty of individuals in certain subgroups to achieve recourse, i.e. receive the desired outcome, considering the subgroup as a whole, and introduces notions of subgroup fairness that are robust, if not oblivious, to the cost of achieving recourse. The notions are the following:

- **Equal Effectiveness**

  The classifier is considered to act fairly for a population group if the same proportion of individuals in the protected subgroups can achieve recourse.

- **Equal Effectiveness within Budget**

  The classifier is considered to act fairly for a population group if the same proportion of individuals in the protected subgroups can achieve recourse with a cost at most c, where c is some user-provided cost budget.

- **Equal Cost of Effectiveness**

  The classifier is considered to act fairly for a population group if the minimum cost required to be sufficiently effective in the protected subgroups is equal.

- **Equal (Conditional) Mean Recourse**

  This definition extends the notion of burden from literature (reference) to the case where not all individuals may achieve recourse. Omitting some details, given any set of individuals, the conditional mean recourse cost is the mean recourse cost among the subset of individuals that can actually achieve recourse, i.e. by at least one of the available actions. Given the above, this definition considers the classifier to act fairly for a population group if the (conditional) mean recourse cost for the protected subgroups is the same.

- **Fair Effectiveness-Cost Trade-Off**

  This is the strictest definition, which considers the classifier to act fairly for a population group only if the protected subgroups have the same effectiveness-cost distribution (checked in the implementation via a statistical test).

In Figure 5, we can see some of the subgroups produced by FACTS that exhibit maximum bias in terms of the corresponding metric used.

```
Subgroup 1
If age=(41.0, 50.0],marital-status=Never-married,race=White,relationship=Not-in-family:
    Protected Subgroup = 'Male', 1.34% covered
        No recourses for this subgroup.
    Protected Subgroup = 'Female', 1.47% covered
        Make marital-status=Married-civ-spouse,relationship=Married with effectiveness 70.49%
    Bias against 'Male' due to Equal Cost of Effectiveness (threshold = 0.7). Unfairness score = inf.
```

```
Subgroup 2
If worklass=Private,hours-per-week=FullTime,marital-status=Married-civ-spouse,occupation=Adm-clerical,race=White:
  Protected Subgroup = 'Male', 1.04% covered
    Make hours-per-week=OverTime,occupation=Exec-managerial with effectiveness 70.00%
    Make hours-per-week=OverTime,occupation=Prof-specialty with effectiveness 70.00%
    Make hours-per-week=BrainDrain,occupation=Exec-managerial with effectiveness 70.00%
    Make hours-per-week=BrainDrain,occupation=Prof-specialty with effectiveness 70.00%
    Make Workclass=Self-emp-in,occupation=Exec-managerial with effectiveness 70.00%
    Make Workclass=Self-emp-in,hours-per-week=OverTime,occupation=Exec-managerial with effectiveness 80.00%
    Make Workclass=Self-emp-in,hours-per-week=OverTime,occupation=Sales with effectiveness 70.00%
    Make Workclass=Self-emp-in,hours-per-week=BrainDrain,occupation=Exec-managerial with effectiveness 70.00%
  Protected Subgroup = 'Female', 3.51% covered
    Make Workclass=Self-emp-in,hours-per-week=OverTime,occupation=Exec-managerial with effectiveness 74.51%
    Make Workclass=Self-emp-in,hours-per-week=BrainDrain,occupation=Exec-managerial with effectiveness 74.51%
  Bias against 'Female' due to Equal Choice for Recourse (threshold = 0.7). Unfairness score = 6
```

```
Subgroup 3
If age=(41.0, 50.0],occupation=Sales:
    Protected Subgroup 'Male', 1.18% covered
        Make occupation Craft-repair with effectiveness 0.00%
        Make occupation=Adm-clerical with effectiveness 0.00%
        Make occupation=Tech-support with effectiveness 19.23%
        Make occupation=Prof-specialty with effectiveness 28.21%
        Make occupation=Exec-managerial with effectiveness 39.74%
    Protected Subgroup 'Female', 1.56% covered
        Make occupation=Craft-repair with effectiveness 0.00%
        Make occupation=Adm-clerical with effectiveness 0.00%
        Make occupation=Tech-support with effectiveness 0.00%
        Make occupation=Exec-managerial with effectiveness 6.94%
        Make occupation=Prof-specialty with effectiveness 6.94%
        Make age=(50.0,90.0] with effectiveness 6.94%
        Make age=(50.0,90.0],occupation=Prof-specialty with effectiveness 6.94%
        Make age=(50.0,90.0],occupation=Craft-repair with effectiveness 6.94%
        Make age=(50.0,90.0],occupation=Adm-clerical with effectiveness 6.94%
        Make age=(50.0,90.0],occupation=Exec-managerial with effectiveness 6.94%
    Bias against 'Female' due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf.
```

Figure 5: Subgroup counterfactuals produced by FACTS in the Adult dataset .

3.4 Criteria for the assessment of bias detection methods

In this section, we briefly discuss a series of criteria that we have drafted for assessing the suitability of fairness policies and bias detection methods. These criteria have been formulated considering substantial risks, gaps, and issues identified in the literature at the intersection of ethics, law, and algorithmic fairness in decision-making, such as indirect/proxy discrimination, intersectional discrimination, and other related concerns.

We note that this section serves as an introduction to these criteria; Sections 4 and 5 further elaborate on these criteria, examining relevant existing law, precedents, and use cases, as well as assessing prominent bias detection methodologies on their grounds.

3.4.1 Equal treatment vs equal outcome - disparate treatment vs disparate impact - affirmative action

Title VII of the Civil Rights Act of 1964 (USA)[21] [23] prohibits discrimination in employment based on the sensitive attributes of race, color, religion, sex, and national origin. In this setting, discrimination may be expressed in the form of *disparate treatment* or *disparate outcome*.

[21] https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964

- **Disparate treatment** refers to cases where intentional discrimination based on a sensitive attribute happens. In our running example, a job position excluding immigrants would constitute disparate treatment, since it discriminates against groups of individuals based on their national origin.

- **Disparate impact** refers to cases where even though there exists no explicit discrimination on grounds of any sensitive attributes, and the decisions are taken via facially neutral practices, nevertheless specific protected groups end up being (disproportionally) disadvantaged. In such cases, discrimination might be unintentional and can be attributed to causes such as structural/historical bias, poor rule/model selection, etc. In our scenario, requiring excellent knowledge of a country's official language, for a job that requires manual labor, might comprise disparate outcomes for groups defined by national origin. As aptly stated in [23]:

  "*Disparate impact theory is relevant to predictive algorithms because these tools may disproportionately screen out racial minorities from employment opportunities, even if the employer did not intend to discriminate when adopting the tool.*"

A similar distinction is made with respect to the notion of equality a policy, practice or method aims to achieve [33]. In particular:

- **Equality of opportunity** prescribes that all individuals are given the same chances to achieve a favorable outcome. In our example, this means that a recommendation system should assign an "accept" label to candidates based on objective criteria and labeled training data, that do not take into account sex information in its decision.

- **Equality of outcome** prescribes that all protected (sub)groups equally/proportionally obtain the favorable outcome. In our example, the ratio of males to females that obtain the "accept" label should be proportionate to the respective candidate's ratio, even if this conflicts with the rating produced by the recommender.

  Equality of outcome is a notion that is based on the recognition of **structural** inequalities and **historical** bias in procedures and datasets and is achieved via instruments such as **affirmative action**[22] (or **positive action**, **positive discrimination**), which aim at alleviating/eliminating such inequalities against sensitive subpopulations. In our case, affirmative action or a company's policy would require a minimum quota in female "acceptances" for every job.

3.4.2 Handling of proxy variables and correlations - indirect discrimination

A substantial issue when it comes to auditing discrimination, algorithmic or not, lies in cases of **proxy discrimination**. This is the case where bias in the data, or a trained ML model, is expressed not directly via sensitive attributes, but indirectly, via proxy variables, that are to some extent correlated with the respective sensitive attribute [3, 30]. Examples of indirect/proxy discrimination can be found in *height* and *maternity leave* attributes serving as proxies for the *sex* sensitive attribute, *residence/location* attributes serving as proxies for the *race* sensitive attribute, etc.

[22] https://plato.stanford.edu/entries/affirmative-action/

Due to the commonly encountered misunderstanding that, upon sensitive attributes are excluded from an AI model's training, fairness is ensured (also called *fairness by unawareness* [27]), bias can be perpetuated via proxy discrimination. That is, even if sensitive attributes are removed, the bias of the training data can still be transferred into the trained model: the ML algorithm will try to learn patterns that relate the labels of the data, that express this bias, with remaining features that correlate with the removed sensitive attribute. Take for example a training dataset on hiring, that is significantly biased against female individuals, i.e., in a binary classification setting (*hire* or *no-hire*), for very similar jobs, almost exclusively male individuals are hired and female ones are rarely hired. Suppose that the model owner removes the sex attribute from each individual and uses the training dataset to train a binary classifier for recommending whether to hire or not a new applicant, whose sex is, of course, also unknown. While the sex attribute is absent from each individual, there most probably exist other attributes that are correlated with it, such as *university name* or *years of experience after graduation*. Specific values for these attributes will also be correlated with the biased labels. For example, individuals from certain universities, which are traditionally attended by proportionally more female students, will mostly be assigned the *no-hire* label in the training dataset. The ML algorithm, during its training, will learn such patterns, transferring this implicit bias into the trained model. Thus, for a new individual candidate, even if the sex is unknown, if they have attended the specific universities, they are most probably going to be assigned the biased, *no-hire* label.

The above also serves as an example of a related issue of implicit bias, termed as *discrimination by association* [40, 12]. This issue appears when individuals are mistakenly categorized as part of a protected group, which faces discrimination, and consequently experience the same type of discrimination. In our example, the training data, and the derived ML model are biased towards female individuals and, by correlations, also towards individuals that have attended specific universities, even if they are males.

3.4.3 Handling of intersectional-subgroup fairness

Another important issue arises when considering subpopulations defined by more than one attribute, with at least one of them being a sensitive one. In such cases, *subgroup fairness* [21] or *intersectional discrimination*[23] or multi-dimensional discrimination [32] is audited. Consider for example a scenario where an AI system decides whether a person is promoted based on a specific feature set and we want to audit the fairness of the system concerning two sensitive attributes: *gender*, with values $\{$*male*, *female*$\}$ and *race*, with values $\{$*Caucasian*, *non-Caucasian*$\}$. In our scenario, the unprotected groups are defined by *female* and *non-Caucasian* respectively. It might be the case that auditing fairness individually on the two sensitive attributes finds the promotion decisions of the system fair, however, by further examining the result, one might identify that non-Caucasian males and Caucasian females are disproportionally unfavored compared to the other two subgroups, i.e., Caucasian males and non-Caucasian females.

Subgroup bias presents quite a few challenges. First of all, it is often the case that bias is magnified for specific subgroups; this is not only due to preexisting discrimination towards individuals belonging to these groups but also due to their under-representation in datasets and

[23] https://www.coe.int/en/web/gender-matters/intersectionality-and-multiple-discrimination

models. That is, since these groups comprise only a very small subset of the general population, they are often not properly represented in sampling/training data-gathering processes. Another issue related to this data sparsity is the uncertainty in evaluating bias for these subgroups when auditing a dataset or algorithm: since very few instances representing a specific subgroup might be found in an audited dataset, the significance of the findings can be questionable. Finally, computational issues arise when trying to drill down to more granular subgroups, since complexity increases exponentially [21].

3.4.4 Handling of feedback loops

Feedback loops, in our context, comprise self-repeating processes that can potentially reinforce and perpetuate preexisting bias [22]. Consider our running example of a hiring recommendation system. If such a system is initially trained on a biased dataset, then its recommendations will probably reproduce the bias (if no fairness-correcting action is taken). Then, these new recommendations can be used as additional training data, that also carry bias. Further, applying the system in real-world domains and continuously rejecting female candidates in favor of male ones, might discourage individuals from the formerly protected groups from applying for specific job positions. It is well recognized in the literature that the pattern recognition and learning mechanisms applied by many AI systems can facilitate the creation of feedback loops [22, 3, 12].

3.4.5 Robustness to manipulation

Discrimination can often be intentional, meaning that the system/data/application owner is aware of preexisting bias and does, or explicitly introduces bias and applies no bias-correction actions or even tries to hide it. Masking bias can be achieved through various ways, including gerrymandering [10], as discussed above, as well as by manipulating the output of fairness auditing/explainability methods to render the audited model seemingly fair, while it is not. The work of [10] priminently demonstrates how a classifier can be retrained in an adversarial way, to maintain the same level of accuracy, and at the same time suppress the explicit contribution of sensitive attributes, so that a large set of explainability methods are tricked into deciding that its outputs are fair, while they are not.

3.4.6 Sampling requirements

Since the data we are working with are massive, we will need to confine ourselves to a smaller representative portion of it, i.e., a sample. By **Sample Complexity**, we refer to the minimum number of portions taken from a population to reliably account for bias.

In essence, bias detection involves calculating distances between two probability distributions derived from our data. This is commonly known as *two-sample problem* or *homogeneity testing*. One probability distribution describes our population, while the second, loosely speaking, pertains to the predictions made by the AI algorithm under study. There are different such distances: *Hellinger, Wasserstein (OT, cf. 3.3.3), Maximum Mean Discrepancy (MMD)*, etc. They will be calculated with an accuracy increasing in the number of samples taken.

Since we do not know the underlying probability distribution of our population and only have access to it through the limited instances in the dataset, we will give an approximation of

the distances between distributions (a measure of the bias), rather than an exact result. We further describe the relationship between the number of samples, n, and the error detecting the bias as in [35], using the MMD distance. We know that n must be smaller than a function of the error:

$$n \leq \frac{k}{|\text{error}|^2} \ , \ \text{k being a constant.} \tag{1}$$

At the same time, it is greater, up to a certain probability, ϵ, than another expression:

$$\text{Prob } \{n \geq \frac{1}{2^7|\text{error}|^2}\} > \epsilon. \tag{2}$$

In the previous reference, the probability takes the value $\frac{1}{5}$, but it can be shown that by making use of *Probability Amplification* techniques, the error $,\epsilon$, can be decreased at the expense of further computational cost.

We can graphically represent the actual behavior of our variables as lying between the two curves depicted in Figure 6. As these bounds show, to achieve an acceptable error, a high

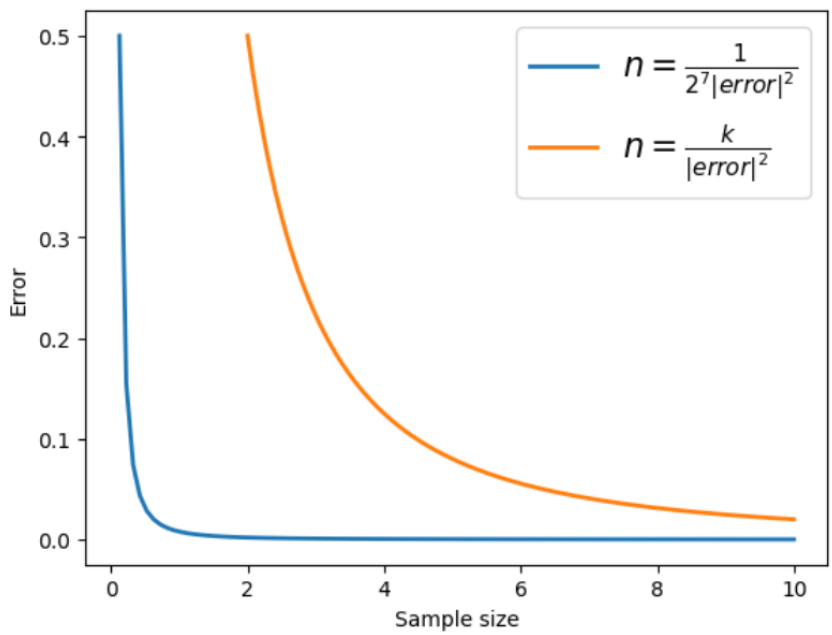

Figure 6: Lower and upper bounds of error detecting the bias with respect to sample size.

number of samples must be considered, with its corresponding computational cost.

Some approaches aim to enhance performance by relaxing certain assumptions, as seen in the Nyström method for the MMD distance. The distance computed using this method may not meet the strict requirements of a proper distance but has proven effective in practice, as recently guaranteed in [7].

3.4.7 Transparency and explainability

Explainability is a concept highly intertwined with fairness. Explainability allows AI systems to be more transparent, reliable and trustworthy and can help towards auditing bias and pursuing

fairness. Although many AI systems are black box, there exists a plethora of model-agnostic explainability methods that can produce diverse types of explaination in terms of complexity and form applied on top of such models. Various prominent studies [12, 15] emphasize the utility of explainability as a tool for designing, debugging and auditing AI systems with respect to fairness.

4 A study of fairness policies, precedents and use cases

In this section, we first present a study on existing cases of law, as well as real world application use cases that can be considered, implicitly or explicitly relevant to algorithmic fairness and can be used to draw intuitions about directions and best practices with respect to fairness policies. Then, we present a broader overview of existing policies, frameworks and tools pursuing ethical AI.

4.1 EU precedents relevant to algorithmic fairness

Next, we present a list of EU judicial precedents that are relevant to our analysis. We note that the list is not exhaustive, since identifying all precedents that might be relevant to AI fairness is out of scope of the deliverable and the project. Instead, we focus on a small set of prominent cases that are widely discussed in the relevant literature and that highlight existing gaps between EU law on discrimination and algorithmic fairness or indicate potential routes for resolution.

4.1.1 Intersectional-subgroup fairness

Intersectional, or subgroup fairness is a dimension that is admittedly poorly handled by the EU law and Court [12, 32]. Next, we present two exemplary cases that highlight the limitations of the current legal framework in addressing a highly significant dimension of algorithmic fairness.

The first example comes from the *Parris v Trinity College Dublin* case of 2016 [24], where intersectional discrimination regarding sexual orientation and age was examined. In this case, receiving survivor's pension from the same sex partner of an individual was denied due to the law not allowing the partners to enter a civil partnership early enough in order to satisfy age restrictions required to receive the pension. In this case, "*The CJEU found that a provision in a scheme's rules which required members to marry before the age of 60 for full survivors' benefits to be payable did not constitute discrimination on the grounds of age or sexual orientation. This was despite the fact that it was legally impossible for the claimant to enter into a civil partnership or same sex marriage before reaching that age.*"[25].

Even though the court rejected the case of the plaintiff, failing to recongize intersectional discrimination, the Advocate General recognized the existence of indirect discrimination based on sexual orientation as it was legally impossible for homosexuals born in Ireland before 1951 to enter into such a civil partnership or marriage before they reached the age limit. [12] points out that the Advocate General opinion *has suggested that recognising the concept of intersectionality could enrich the Court's assessments of discrimination*.

[43] further comments that "*At the doctrinal level, some encouraging signs can be read in the Court's jurisprudence despite the failures highlighted above. In Parris, the Court at least signalled that the litigation of intersectional discrimination is not precluded and that claims invoking several grounds of discrimination simultaneously can be reviewed. In addition, even though not followed by the Court, the opinion rendered by AG Kokott in Parris shows awareness of, and willingness to address, intersectional discrimination. The AG opinion warns that 'the*

[24] https://curia.europa.eu/juris/liste.jsf?language=en&num=C-443/15

[25] https://www.sackers.com/pension/parris-v-trinity-college-dublin-cjeu-24-november-2016/

Court's judgment will reflect real life only if it duly analyses the combination of those two factors, rather than considering each of the factors of age and sexual orientation in isolation'. Further, it acknowledges that 'the combination of two or more different grounds [. . .] is a feature which lends a new dimension to a case'. It also confirms that an appropriate assessment of such a case of discrimination should take the synergy of these different axes of discrimination into account as opposed to splitting the analysis based on each ground in separation."

Another examplary case is the one of Odar v Baxter Deutschland GmbH of 2012[26]. There, intersectionality was considered with respect to age and disability. In particular, the plaintiff had to retire early due to severe disabilities and on the same time, a special formula was on effect that reduced compensation on an increasing basis as workers got closer and closer to state pension age[27]. This resulted to the plaintiff receiving a significantly reduced redundancy compensation. The European Court of Justice decided that the discrimination regarding age was not unlawful, while discrimination regarding disability could not be justified. Thus, while, in this case, the two sensitive attributes were examined in isolation, [12] remarks that "*the Court has implicitly recognised how disadvantage arises from the interaction of age and disability based discrimination, acknowledging 'the risk that severely disabled persons may have financial requirements arising from their disability which cannot be adjusted and/or that, with advancing age, those financial requirements may increase*".

The case of Leger vs Ministre des Affaires sociales of 2015[28], although in another context (blood donation), provides similar findings, where discrimination is considered in the intersection of sex and sexual orientation. In this case, blood donation was refused by a doctor to a man with same-sex sexual relations. The case examined whether a permanent ban on blood donation for men having same-sex sexual relations comprised discrimination on sexual orientation, finding that this ban is justifiable under specific circumstances [29]. On this decision, [43] notes that the opinion of the Advocate General in the case "*acknowledges the existence of 'clear indirect discrimination consisting of a combination of different treatment on grounds of sex — since the criterion in question relates only to men — and sexual orientation — since the criterion in question relates almost exclusively to homosexual and bisexual men'. These traces of doctrinal sensitivity for the problem of intersectional discrimination thus open potential legal pathways for a better handling of algorithmic discrimination in its intersectional manifestations*".

The failure of the current EU legal framework to properly handle intersectional discrimination is further highlighted in the case of Z. v A Government Department and the Board of Management of a Community School of 2014[30]. There, discrimination on the intersection of gender and disability is examined: a woman was denied maternity leave, after giving birth to a baby via surrogacy, with the justification that she had never been pregnant and the baby had not been adopted either. [43] notes that "*Instead of examining how the inherently gendered form of disability at stake in this case resulted in discriminatory effects – depriving a mother from social protection –, the Court resorted to a formalistic comparison test, separating the question of discrimination on grounds of sex from that of discrimination on grounds of disability.*

[26] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62011CJ0152

[27] https://www.agediscrimination.info/case-reports/2012/12/14/odar-v-baxter-deutschland-gmbh

[28] https://curia.europa.eu/juris/liste.jsf?num=C-528/13

[29] https://europeanlawblog.eu/2015/08/26/gay-blood-bad-blood-a-brief-analysis-of-the-leger-case-201

[30] https://curia.europa.eu/juris/liste.jsf?language=en&num=C-363/12

This reasoning obfuscated the disadvantage produced by the interaction between ableism and a strictly biological understanding of motherhood as ensuing from pregnancy. As a result of this failure to consider intersectional discrimination as 'greater than the sum of' its parts, the Court found no discrimination"

4.1.2 Proxy variables and correlations

Another important issue in algorithmic fairness, that also arises often in EU law cases is that of **indirect discrimination** via **proxy variables** or correlations. Next, we enumerate a set of prominent cases that demonstrate how EU law has handled such cases.

The first case regards proxy discrimination on the gender attribute, via the pregnancy attribute: Elisabeth Johanna Pacifica Dekker v Stichting Vormingscentrum voor Jong Volwassenen (VJV-Centrum) Plus, of 1990. There, an employer did not hire the best candidate with the justification that she was pregnant, something that would cause losses to the employer. The European Court of Justice decided in favor of the candidate (plaintiff) judging that the employer's decision was in breach of the Equal Treatment Directive. [43] remarks that this is a representative case where discrimination on the proxy variable (pregnancy) has been recognized as direct sex discrimination. [12] further discusses that "*Beyond the identification of protected grounds in algorithmic decision making, the correlation and proxy challenge is relevant in relation to the definition of the protected grounds. It raises the question of how narrowly or widely the protected grounds should be defined. If protected grounds are given an extensive interpretation, relevant proxies could also be covered by their meaning. The classic example is pregnancyrelated discrimination, which is so clearly and closely related to sex-based discrimination that 'pregnancy' is generally regarded as a proxy for 'being a woman'. In fact, the Court of Justice of the EU has treated pregnancy-related discrimination as direct sex discrimination.222 Another classic example is holding a foreign passport, which is a clear proxy for 'being of a different nationality'. Usually there is an almost 100 % overlap here between the 'actual' protected ground and its proxies, meaning that the use of the proxy covers almost exactly the same group of persons as using the actual ground would do. Similarly, when there is a close connection between individual preferences and affinities, and protected grounds, belonging to a group with a certain 'affinity' (e.g. having an interest in particular religious matters) might be nearly the same as belonging to a group characterised by a particular personal trait (e.g. adhering to a certain religion)*".

4.2 EU use cases relevant to algorithmic fairness

In this subsection, we present some prominent examples of deployment of AI systems for decision making in European countries and discuss some insights, derived from them, regarding AI-fairness policies and best practices.

4.2.1 Equal treatment vs equal outcome - disparate treatment vs disparate impact - affirmative action

Starting from 2021, Austrian Labor Market Service (AMS) deployed the Labor Market Opportunities Assistance System (AMAS), which analyses statistics from previous years to classify new job seeker into three categories, affecting the training resources that will be dedicated for each individual. The features considered by the system include: age, group of countries, gender, education, care responsibilities and health impairment as well as past employment, contacts with the AMS and the labor market situation in the place of residence[31] [12].

The system has received critic focusing on the lack of transparency in the training data selection process, the absence of procedures for bias auditing and correction and the decision to prioritize training effectiveness and accuracy as its sole objective. A particular point of criticism from the Institute for Technology Assessment of the Austrian Academy of Sciences[32] focuses on the fact that *the system does not offer any clues to prevent possible structural unequal treatment*. This, directly relates to the discussion on whether an **equal opportunity or equal outcome** should be adopted and to which extent, by an AI system that helps decision making in the field of **employment/hiring/labour market**.

4.2.2 Proxy variables and correlations - indirect discrimination

The aforementioned Labor Market Opportunities Assistance System (AMAS) for classifying job seekers has also been found to express indirect discrimination, since the Austrian public employment service stated that the system "*has shown that caring responsibilities affect women's labour market opportunities but not men's*". This comprises a case of discrimination by **proxy**, since *caring responsibilities*, through which bias is expressed, serves as a proxy for gender.

"My employability" system used by the Croatian Employment Service (CES)[33] has also been found to potentially discrimination against women, based on a third feature. In particular, it has been reported [12] that specific values of the feature users' children have negative impact on the system's decisions for women, whereas no such impact is observed for men, revealing another potential case of discrimination by **proxy**.

4.2.3 Feedback loops

VDAB[34] is an AI system developed in Belgium to assist residents of Flanders (a region in Belgium) in finding jobs. An issue was raised [12] on whether VDAB should take into account

[31] https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus
[32] https://www.oeaw.ac.at/ita/
[33] https://stapweb.hzz.hr/
[34] https://www.vdab.be/

the apparent preference of women candidates for temporary jobs in its recommendations. Doing so, there might be a risk of helping to perpetuate a historical bias by facilitating a *feedback loop*: the system consistently recommends more temporary jobs to women, resulting in women more often engaging in temporary jobs.

In another example, personalized advertising in Italy is reported to perpetuating ethnicity and social bias, by discriminating on its advertisements to the respective subgrous and thus creating new ghettos and luxury districts [12][35].

4.2.4 Transparency and explainability

SyRI comprised a social security fraud detection system deployed in the Netherlands which was eventually banned due to issues of **transparency** and verifiability[36]. Further, it has been reported that it has been primarily used in low-income neighborhoods, exacerbating bias towards demographics living in these areas (**feedback loops**) and introducing/increasing **proxy** bias via the location attribute.

In another example, the Parcoursup system[37] deployed in France for assisting the distribution of candidate students to French universities, was also criticized for its lack of **transparency**. Further concerns of **proxy** discrimination were expressed due to the system using income and residency attributes for its decision making.

[35] https://www.agendadigitale.eu/sicurezza/privacy/algoritmi-che-discriminano-ecco-perche-serve-lapp
[36] https://uitspraken.rechtspraak.nl/#!/details?id=ECLI:NL:RBDHA:2020:1878
[37] https://www.parcoursup.fr/

4.3 Ethical AI: policies, frameworks & tools

4.3.1 Available policies and accountability tools for algorithms & AI applications

This section describes a number of different policy mechanisms through which governments try to achieve algorithmic accountability in the public sector. As a relatively recent addition to technological governance, these policies vary widely, as does the vocabulary used to describe them[38]. The following typology indicates the forms of algorithmic accountability policies that are currently taking shape in the public sector.

Principles and guidelines. Some policy documents provide non-binding normative guidance, in the form of principles and values, that public services should follow. These documents vary in form, but generally identify high-level policy objectives and how these may be implicated by the use of algorithmic systems within public organisations. In some cases, such as in the UK's data ethics framework[39] or the Australian Ombudsman's good practice guide to automated decision making[40], these guidelines also offer implementation guidance. These guidelines provide normative standards against which agencies, and in some cases the public, can assess their own practices.

Bans and moratoria. Some jurisdictions have banned or prohibit the use of certain types of "high-risk" algorithmic systems. In some cases, such as Morocco's policy on facial recognition, the bans are framed as temporary moratoria, intended to end once appropriate safeguards and accountability mechanisms are designed and implemented[41] . Bans and moratoria have been significantly applied to facial recognition technologies used by law enforcement authorities, and in some cases by local governments in the US, such as Portland[42], Oakland[43] and San Francisco[44]

Public transparency. Transparency mechanisms provide information about algorithmic systems to the general public (e.g. to affected persons, the media or civil society), so that

[38]This section summarizes research results from the Ada Lovelace Institute, AI Now Institute and Open Government Partnership. (2021) *Algorithmic Accountability for the Public Sector*

[39]Department for Digital, Culture, Media and Sport (2021). *Data Ethics Framework*. UK Government. `www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework`

[40]Commonwealth Ombudsman (2019). *Automated Decision-Making Better Practice Guide*. Government of Australia. https://www.ombudsman.gov.au/publications/better-practice-guides/auto- mated-decision-guide.

[41]National Control Commission for the Protection of Personal Data, Morocco. (2020) 'Press release of 30/03/2020: Press release accompanying the publication of deliberation No. D-97-2020 du 26/03/2020'. (In French) `https://www.cndp.ma/fr/presse-et-media/communique-de-presse/661-communique-de-presse-du-30-03-2020.html`

[42]City of Portland, Oregon. (2020). *city council approves ordinances banning use of face recognition technologies by City of Portland bureaus and by private entities in public spaces*, https://www. portland.gov/smart-city-pdx/news/2020/9/9/city-council-approves-ordinances-banning-use-face-recognition

[43]City of Oakland, California. Chapter 9.64 - Regulations on city's acquisition and use of surveillance technology, Title 9 - Public Peace, Morals and Welfare, *Oakland, California Code of Ordinances*.: https://library.municode.com/ca/oakland/codes/code_of_ordinances?nodeId=TIT9PUPEMOWE_CH9.64REACUSSUTE

[44]City and County of San Francisco. (2019). chapter 19B: Acquisition of Surveillance Technology, *San Francisco Administrative Code* https://codelibrary.amlegal.com/codes/san_francisco/lat- est/sf_admin/0-0-0-47320

individuals or groups can learn that these systems are being used and demand answers and justifications for such use. Examples of such transparency efforts include:

1. Public registries of algorithmic systems in Ontario[45] , Amsterdam[46], Helsinki[47], and cities in France, such as Antibes, Lyon and Nantes[48] , which address civil society and citizens.

2. Source code transparency requirements that apply to computational algorithmic systems and have been implemented under the Canadian Automated Decision Making (ADM) Directive[49] .

3. Explanations of the algorithmic logic (which are supposed to allow the public and policy makers to understand how an algorithmic decision was made). This is a legal requirement under French law in the Digital Democracy Bill[50].

Impact assessments. Impact assessments include a wide range of accountability mechanisms that have been applied in scientific and policy areas with a wide range of scope, such as environmental protection, human rights, data protection and privacy. The aim is to mitigate the harmful impacts of a particular initiative or development, identify risks and address them before implementation.

Algorithmic Impact Assessments (AIAs) are mechanisms intended for public services to better understand, categorize and respond to potential harms or risks arising from the use of algorithmic systems, usually before they are used. AIAs vary considerably, but were originally created as a way to allow stakeholders to define and construct a matrix of harms, benefits and risks in order to assess in advance whether the use of an algorithmic system is appropriate in a particular context.

It is assumed that AIAs provide affected communities in particular with greater involvement in the uses of algorithmic systems by public services and influence on how they respond to potential failures[51] . In practice, however, most AIAs currently in use have not meaningfully engaged these communities and have been primarily implemented for internal self-assessment by public services, making them often unavailable to the public. In some cases, for example, under the Canadian Guidance on Automated Decision Making[52] or the New Zealand Algorithmic

[45]Ontario. *data catalogue*. available at: `https://data.ontario.ca/group/artificial-intelligence-and-algorithms`

[46]City of Amsterdam Algorithm Register Beta: *What is the Algorithm Register?:* `https://algoritmeregis-ter.amsterdam.nl/en/ai-register/`

[47]City of Helsinki AI Register. *What is AI Register?* Available at: `https://ai.hel.fi/en/ai-register/`

[48]Pénicaud, S. (2021). 'Building Public Algorithm Registers: lessons learned from the French approach'. *Open Government Partnership Blog*. 12 May Available at: `https://www.opengovpartnership.org/stories/building-pub-lic-algorithm-registers-lessons-learned-from-the-french-approach/`.

[49]Treasury Board of Canada Secretariat, Government of Canada. (2019). *Directive on Automated Decision-Making* https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

[50]Republique Francaise. (2016) *The Digital Republic bill - Overview.* https://www.republique-numerique.fr/pages/in-english

[51]Metcalf, Jacob, et al. (2021). 'Algorithmic impact assessments and accountability: the co-construction of impacts.' *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* `https://dl.acm.org/doi/abs/10.1145/3442188.3445935`

[52]Treasury Board of Canada Secretariat, Government of Canada. (2019). *Directive on Automated Decision-Making*. Available at: `https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592`

Charter[53] , the results of AIAs determine the final level of regulatory control applied to specific algorithmic systems.

Audits and regulatory inspection. Audits refer to a set of mechanisms designed to provide insight into the operation and potential impact of an algorithmic system.

1. *Technical audit*: a narrowly targeted test of a specific hypothesis about a system by examining its inputs and outputs - for example, whether it exhibits racial bias in the outcomes of a decision.

2. *Regulatory audit:* in this context, audit refers to a regulatory inspection and compliance exercise, such as financial audit. Increasingly, regulatory audits are also designed to capture the broader societal consequences of a system's use and to assess its operation against an established regulatory standard in order to identify potential areas of concern[54].

Technical audits and regulatory inspections may vary in their scope and application[55], but in general they are based on the assumption that inspections help to provide an independent account of how algorithmic systems operate, and to account for any deficiencies, biases or errors in the system.

While audits are an important mechanism for public sector accountability and, in combination with other approaches, hold great promise for algorithmic systems, they have not been formalised as standard policy mechanisms for the use of algorithmic systems by the public sector. To date, they remain largely ad-hoc exercises conducted within the broader scope of specific regulatory or administrative bodies, including statutory auditors in Sweden[56], in the Netherlands[57] and in France[58] . The UK Information Commissioner's Office has also encouraged internal regulatory auditing by organisations using AI, including both compliance audits and technical 'bias' audits, in its draft Guidance on the AI Auditing Framework[59] .

[53]New Zealand Govt. (2020). *algorithm charter for Aotearoa New Zealand*. `https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf`.

[54]Ada Lovelace Institute and DataKind UK. (2020). *Examining the Black Box: Tools for Assess- ing Algorithmic Systems. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems*

[55]Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK. (2020). *auditing machine learning algorithms: a white paper for public auditors*. Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK. available at: `https://www.auditingalgorithms.net/`

[56]Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK. (2020). *auditing machine learning algorithms: a white paper for public auditors*. Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK `https://www.auditingalgorithms.net/`

[57]Swedish National Audit Office. (2020). *Automated Decision-Making in Public Administration* `https://www.riksrevisionen.se/en/audit-reports/audit-reports/2020/automated-decision-making-in-public-administration-\/-\/-effective-and-efficient-but-inadequate-control-and-follow-up.html`.

[58]Netherlands Court of Audit. (2021). *understanding algorithms*. available at: `https://english.rekenkamer.nl/publications/reports/2021/01/26/understanding-algorithms`

[59]UK Information Commissioner's Office. (2020). *draft guidance on the AI Auditing Framework.* `https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf`.

External/independent oversight bodies. Independent oversight mechanisms aim to ensure accountability by monitoring the actions of public bodies and making recommendations, sanctions or decisions on their use of algorithmic systems. Oversight mechanisms vary widely in form and function. Some mechanisms are based on legislative oversight, such as the legislation on community control of police surveillance in the US.[60] Others, such as the West Midlands Police Data Ethics Committee, UK, operate in an advisory capacity without being specifically delegated legal powers[61] responsibilities.

Rights to be heard and rights of appeal. Some policies require that decisions taken using algorithmic systems adhere to certain procedures as a means of ensuring fairness and providing a forum for individual recourse in the event of a biased or incorrect decision. These procedures, which include notification of the decision, the provision of a hearing, the opportunity to present evidence, and/or the right to appeal the decision to a neutral forum, are intended to provide a forum for affected individuals or groups to discuss or challenge specific decisions that affect them. The best known of these are the notice, hearing and right to explanation requirements of automated decisions provided by the GDPR in the EU[62].

Conditions for the conclusion of contracts. The conditions of public procurement have been an important area of intervention for transparency and accountability. Some policies attempt to translate these general rules of transparency and accountability into algorithmic systems. When governments acquire algorithmic systems from private vendors, specific procurement conditions may apply that restrict the design and development of an algorithmic system (e.g. to ensure that a system being considered for procurement is transparent and non-discriminatory).

These contractual preconditions are intended to ensure that governments acquire only systems that comply with transparency, fairness or other requirements and that, should a vendor fail to comply, the vendor is subject to contractual liability. Procurement conditions have been established as policy mechanisms by the City of Amsterdam in the Netherlands[63], have been promoted by the UK government through its AI procurement guidelines[64] and by the Tamil Nadu state government in India in its policy on safe and ethical use of AI[65]

An example of such terms are the following articles from the "Standard Clauses for Municipalities for Fair Use of Algorithmic Systems" of the Municipality of Amsterdam:

Article 5. Transparency on Algorithmic Application (*abridged*)

[60]ACLU. (2021). community control over police surveillance (CCOPS) model bill. Available at: `https://www.aclu.org/legal-document/community-control-over-police-surveillance-ccops-model-bil`

[61]West Midlands Police and Crime Commissioner. (2021). *ethics committee*, `https://www.westmidlands-pcc.gov.uk/ethics-committee/`

[62]Article 29 Working Party. (2018). Guidelines on Automated individual decision-making and Profiling for thpurposes of Regulation 2016/679. `https://ec.europa.eu/newsroom/article29/items/612053/en`

[63]Municipality Amsterdam. (2020). *Standard Clauses for Municipalities for Fair Use of Algorithmic Systems.* `https://www.amsterdam.nl/innovatie/`

[64]UK Government. (2020). *guidelines for AI Procurement.* `https://www.gov.uk/government/publications/guidelines-for-ai-procurement`

[65]Government of Tamil Nadu. (2020). *Safe and Ethical AI Policy*. Available at: `https://elcot.in/sites/default/files/AIPolicy2020.pdf`

5.1 The Contractor shall at the first request of the Municipality to provide the Municipality with Procedural Transparency. The Municipality has the right to share and disclose the information provided by the Contractor in this context with third parties. At the request of the Municipality the Contractor shall complete a register for Algorithmic Applications to be designated by the Municipality.

5.2 The Contractor shall, at the first request of the Municipality, provide the Municipality with Technical Transparency to enable the Municipality to conduct an audit as referred to in Article 8. The Municipality shall only request and use such information if and to the extent necessary for the application of Article 8.

5.3 When applying article 5.2, the Contractor may choose not to surrender the source code of the Algorithmic application to the Municipality, but to an independent third party to be designated and engaged by the Municipality, which will conduct the audit referred to in Article 8 on behalf of the Municipality. Any additional costs resulting from this shall be borne by the Contractor. The Municipality may require the Contractor to make an advance payment in connection with the independent third party's costs.

5.4 The Municipality must at all times have the opportunity to explain the operation of the Algorithmic Application (Explainability). The Contractor is obliged to cooperate fully in making the Algorithmic Application Explainable and to provide the Municipality with all the information required for this purpose. The Municipality has the right to share and disclose the information provided by the Contractor in that context with third parties.

Article 8. Audit or other type of control

8.1 The Contractor shall at all times be obliged to cooperate in an audit carried out by or on by or on behalf of the Municipality for an audit or other type of audit carried out by or on behalf of the Municipality in which an assessment is made of the Contractor's compliance with the conditions stipulated in the Agreement. Such cooperation shall include providing Technical Transparency, providing insight into the risk management strategy implemented, making Contractor personnel available to conduct interviews and providing access to Contractor's sites.

8.2 The Municipality shall prepare (or cause to be prepared) a report recording the conclusions of the audit. In the report, the Municipality shall record the extent to which the Contractor complies with the obligations under the Agreement. If the Municipality establishes that the Contracted Party does not comply with the obligations under this article, the Contractor will be obliged to comply within the the reasonable period set by the Municipality in the report to remedy the defects identified by the Municipality. If the Contractor does not remedy the defects identified by the Municipality within the remedy period specified in the report, the Contractor shall be in default by operation of law.

8.3 The Municipality shall have the right to publish the conclusions of the report referred to in Article 8.2. In the event of a conflict between Article 5.2 and Article 8.3, Article 8.3 shall take precedence.

8.4 The Municipality has the right to carry out an audit (or have one carried out) no more than once per calendar year.

8.5 The Municipality may decide to have the audit conducted (in part) by an independent auditor.

8.6 The costs of any auditor to be engaged by the Municipality shall be borne by the Municipality. Any costs incurred by the Contractor will incur in the context of the audit for work other than providing Technical Transparency or Procedural Transparency, the Municipality shall pay the Contractor a reasonable fee. A dispute over the amount of such fee shall never be grounds for the Contractor to suspend its obligations under these terms and conditions. Such fee need not be paid by the Municipality if the audit reveals that the Contractor is not, or has not, complied with these Terms and Conditions in material respects.

4.3.2 Overview of existing frameworks for assessing the ethical implications of AI applications use

Tools and policies such as those presented in the previous section are organised according to the scope of the AI and the requirements in evaluation frameworks.

The use of AI applications raises ethical concerns about privacy, transparency, bias and accountability. For example, AI applications may collect personal data from users without their consent or knowledge, which could be used for targeted advertising or other purposes. They may also perpetuate harmful biases or discriminatory practices if not properly designed and trained. In addition, AI applications may lack transparency and accountability, making it difficult for users to understand how they operate and who is responsible for their actions.

Given the potential risks and benefits of AI applications, it is important to conduct ethical impact assessments to identify and mitigate potential harms and ensure that benefits are fairly distributed. This includes developing ethical frameworks and guidelines for the design, development and deployment of AI AI applications, as well as assessing their impact on various stakeholders.

Unesco recommendation on the ethics of artificial intelligence. The United Nations Educational, Scientific and Cultural Organization (UNESCO) has issued a recommendation[66] on the ethics of artificial intelligence (AI), which provides guidelines for the development of AI that is transparent, inclusive and respectful of human rights.

The UNESCO recommendation underlines the need for AI to be designed and used in a manner consistent with human dignity, human rights, cultural diversity and non-discrimination. It calls for the protection of privacy, personal data and intellectual property rights, as well as the promotion of transparency and accountability in the development and use of AI.

The Recommendation also underlines the importance of ensuring that AI is developed and used in a way that benefits all individuals and society as a whole and that it does not exacerbate existing inequalities or lead to new forms of discrimination.

UNESCO recommends that Member States establish national and international frameworks to promote the ethical development and use of AI, including the establishment of multi-

[66] `https://unesdoc.unesco.org/ark:/48223/pf0000380455`

stakeholder advisory committees to provide guidance and oversight on the development and deployment of AI.

Asilomar AI Principles. The Asilomar AI Principles[67] are a set of guidelines for the development and use of artificial intelligence (AI) systems. They were created by a group of leading AI researchers and industry experts at the Asilomar Conference on Beneficial AI in January 2017.

The principles consist of 23 points, grouped into three categories: research issues, ethics and values, and long-term issues. Some of the key principles include:

- *Research questions:* The authorities call for AI research to focus on creating AI systems that are robust and verifiable to ensure that they are safe and reliable. They also stress the importance of transparency in AI research so that researchers and others can understand how AI systems work and what their limitations are.
- *Ethics and values:* The principles emphasise the importance of developing AI systems that are aligned with human values and do not harm individuals or society as a whole. They call for AI to be used to enhance rather than replace human capabilities and for AI systems to respect the privacy and dignity of individuals.
- *Long-term issues:* The authorities recognise that AI has the potential to transform society in important ways and call for careful consideration of the social implications of AI. They call for the development of AI systems that are sustainable and equitable and that promote human well-being.

The Asilomar Principles for AI represent a consensus view among leading AI experts on the key issues that need to be addressed to ensure that AI is developed and used in a way that is beneficial to humanity.

The Montreal Declaration on Responsible AI. This declaration was developed by a group of AI experts and includes a set of principles for the responsible development and use of AI, including AI applications[68]. The declaration consists of ten principles, which are as follows:

- *Prosperity*: The development and use of AI should prioritise the well-being of all sentient beings.
- *Respect for autonomy:* AI should respect and preserve human autonomy, agency and decision-making.
- *Protection of privacy:* The development and use of AI should protect privacy and promote confidentiality and security of personal data.
- *Solidarity*: AI should be developed and used for the common good and benefit of all and should not be used to exacerbate existing inequalities.

[67] `https://futureoflife.org/open-letter/ai-principles/`

[68] The Montreal Declaration for Responsible AI: `https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/`

- *Democratic participation:* AI systems should be designed and used to enhance democratic participation, not to undermine it.
- *Social and environmental responsibility:* The development and use of AI should be guided by a sense of social and environmental responsibility.
- *People-centred values:* AI should be designed and used in ways that reflect human-centred values such as fairness, inclusiveness and transparency.
- *Responsible innovation:* The development and use of AI should be guided by principles of responsible innovation, including transparency, inclusiveness and accountability.
- *Explicability and interpretability:* Artificial intelligence systems should be designed and used in ways that allow for explanatory and interpretability in order to enhance trust and accountability.
- *Accountability:* Those responsible for the development and deployment of AI systems should be accountable for their impact on society and the environment.

The Fairness, Accountability, and Transparency framework. It is a set of guidelines for the development and evaluation of AI systems to ensure that they are fair, transparent and accountable. The framework was developed in response to growing concern about the potential negative impacts of AI on society, such as bias and discrimination, lack of transparency and lack of accountability.

Fairness refers to the equal treatment of individuals or groups, regardless of their race, gender, age or other characteristics. The FAT Framework emphasises the need to ensure that AI systems are designed and implemented in a way that does not perpetuate or exacerbate existing prejudice or discrimination. This includes considering the data used to train the system, the algorithms used to make decisions and the impact of the system on different groups of people.

Accountability refers to the ability to hold an AI system accountable for its actions and decisions. The FAT Framework emphasizes the need to ensure that AI systems are designed and implemented in a way that allows for traceability and auditability. This includes documenting the data used to train the system, the algorithms used to make decisions, and the decisions made by the system.

Transparency refers to the ability to understand the internal workings of an AI system. The FAT Framework emphasises the need to ensure that AI systems are designed and implemented in a way that allows for transparency and explanatory power. This includes providing clear documentation of the data used to train the system, the algorithms used to make decisions and the decisions made by the system.

Overall, the FAT framework provides a set of principles and best practices for developing and evaluating AI systems that are fair, transparent and accountable. By following these guidelines, developers can help ensure that their AI systems are designed and implemented in a way that benefits society as a whole.

DEON. In addition to the existence of general frameworks such as FAT, there are frameworks that have been specifically investigated and compared for use in the development and operation of interactive AI applications[69] .

The DEON checklist[70] (Duty Ethics and Online Networks) is a framework for assessing the ethical implications of a particular technology or system, with a focus on artificial intelligence and machine learning. The checklist was developed to provide a structured way of analysing the ethical issues that arise in the design, development, deployment and use of AI systems. The name "DEON" is an acronym for the four categories of ethical concerns addressed by the checklist:

- **Duty:** This category addresses the responsibilities and duties of those involved in the development and deployment of AI systems. It includes questions such as:
  - What are the ethical principles and values that should guide the development and deployment of this AI system?
  - Who is responsible for ensuring that the system adheres to these principles and values?
  - What measures can be taken to ensure that the system does not cause harm to individuals or society as a whole?
- **Ethical principles:** This category includes an assessment of the ethical principles and values associated with the AI system under development. It includes questions such as:
  - Does the system promote the values of autonomy, privacy and justice?
  - Is the system transparent and accountable in its decision-making processes?
  - Does the system take into account the needs and interests of all stakeholders, including marginalised groups?
- **Results:** This category refers to the possible outcomes or consequences of the AI system, both intended and unintended. It includes questions such as:
  - What are the potential risks and benefits of the system?
  - Will the system lead to greater social inequality or will it reinforce existing prejudices?
  - How will the system affect the privacy and autonomy of individuals?
- **Rules:** This category includes the evaluation of social norms and values related to the development and deployment of the AI system. It includes questions such as:
  - What are the social norms and expectations around the use of this AI system?
  - How can the system be perceived by different stakeholders and how can these perceptions affect its effectiveness and use?
  - What impact could the system have on existing social norms and values?

[69] Atkins, S., Badrie, I., & Otterloo, S.V. (2021). Applying Ethical AI Frameworks in practice: evaluating conversational AI AI applications solutions. *Computers and Society Research Journal*.

[70] DEON. an ethics checklist for data scientists, 2018. `https://deon.drivendata.org/#default-checklist`, accessed on 2021/09/15.

Assessment List for Trustworthy Artificial Intelligence (ALTAI). The Assessment List for Trustworthy Artificial Intelligence (ALTAI[71]) is a framework developed by the European Commission's High Level Expert Group on Artificial Intelligence (AI HLEG) to assess the trustworthiness of AI systems. It provides a set of basic requirements that AI systems must meet in order to be considered trustworthy. The ALTAI framework is based on the following four ethical principles: respect for human autonomy, prevention of harm, fairness and explanatory power.

The ALTAI framework consists of three main parts:

- **A list of requirements:** the ALTAI framework provides a list of requirements that AI systems must meet in order to be considered trustworthy. These requirements are grouped into the following categories:
  - *Human representation and supervision:* AI systems should support human decision making and control and be designed in a way that allows humans to understand and manage their behaviour.
  - *Technical robustness and safety:* AI systems should be designed to be robust, safe and reliable and to minimise the risk of unintended failure.
  - *Privacy and data governance:* AI systems should respect privacy and data protection rights and be designed in a way that allows individuals to exercise control over their personal data.
  - *Transparency:* AI systems should be transparent and explainable and decision-making processes should be open to scrutiny.
  - *Diversity, non-discrimination and fairness:* AI systems should be designed in a way that ensures fairness, non-discrimination and diversity.
  - *Social and environmental well-being:* AI systems should be designed to have a positive impact on society and the environment.
- **A set of evaluation methods:** the ALTAI framework provides a set of evaluation methods that can be used to assess the reliability of AI systems. These include:
  - *Technical tests:* This is the testing of the technical robustness, safety and reliability of the AI system.
  - *Human assessment:* Includes the evaluation of the impact of the AI system on human autonomy, decision-making and control.
  - *Ethical evaluation:* It involves assessing the compliance of the AI system with ethical principles such as fairness, non-discrimination and privacy.
  - *Social impact assessment:* includes the assessment of the impact of the AI system on society and the environment.

[71] European Commission: High-Level Expert Groupon Artificial Intelligence.Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment: Shaping Europe's digital future., 2020. `https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment`.

- **A set of practical guidance and tools:** the ALTAI framework provides practical guidance and tools to help developers and users of AI systems to implement the requirements and assessment methods. This includes a set of best practices, technical guidelines and tools for assessing and mitigating risks associated with AI systems.

  The ALTAI framework is intended to be a living document that will be updated and improved over time as AI technologies evolve and new challenges arise. It is designed to be flexible and adaptable to different contexts and applications and to be used by a wide range of stakeholders, including developers, users, regulators and policy makers.

Microsoft Conversational AI Guidelines.[72] Microsoft's guidelines for interactive AI provide a set of recommendations for creating AI applications and other interactive AI systems that are ethical, responsible and inclusive. The guidelines were developed in collaboration with Microsoft's AI experts and are based on the principles of fairness, trustworthiness, privacy and inclusiveness. The guidelines are organised around six key principles:

- **Design for privacy and security:** conversational AI systems should be designed with privacy and security in mind from the outset. This includes using secure communication protocols, encrypting data at rest and in transit, and minimising the amount of personal data collected and stored.

- **Transparency:** Users should be able to understand how the system works, what data is collected, and how this data is used. Conversational AI systems should provide clear and concise explanations of their functionality and limitations.

- **Reliability and safety:** they should be reliable and safe to use. This includes testing the system for accuracy and consistency and ensuring that it is not biased or prejudiced.

- **Promoting inclusiveness:** Interactive AI systems should be designed to be inclusive, accessible and easy to use for all users. This includes designing for the different needs of users, including people with disabilities or who may speak different languages.

- **Respect for user preferences:** Systems should respect users' preferences in terms of privacy, data exchange and communication style. Users should be able to control their interactions with the system and choose the level of engagement with which they feel comfortable.

- **Be responsible:** Interactive AI systems should be designed with accountability in mind. This includes providing clear channels for feedback and complaint resolution and ensuring that the system can be audited and monitored for compliance with ethical and legal standards.

[72] Lili Cheng. The Official Microsoft Blog:Microsoft introduces guidelines for developing responsible conversational AI, 2018. `https://blogs.microsoft.com/blog/2018/11/14/microsoft-introduces-guidelines-for-developing-responsible-conversational-ai/`, accessed on 2021/09/15.

Artificial Intelligence Impact Assessment (AIIA). Artificial Intelligence Impact Assessment (AIIA) is a framework designed to help organisations identify and address the potential risks and benefits of implementing AI systems. The AIIA is intended to be used as a tool to assess the ethical, social and economic impact of AI in a given context. The AIIA framework consists of four main steps:

- **Determining the scale:** The first step involves defining the scope and objectives of the AI system being evaluated. This includes identifying stakeholders and potential impacts, as well as any relevant legal or ethical considerations.
- **Impact assessment:** the second step involves assessing the potential impact of the AI system on various stakeholders, such as customers, employees and society as a whole. This includes assessing the risks and benefits of the AI system, as well as any unintended consequences.
- **Mitigation:** The third step involves identifying and implementing measures to mitigate any negative impacts of the AI system. This may include adapting the design of the AI system, implementing new policies and procedures or providing additional training and support to affected stakeholders.
- **Monitoring and review:** the final step involves monitoring the AI system over time to ensure that it continues to work as intended and that any negative impacts are addressed. This may include ongoing data collection and analysis, as well as periodic reviews and updates of the AI framework itself.

The AIIA framework is intended to be flexible and adaptable to different AI contexts and applications. It is designed to be used by a wide range of stakeholders, including policy makers, industry leaders and civil society organisations.

The above ethical frameworks and guidelines provide valuable guidance for the development and use of AI applications and can help ensure that these technologies are developed in a responsible and ethical manner. However, it is important to note that these frameworks and guidelines are not prescriptive and that ethical considerations regarding AI applications will continue to evolve as the technology evolves. Therefore, ongoing ethical impact assessments and stakeholder engagement will be necessary to ensure that AI applications continue to be developed and used in an ethical and responsible manner.

The European Commission proposal. Following an initial approach focusing on general ethical principles and the White Paper on AI[73], in April 2021 the European Commission proposed an EU Regulation on AI[74]. This proposal introduces two new elements: a move away from more uncertain ethical grounds towards the adoption of "hard" legal rules and the adoption of a Regulation in the absence of national AI laws or different approaches between EU Member States.

[73] European Commission (2020d) White Paper on Artificial Intelligence - A European Approach to Excellence and Trust, COM(2020) 65 final. `https://ec.europa.eu/info/files/white-paper-` artificial-intelligence-european-approach-excellence-and-trust_en.

[74] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

The latter aspect underlines the EU legislator's concerns about the rapid growth of AI, the limited competitive strength of the EU in this area in terms of market share and the need to address the growing public concerns about AI, which could hinder its development.

As in the case of data protection, the EU proposal is therefore framed within the interests of the internal market, while protecting fundamental rights.This focus on the market and competition seems to be the main rationale behind the regulation of another unregulated area, aimed at encouraging AI investment in the EU.

It also clearly follows from the four objectives of the proposed Regulation: i) to ensure that AI systems placed on the market and used in the Union are safe and respect the existing legislation on fundamental rights and values of the Union; ii) to ensure legal certainty to facilitate investment and innovation in the field of AI; iii) to strengthen the governance and effective enforcement of the existing legislation on fundamental rights and security requirements applicable to AI systems; and iv) to ensure that AI systems are safe and secure.

The European Commission's proportional, risk-based approach identifies four levels of risk: i) extremer-risk applications, which are banned; ii) high-risk applications, subject to compliance assessment; iii) a limited number of applications that have significant potential for manipulation of individuals, obligated to comply with specific transparency obligations; iv) non-high-risk uses, which are addressed by codes of conduct designed to promote compliance with key requirements of the proposed Regulation; v) applications that are not high risk, which are addressed by codes of conduct designed to promote compliance with key requirements of the proposed Regulation. Of these provisions, the most important in terms of human rights impact assessment are the provisions for high-risk applications.

The AI proposal defines high-risk AI systems as those that pose significant risks to the health, safety or fundamental rights of individuals. This includes AI applications in areas such as healthcare, transport and customer service, where AI applications are commonly used. For such high-risk AI applications, the law requires developers to carry out a thorough risk assessment and implement technical and organisational measures to mitigate the identified risks. This includes ensuring that the AI system is transparent, explainable and subject to human oversight.

In addition, the proposal requires that AI systems are developed and used in accordance with ethical principles, such as those described in the European Commission's High Level Expert Group on AI (HLEG) Ethical Guidelines for Trustworthy AI. These guidelines underline the importance of human agency and oversight, non-discrimination, transparency and accountability in the development and use of AI systems.

Research comparison of these frameworks in interactive AI applications - AI applications[75] showed that the current generation of responsible AI guidelines is a poor way to evaluate the ethics of AI applications as users. Many of the principles are vague, difficult to understand and often irrelevant to AI applications. AI applications-specific frameworks, such as Microsoft's, can be useful tools for developers and are easier to apply by users from more general philosophical frameworks. However, much more transparency is required so that the average user can see that the AI applications are complying with what is defined in a framework..

[75]Atkins, S., Badrie, I., & Otterloo, S.V. (2021). Applying Ethical AI Frameworks in practice: evaluating conversational AI AI applications solutions.*Computers and Society Research Journal*.

Key ethical harms and concerns tackled by these initiatives. All of the initiatives listed above agree that AI should be researched, developed, designed, deployed, monitored, and used in an ethical manner – but each has different areas of priority. This section will include analysis and grouping of the initiatives above, by type of issues they aim to address, and then outline some of the proposed approaches and solutions to protect from harms.

A number of key issues emerge from the initiatives, which can be broadly split into the following categories:

1. **Human rights and well-being**

   Is AI in the best interests of humanity and human well-being?

2. **Emotional harm**

   Will AI degrade the integrity of the human emotional experience, or facilitate emotional or mental harm?

3. **Accountability and responsibility**

   Who is responsible for AI, and who will be held accountable for its actions?

4. **Security, privacy, accessibility, and transparency**

   How do we balance accessibility and transparency with privacy and security, especially when it comes to data and personalisation?

5. **Safety and trust**

   What if AI is deemed untrustworthy by the public, or acts in ways that threaten the safety of either itself or others?

6. **Social harm and social justice**

   How do we ensure that AI is inclusive, free of bias and discrimination, and aligned with public morals and ethics?

7. **Financial harm**

   How will we control for AI that negatively affects economic opportunity and employment, and either takes jobs from human workers or decreases the opportunity and quality of these jobs?

8. **Lawfulness and justice**

   How do we go about ensuring that AI - and the data it collects - is used, processed, and managed in a way that is just, equitable, and lawful, and subject to appropriate governance and regulation? What would such regulation look like? Should AI be granted 'personhood'?

9. **Control and the ethical use – or misuse – of AI**

   How might AI be used unethically - and how can we protect against this? How do we ensure that AI remains under complete human control, even as it develops and 'learns'?

10. **Environmental harm and sustainability**

    How do we protect against the potential environmental harm associated with the development and use of AI? How do we produce it in a sustainable way?

11. **Informed use**

    What must we do to ensure that the public is aware, educated, and informed about their use of and interaction with AI?

12. **Existential risk**

    How do we avoid an AI arms race, pre-emptively mitigate and regulate potential harm, and ensure that advanced machine learning is both progressive and manageable?

5 Suggested fairness methods, practices and strategies

In this section, the concepts and recommendations identified in prominent recent literature, drafted by widely recognized scholars working at the intersection of law and ICT, are summarized, believed to be the optimal approach for developing best practices and policies on AI fairness. Subsequently, individual subsections are dedicated to each of these works, outlining the main concepts, principles, as well as the best practices and recommendations found within them. The focus is on presenting the findings of three prominent works, deemed comprehensive in addressing various aspects (gaps, challenges, recommendations) to bridge the gap between law and algorithms, ultimately achieving fairness in AI. Additionally, a proposed template for developing AI fairness policies is introduced, accompanied by a case study on hiring, which takes into account the knowledge and insights reported in the previous sections and can serve as a generic methodology that can be specialized according to the particularities of different application fields and use cases.

5.1 Suggestions from the law literature

5.1.1 Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non Discrimination Law - Wachter et al 2021

Concepts and principles

This study [42] emphasizes a lot on the trade-off between equal treatment (or equality of opportunity) versus equal outcome and how it translates with regards to the spirit of EU law. These concepts can by roughly mapped to the concepts of **formal equality** and **substantive equality** discussed in the study. Formal equality aims at ensuring that a system does not introduce additional bias to each decisions, however, it includes the implicit assumption that *the status quo is fair*. That is, formal equality assumes that there is not **historical/structural bias** in the data, thus the AI algorithm is not expected to learn such bias from the data. On the other hand substantive equality accounts for historical biases and aims at reducing them.

Although the concepts of substantive equality and equality of outcome and **affirmative action (alt. positive action)** are close, the authors emphasize their subtle difference. Videlicet, affirmative action aims at adjusting the output of a system so that positive outcomes are equitably and proportionally distributed to all protected groups defined by a sensitive attribute. Essentially, this translates to increasing the positive outcomes for members of disadvantaged/underrepresented groups. Instead, substantive equality aims to adjust, correct and improve the procedures, by taking into account the pre-existing, historical bias (in cases such exists) with the aim to gradually eliminate such biases.

The authors claim that the goal of the EU non-discrimination law is actually substantive equality. Formal equality comprises the norm of EU law, however, "*substantive equality comprises another fundamental normative concept in EU non-discrimination law. According to substantive equality or "de facto equality", true equality can only be achieved by accounting for historical inequalities which actively ought to be eroded.*"

Given the above discussion, the authors categorize existing algorithmic fairness definitions (alt. measures, metrics) into **bias preserving** and **bias transforming**. Bias preserving measures aim at ensuring that an AI system does not introduce additional bias in a decision

making process, being inline with equality of opportunity (treatment) and formal equality. Bias transforming measures aim at *leveling the playing field* between protected groups, by pursuing substantive equality. Figure 7 shows this categorization as presented in [42]. Roughly speaking, statistical measures that take into account solely the model's predictions in measuring fairness are categorized as bias transforming, while the ones that take into account both predictions and actual labels (existing in historical, labeled datasets) are considered as bias preserving. Further, most measures that examine or account for causal relations between attributes and outcomes, including counterfactual-based methods, are also categorised as bias transforming ones.

Fairness metric	Bias preserving?
1. Group fairness, Statistical (demographic) parity	✗
2. Conditional statistical (demographic) parity, Conditional independence	✗
3. Predictive parity, outcome test	✓
4. False positive error rate balance	✓
5. False negative error rate balance, Equal opportunity	✓
6. Equalized odds	✓
7. Conditional use accuracy equality	✓
8. Overall accuracy equality	✓
9. Treatment equality	✓
10. Test-fairness or calibration	✓
11. Well-calibration	✓
12. Balance for positive class	✓
13. Balance for negative class	✓
14. Causal discrimination (direct discrimination)	*
15. Fairness through unawareness	*
16. Fairness through awareness	✗
17. Counterfactual fairness	✗
18. No unresolved discrimination	✗
19. No proxy discrimination	✗
20. Path based causal reasoning	✗

Figure 7: Categorization of fairness metrics into bias preserving and bias transforming by [42]. Bias preserving measures, i.e. measures adhering to equal treatment and formal equality are marked with a "check mark". Bias transforming measures, aiming to substantive equality (close to equal outcome/affirmative action) are marked with an "X".

The authors take a clear stance in favor of bias transforming metrics, regarding them as the most promising means to achieve the desirable state of substantive equality. Nevertheless, they recognize the utility of both these types of measure, emphasizing their merit with respect to different scenarios and objectives:

- **Bias preserving measures**. Scenarios where such measures are suitable are when the user wants to test or "debug" a system on whether it introduces additional, "technical" bias. Further scenarios comprise cases when the owners/experts can be sure that ground

truth labels can be exactly verified by human experts and are not subject to debate; no structural bias exists; as well as no affirmative action policies are in place. Finally, cases where (indirect) discrimination between protected subgroups can be (legally) justified.

- **Bias transforming measures**. Scenarios requiring such measures naturally include application domains where structural/historical bias has been verified and cannot be legally justified. Usually, in such cases, affirmative action policies are in place, trying to compensate for direct or indirect discrimination towards disadvantaged protected groups.

Further, the authors emphasize that, currently, there is not consensus in the legal scholarship on how exactly substantial equality can be translated to practical steps, also in relation to EU law. In particular, they pose a set of questions that need to be answered and agreed upon, so that substantial equality can be effectively implemented in the frame of algorithmic fairness. These questions, directly quoted from [42] are:

- What is the end goal of non-discrimination law? To rectify historical harms and combat traditional power hierarchies? To achieve equality of distribution of goods for all? To accommodate diversity?
- What role (passive or active) is expected of the regulator, the legislator and the private and public sector?
- Should there be a practice and pre-emptive duty of the public and the private sector to dismantle inequality?
- Can this happen at the expense of dominant groups, potentially leading to positive discrimination?
- When can disparity be legally justified?
- How should the law address intersectional discrimination

The above questions are mainly addressed to policy makers, law experts and regulators and are crucial towards clarifying and specializing EU law with respect to AI fairness. Next a series of more "practical" suggestion are provided.

Suggestions and best practices

The authors of [42] position in favor of bias transforming measures towards achieving substantial equality, as a general principle to guide fairness in AI. Nevertheless, they provide a series of methodological suggestions and best practices for handling such delicate procedures. Next, the major ones are enumerated.

- **What is the objective?** The first step toward pursuing AI fairness in a specific field would be to decide what is wanted to be achieved in any given case: (a) Does the user want to test/debug an ML model so as to understand its behavior, potential biases in the training data, etc? (b) Or do they want to use the model for decision making. And if so, do they need to adhere to format equality, or do they want to achieve substantial equality. In case of (a), the recommendation is that the user may use either bias transforming, or bias preserving measures, according to their needs. In case of (b), then the following strps need to be taken into account

- **Does bias exist in the specific application domain?** This question aims to clarify if and what type of bias exists, which is crucial for deciding the next steps. Specifically, there might exist inequality in existing training data, but this might have been legally justified (e.g. in cases where physical attributes, which might favor males, serving as proxies for sex) can be legally taken into account in a decision making scenario. This steps also considers issues such as intersectional/subgroup fairness, or misrepresentation of minority subgroups in the data or in the (labelled) positive outcomes, and how relevant these issues might be for the specific domain.

- **Case by case examination.** To decide the next steps, the context, requirements, and particularities of the specific use case scenario to be addressed should be carefully examined. This includes the application domain (e.g. employment, education, finance, etc.), as well as other contextual factors, such as country, social circumstances, etc. The authors note that different types of bias, regarding the involved types of sensitive attributes, the intensity of discrimination, existing remedial policies in action etc, need to be taken into account in weighting subsequent fairness measuring and correction procedures. The authors use the apt example of blindly applying a drastic bias transforming measure (demographic parity) in decision making systems on loan applications. Such policies could lead to accepting loan applications from individuals that are unable to repay them, leading them to bankruptcy and, this way, actually perpetuating and reinforcing existing bias. Summarizing the procedure described in these three steps, the authors of [42] draft the checklist presented in Figure 8.

- **Training data selection.** The authors point out the importance of proper training data selection, so as to avoid several types of bias such as historical bias lying in the labels or bias caused by misrepresentation of protected (sub)groups. On the other hand, they note that, as the granularity of gathered profiles increases, privacy issues arise and need to be taken into account during the data collection process

- **Transparency and explainability.** Although not explicitly, the importance of transparency and explainability is discussed by the authors. Notably, they discuss that when in a specific case there exists the dilemma of formal equal treatment versus equal outcome, then it is important that a decision making system that adopts a specific strategy to be accompanied by summary statistics that reveal whether structural, historical or social bias exists in the specific application domain.

- **Proposed fairness measures.** Starting from the observation that "*some of the tests used by the European Court of Justice and Member State courts to measure indirect discrimination match the metric of demographic parity from algorithmic fairness*", the authors propose the adoption of two bias transforming measures. In particular, in cases where there exists justified indirect discrimination, the measure of Condidional Demographic Parity (alt. Conditional Independence) should be adopted. This measure ensures that any differences in a system's performance with respect to a protected attribute can be attributed to the conditioning variable(s). For a broader use, the authors propose the measure of Conditional Demographic Disparity, which examines the difference in acceptance and rejection rated between protected groups, separately for each dataset

partition defined by a conditioning variable and then aggregates the measured differences (disparities) to produce a stratified measure of disparity for the whole set. The authors firmly believe that these measures move toward serving substantial equality and, on the same time, reduce the risks of blindly applying more drastic measures, such as demographic parity.

Q1: Are you using fairness metrics to solely diagnose disparity, but are not making substantive decisions about individuals?

Yes: Both bias preserving and transforming metrics can be used.
No: Go to Question 2.

Q2: Are you deploying a system to make decisions in an area known to have unacceptable historical social inequality?

Yes: Go to Question 3.
No: Recommend investigation of possible bias in use case before choosing a metric. In cases where historical inequality does not exist, or known disparity has been deemed legally justified, both bias preserving and transforming metrics can be used.

Q3: Are you deploying the system and in a legal jurisdiction that **solely** promotes formal equality?

Yes: Both bias preserving and transforming metrics can be used.
No: Go to Question 4.

Q4: Are you deploying the system and in a legal jurisdiction that promotes substantive equality?

Yes: Recommend using a bias transforming metric.
No: Both bias preserving and transforming metrics can be used.

Figure 8: Checklist for selecting the appropriate type of bias measures, as drafted by [42]

- **Cross-sectorial synergies.** Closing their analysis, the authors point out that, in order for the aforementioned suggestions and practices to become feasible and effective, cross-sectorial collaboration between "*computer scientists and developers, lawyers, ethicists, social scientists, regulators, the general public*" is a prerequisite.

5.1.2 Algorithmic discrimination in Europe Challenges and opportunities for gender equality and non-discrimination law - Gerards & Xenidis 2021

Concepts and principles

Gerards and Xenidis [12] provide an extensive review of the current status in EU, regarding algorithms discrimination. They first summarize important concepts of AI algorithms relating to fairness and discrimination and try to identify mappings, links and gaps between algorithmic notions and the EU legal framework, examining law and precedents. They then present representative use cases of application of AI systems in various fields (e.g. hiring, education) and identify issues and opportunities towards handling the challenges of AI fairness. Finally, they enumerate a series of already applied best practices, on the policy level, as well as propose a framework for handling algorithmic discrimination. As with the previous study, we next focus a subset of concepts and recommendations identified in [12] that fit our analysis; nevertheless, we encourage the reader to consult [12] since it performs an insightful discussion on the field and offers further references to the relevant scientific literature, the legal framework and to relevant real world application cases.

The authors first identify several dimensions of AI algorithms that are relevant to AI fairness, many of which are already discussed in Section 3.4. First of all, they point out the important role of the **human factor** in algorithmic fairness. In particular, they discuss that AI algorithms are feed with often human-prepared and -label data, and are retrained based on human feedback, thus are prone into incorporating existing human/social bias.

Further, they emphasize the challenges imposed by the existence of correlations in the data and, in particular, **proxy variables**. They point out that, since AI algorithms are very good at learning patterns and correlations within the data, they make easier for bias to lurk implicitly into the AI models, via proxy variables, even when explicitly sensitive attributes are excluded from the analysis. A further danger is that the implicit bias represented by proxies is this way exacerbated, since correlations can be reinforced by the AI models.

Another important factor the authors identify regards the **quality of the training data**. They note that it is often the case that training data under- or mis-represent particular minority groups, which leads to introducing or even increasing existing bias. Also, the quality of the data might be low, regarding e.g. measurement errors or human bias affecting the quality of data labels. The authors emphasize that issues with data quality, along with the existence of correlations can lead to **feedback loops** that reinforce already existing structural discrimination. Further, the **scale and speed** of data obtained these days present additional challenges regarding bias detection, affecting all stages, from designing and developing to using an algorithm. For example, if a system developer fails to properly check and validate (vast and fastly incoming) training data or (vast and fastly produced) algorithm output data, bias might go unnoticed.

Transparency and explainability is another important dimension discussed in the study. Lack of these qualities hinders the system developers/owners from "debugging" their systems with respect to fairness, as well as users and auditors from identifying evidence of discrimination.

A final dimension recognized in the study revolved around responsibility: since a large number of individuals and potentially organizations, with different roles, are involved in the design, developement, validation and deployment of an AI system, it becomes rather difficult to distribute the responsibility each contributor has to the potentially unfair outcomes of the system.

Another significant discussion provided in the study concerns the explanation of the conceptual relation between **bias and discrimination** and **fairness and equality**. There, the authors note that the bias and fairness and grounded in the fields of statistics and ethics,

while discrimination and equality are mainly used in the legal field. "Bias" is a wider term than "discrimination", with the former referring to any type of systematic error observed by a system, that might have a statistical, cognitive, societal, structural or institutional origin. In contrast, algorithmic discrimination in the context of EU law "*only pertains to the unjustified unfavourable treatment of, or disadvantage experienced by, specific categories of population protected by the law*[76] *either explicitly (e.g. protected grounds) or implicitly (e.g. general or open-textured non-discrimination clauses)*" [12].

Likewise, "fairness" is a broader term than "equality", with the former being based on moral and ethical principles and having various interpretations (e.g. fairness definitions), and the latter being limited to the six protected attributes defined by EU law. The authors point out the importance of bridging the gaps between the broader, statistical and ethical terms of fairness and bias an the EU law, including answering questions such as "*whether the EU legal framework adequately captures algorithmic discrimination, how algorithmic discrimination challenges this legal framework and where potential frictions and inadequacies arise.*".

Suggestions and best practices

[12] proceed into summarizing and suggesting best practice solutions and tools, as well as a specific methodological framework towards achieving AI fairness. Regarding solutions and tools, they distinguish three categories: legal, knowledge-based and technology-based ones:

- **Legal solutions** The authors propose the "*targeted adaptation and purposive interpretation of existing legislation, doctrinal practices and institutional arrangements*", so that the EU law better accounts for the particularities and issues rising in the context of AI fairness. Specifically, the authors focus on the concepts of:
  - *Flexible view on protected grounds.* The authors propose that obtaining a more flexible viewpoint, beyond the currently strict consideration of the six protected attributes (racial and ethnic origin, sex, religion and belief, disability, age and sexual orientation) would facilitate in better handling the important issues of proxy and intersectional discrimination. As stated in [12] this "*could help judges tackle algorithmic discrimination by contextually defining 'new' protected grounds in order to expand the non-discrimination protection to proxy-based or intersectional forms of algorithmic discrimination*".
  - *Widening the scope of the gender equality and non-discrimination directives.* By this, the authors propose that there should be a reconsideration and potential expansion of protected attributes in application fields where currently only a limited set of such attributes are considered.
  - *Re-interpret the instruction to discriminate.* Examining and clarifying this concept in the context where an AI system assists in decision making would facilitate the allocation of responsibility for potential discrimination in the respective real world scenarios.
  - *Ease the burden of proof.* Since it is most often the case that AI systems quite complex to understand and/or "black box", meaning that access to their internals

[76] racial and ethnic origin, sex, religion and belief, disability, age and sexual orientation

is not allowed, the burden of proof that a system has discriminated against an individual becomes much heavier. The authors suggest shifting the burden of proof to the respondent, in case of lack of auditing procedures, transparency or explainability.

  - *Public and collective approach to monitoring and redress.* The authors suggest that public institutions should be assigned a supervisory role with respect to supervising and remedying algorithmic discrimination. This includes "*reforming these institutions, providing them with adequate resources, investigation, auditing, supervision and sanctioning powers, the ability to render legally binding decisions (in the same manner as data protection authorities) as well as standing rights in courts (as is already the case in some EU Member States)*".
  - *Accreditation, certification and supervision.* The authors emphasize the importance of drafting an accreditation system in European level that would set the principles, guidelines and would monitor individual, country-level organizations and processing for certifying and supervising AI fairness.
  - *Multi-disciplinary legal approach.* The authors proposed that experts from various fields of law (non-discrimination and gender equality law, data protection law, IP law, consumer protection law, etc.) need to be engaged and collaborate in procedures for addressing AI fairness.

- **Knowledge-based solutions** The second set of best practices revolves around informing and educating stakeholder and the public, so as to raise awareness on AI in general and AI fairness related issues, risks and solutions in particular. The authors propose that individuals that are involved in any step of the design, development or deployment of an AI system (e.g. designers and developers, system owners, decision makers, end users) should be educated on the basic characteristic of such systems and the ethical and legal risks their use entails. This suggestion goes along with the more general principle of increasing and facilitating the explainability of AI systems, towards any type of stakeholder involved. They suggest that mechanisms such as open databases reporting tools, scoreboards, etc. can be of use in this educational and awareness raising process. Finally, they emphasize that all types of stakeholders (regulators, judges, equality bodies, etc.) need to be training on the use of AI and on the risks of algorithmic discrimination. Education institutions of all levels should play a part in this process.

- **Technology-based solutions.** The third set of suggestions focuses on technological solutions, based on the fundamental differences between algorithmic and human-based decision making recognized above. The authors distinguish between *ex ante* and *ex post* strategies. Regarding the former, they emphasize the importance of techniques that are able to identify and quantify correlations between sensitive attributes and proxy variables and facilitate the examination of which correlations are justifiable and which comprise implicit discrimination. Further, to facilitate the above process, they claim that the use of sensitive attributes during the development of AI models should be permitted. Finally, the point out the utility of data augmentation strategies, for increasing the diversity of data, and data cleaning, for debiasing data. Regarding ex post strategies, they emphasize the importance of auditing mechanisms and procedures.

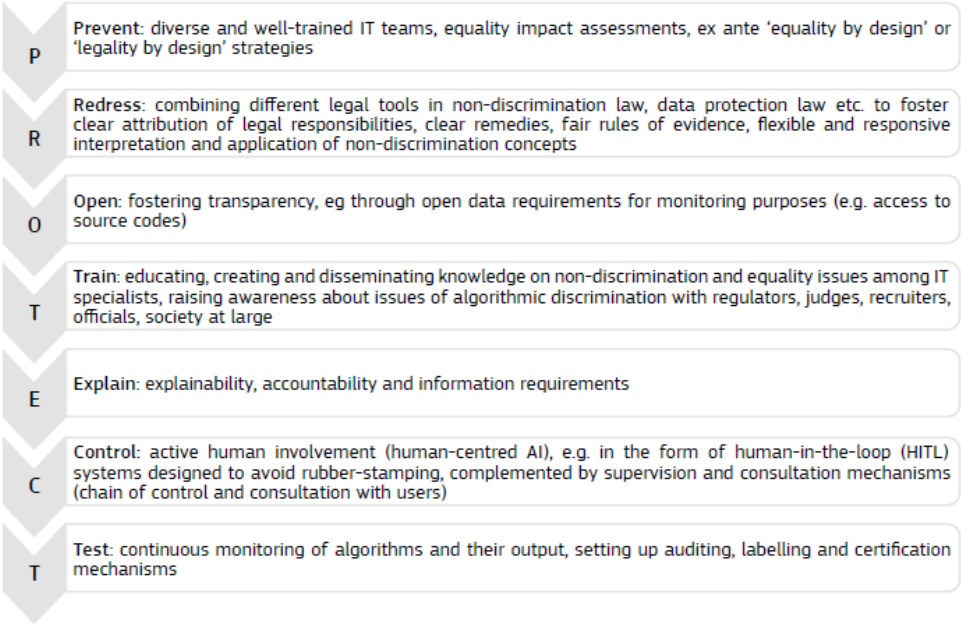

Figure 9: The suggested PROTECT approach for AI fairness by Gerards and Xenidis [12]

The authors' suggestions are summarized in the PROTECT framework they propose, depicted in Figure 9.

5.1.3 Legal perspective on possible fairness measures –A legal discussion using the example of hiring decisions - Hauer et al. 2021

Concepts and principles

Hauer et al. [15] comprises one of the few works that delve into the specifics of algorithmic fairness definitions and try to map them to existing law concepts. They first emphasize a significant gap between law and algorithms, consisting in the **process-oriented** nature of (EU) law, versus the **result-oriented** nature of algorithmic systems. In a process-oriented assessment, each individual case would be examined in detailed and e.g. "*a judge would need to evaluate whether the formulated requirements are appropriate and the resulting bias is acceptable in this specific case*". This is not feasible in an algorithmic setting, since an AI system has learnt to model patterns over large datasets, i.e. large number of individual cases, and a result of an AI system cannot adequately represent all the specifics and the context of the respective case under examination. The authors remark that "*To date, there has been no legal definition of fairness beyond individual decisions of jurisdiction. This is mainly due to the process-oriented assessment of equality. With the upcoming algorithm-based decision making though, it will be essential to find a clear specification of fairness as the assessment will no longer work on a process level. Rather, there will be a decisive shift to a result-based assessment. It is therefore of utmost importance to analyze, whether and how the fairness measures presented in this paper comply with EU anti-discrimination legislation*"

The authors consider two broad categories around algorithmic discrimination: (a) issues caused by the design and development of algorithmic systems and (b) issues related to the training data. Focusing on the latter, they recognize and emphasize that a major challenge is

to be able to decide on "*What distribution of results can be deemed discriminatory?*", which leads to the question of "*if and how different fairness measures can be integrated into existing legislation?*", given that different fairness measures fit different use cases.

Another point they make regards proxy discrimination; the authors emphasize that simply excluding sensitive attributes (i.e. fairness by unawareness) does not guarantee fairness, since it is often the case that other variables are correlated with sensitive ones, thus carrying the corresponding bias. Consequently, maintaining sensitive attributes, even when they do not contribute to the decision making process of an AI system, is crucial to be able to assess fairness.

The authors discuss the spirit of the EU law, as expressed via the General Equal Treatment Act (GET) of Germany[77]. They note that "*the difference between direct and indirect discrimination relies on a process-oriented assessment that looks at the actual decision process and can no longer be upheld for algorithmic decision making. It is therefore necessary to find a unified definition of discrimination that does not differentiate between direct and indirect discrimination. This unified definition must be based on the common properties of both direct and indirect discrimination. The common ground between the two forms of discrimination is that in both cases a person is treated less favorably than another is, has been or would be treated in a comparable situation. How such a unified definition complies with the GET and the underlying directives remains unclear to this date. It is unlikely that such a definition can be simply derived from the GETs or the underlying directives text. Even when factoring in the primary EU law of EU CFR and TFEU, it is doubtful that they allow for such a wide interpretation. It will therefore indeed be necessary to establish a new discrimination term in the EU directives to adapt to the new situation of algorithmic decision making.*."

Then, they point out that discrimination is defined in terms of comparison with a comparator group, however, for algorithmic systems, this comparison is limited: "*Machine learning models are trained with training data sets with which the model learns its prediction. The trained model is then making predictions for new data without being trained anymore. As such, the decision process is defined by the structure of the training set and the conditions set for the model during training. This means that a discrimination of a person can-not be based on a differing individual opinion or preference in the past, because the machine learning model always decides based on a determined algorithm that has been developed from a certain training data set. Therefore, the idea of comparing hypothetical and former treatments in anti discrimination law has to be partly reconceived.*"

Then the authors review a series of algorithmic fairness definitions, on a hiring example, following the categorization of [2], examining six forms of fairness: **Independence**, **Conditional independence**, **Separation**, **Sufficiency**, **Individual fairness** and **Counterfactual fairness**.

Independence. According to Independence, as represented by, for example, statistical parity, the system output must be independent of sensitive attributes. The probability of obtaining a positive outcome should be the same for the different groups identified by the sensitive attributes. Independence does not take into account the actual ground truth, making it a definition suitable for implementing positive discrimination (alt. positive action, affirmative action, equal outcome). The authors note that such a measure cannot be a permanent one but rather a temporary one, since it "*violates the right to develop ones own personality out of*

[77] `https://www.gesetze-im-internet.de/englisch_agg/englisch_agg.html`

ones individual freedom that is inherent in EU primary right.".

Conditional independence. Conditional independence improves upon independence by requiring equal probabilities of positive outcomes between protected groups **within groups** that are similar in some other attributes (attributes the we condition on). The authors present a more positive view on this definition, claiming that "*conditional independence is a highly attractive fairness measure. It is the idea of equality of opportunity that only the characteristics of a person are important, which they themselves can change, or at least influence. When choosing such features as the key factor for fairness, Conditional independence corresponds with the constitutional goal of the individual development of ones freedom. Furthermore, the fact that the ground truth is not regarded, makes the integration of structural imbalances such as racism or sexism at least less likely*". However, they recognize that choosing the proper attributes to condition on is a difficult and case specific task, while it may also perpetuate discrimination on groups that do not fulfil the conditions, in cases where the conditioning variable serve as proxies for structural bias.

Separation. Separation, represented by, for example, equalized odds, demands that the system's prediction is independent from the sensitive attribute, given the ground truth. Although separation adheres to the principle of formal equal treatment, the authors remark that it "*promotes both equal bias and equal accuracy in all demographic groups and thus discriminates in models that perform well only on the majority. Therefore the measure helps to further strengthen existing bias in the data set (the ground truth)*".

Sufficiency. Sufficiency, represented by, for example, calibration, demands that the positive outcome is independent from the sensitive attribute, given the system's prediction. The authors remark that "*the mathematical formula behind Sufficiency allows the false-positive rates to diverge between groups, which can have a severe impact and makes this fairness measure highly debatable*"

Individual fairness. Individual fairness demands that a system's output is similar for similar individuals. The authors present a negative view on this definition. Firstly, they discuss the widely known problem of defining a proper similarity between individuals to be used by the measure. However, it recognizes a more structural issue of the specific definition: "*Individual Fairness is hardly able to map these qualifications which deviate from the general set of characteristics. Rather, it heavily advantages the existing majority groups representing the majority of people who will be positively classified. These majority groups form the property set that Individual Fairness compares people with to determine their similarity. This means that Individual Fairness actually asks from minority groups to adapt to the property set of the majority group... Even if Individual Fairness pretends only to regard the individual, it in fact leads to a decomposition of individuality in favor of a homogeneous majority driven prototype personality. This consequence is not compliant with the ideal of individuality and development of the individual persons freedom and cannot be implemented into a machine learning model in a EU law compliant way*".

Counterfactual fairness. Counterfactual fairness demands that the decision of a system would be the same in an imaginary world where the sensitive attribute would have a different value. The authors remark the main challenge of this definition, which is constructing a causal model (graph) that represents the causal relations between examined attributes, but opt for this metric, since they believe it "*is based on a process-oriented understanding of equality that is then connected with a result-mapping process. The idea of a counterfactually fair decision*

output is fascinating as it simultaneously corresponds with the goal of anti-discrimination law and infringes it in its process."

Suggestions and best practices

The authors focus on examining algorithmic fairness definitions with respect to their compliance with the EU law, as well as to which principle they serve (formal equal treatment, equal outcome, or substantial equal treatment). They single out the measures of **Conditional independence**, **Separation**, **Sufficiency** and **Counterfactual fairness** as ones that can be adopted on different application scenarios, stating that they "*can all possibly be implemented in compliance with EU law dependent on their concrete application.*" However, they distinguish counterfactual fairness as a definition that can most effectively bridge the currently process-oriented viewpoint of EU law with a result oriented approach necessary in algorithmic systems. This can be achieved by assigning the creation of causal models to domain experts and the auditing of these models and the respective decision making systems to public authorities. The authors further comment that counterfactual fairness "*Counterfactual Fairness upsets the whole legal conception of equality, not in questioning individuality but in showing that our process-based assessment is inherently flawed...* ". This is because it suggests that sensitive attributes and their proxies should not be excluded, but utilized in a process-oriented assessment. If properly done, counterfactual fairness could isolate the objective value of proxy variables versus the bias they carry and properly utilize only the former in a decision making algorithm.

Another significant point made by the authors, as with previously presented studies, is the need for cross-sectorial collaboration between lawmakers, politicians and domain experts to decide which fairness measures should be implemented under which conditions, specializing this way EU law. In this setting, consequences and implications in society need to be (re)estimated. Furthermore, since AI systems comprise dynamic and evolving systems, their evaluation/auditing should be a continuous process. Another suggestion, also related to the dynamicity, is that fairness definitions should not incorporated directly into AI models but should be used as an auditing tool, assisting a result-oriented review of such systems by supervisory/auditing authority.

Finally, the authors strongly recommend the exploitation of transparent model or explainability mechanisms that facilitate the understanding for their decision making.

5.2 A proposed template for developing AI fairness policies

In this section, we first provide a guide for developing policies that implement anti-bias and fairness principles in an AI sector, using the EU requirements for trustworthy AI as key guiding principles, and then we provide a case study on recruiting.

Please note that any policy document in such a fast flowing domain should become part of a continuous | **co-creation** | **co-production** | **co-evaluation** | cycle that informs policy design, implementation and improvement. The proposed process will involve the following steps:

1. Scoping
2. Stakeholder Engagement
3. Risk Assessment
4. Policy Development
5. Implementation
6. Evaluation and Improvement

Step 1: Scoping
The scope should include the type of AI applications, the sector in which they are used, and the potential impact on individuals and society, what are the expected benefits and risks of AI, and what are the main ethical principles and values that should guide AI development and use. The process should analyse the existing legal and ethical frameworks that apply to AI at national, regional, and international levels, and assess their strengths and weaknesses, their gaps and overlaps, their compliance with human rights standards, and their alignment with the EU requirements for trustworthy AI.

Step 2: Stakeholder Engagement
Who will be affected by the policy? Stakeholders may include experts in AI, civil society organizations, affected individuals, and other relevant groups. The engagement process should involve dialogue and consultation to identify concerns, gather feedback and input, and build consensus. It can include stakeholders such as AI developers, users, regulators, civil society organisations, experts, and affected groups. It should gather their views and feedback on the current state of AI ethics and regulation, their needs and expectations, their challenges and concerns, and their suggestions for improvement.

Step 3: Risk Assessment
The risk assessment process should consider potential biases and fairness issues and identify the potential impact on individuals and society. The assessment should also consider the potential benefits of the AI application and weigh them against the risks.

Step 4: Policy Development
The policy should be developed based on the EU key requirements for trustworthy AI and

should address the concerns and feedback gathered during stakeholder engagement and risk assessment. The policy should outline specific guidelines for the development and deployment of AI systems, including requirements for transparency, accountability, and respect for human autonomy. Policy development should take into account the set of criteria for assessing whether an AI system meets the EU requirements for trustworthy AI, especially those related to diversity, non-discrimination, fairness, transparency, accountability, human agency and oversight.

It should also be cognizant of current best practices for preventing or mitigating bias and unfairness in AI systems throughout their life cycle (design, development, deployment, use, evaluation, etc.) and a clear description of the roles and responsibilities of different actors involved in or affected by AI systems (developers, users, regulators, auditors, etc.)

Step 5: Policy Implementation
The implementation process should involve training and education for stakeholders involved in the development and deployment of AI systems. It should also include monitoring and evaluation mechanisms to ensure that the policy is being followed and that any biases or fairness issues are detected and addressed, i.e., for monitoring and enforcing compliance with the policy guidelines by using appropriate mechanisms such as audits, inspections, sanctions, incentives and other tools.

Step 6: Evaluation and Improvement
The evaluation process should assess the effectiveness of the policy in achieving its intended outcomes and identify areas for improvement. The policy should be regularly reviewed and updated to reflect new developments in AI technology and changes in societal values and expectations.

5.3 Case Study | Main considerations for an organization developing and deploying ethical AI for recruiting

Any policy regarding AI should apply to all employees, contractors, vendors, and other stakeholders who are involved in the development and deployment of AI systems for job application processes.

The organization should be explicitly committed to creating and deploying AI systems that are free from bias and ensure fairness in job application processes. The organization should implement the following guidelines to ensure that its AI systems are developed and deployed in a manner that is consistent with this commitment:

1. Training Data:

1. Ensure that the training data used for the AI systems is collected from diverse sources to ensure that the data is representative of the population.

2. Take steps to eliminate biases in the training data and use algorithms to detect and mitigate any biases that cannot be eliminated.

2. Algorithmic Bias:

1. Implement measures to ensure that its AI algorithms are free from bias. The organization will use techniques such as, for example, counterfactual analysis to detect and mitigate any biases that are found in the algorithms.

3. Transparency:

1. Provide clear and transparent explanations of how AI systems work for job applications.
2. The organization will make the data used to train its AI systems available to the public to ensure that its AI systems are subject to external scrutiny.

4. Privacy:

1. The organization will ensure that its AI systems comply with applicable data protection laws and regulations.
2. The organization will implement measures to protect the privacy of job applicants whose data is used to train and test its AI systems.

5. Accountability:

1. Assign responsibility for the development and deployment of its AI systems to specific individuals.
2. Implement measures to ensure that these individuals are held accountable for any biases or other issues that arise from the use of the AI systems in job applications.

6. Continuous Monitoring:

1. Implement measures to continuously monitor the performance of its AI systems to ensure that they are working as intended.
2. Implement measures to detect and mitigate any biases or other issues that arise during the job application process.

7. Diversity and Inclusion:

1. Take steps to ensure that its job application process and workforce are diverse and inclusive.
2. Ensure that the development and deployment of its AI systems take into account the needs and interests of all job applicants.

8. Bias Detection:

1. Implement measures to detect biases in its AI systems used for job applications. The organization will use techniques such as fairness metrics and interpretability methods to detect and mitigate any biases that are found.

Implementation

The organization should implement the guidelines set out here this policy document in the following manner:

1. Appoint a designated individual or team to oversee the implementation of this policy.
2. Provide training to all employees, contractors, vendors, and other stakeholders who are involved in the development and deployment of AI systems for job application processes.
3. Regularly review and update this policy document to ensure that it remains up to date with best practices and regulatory requirements.
4. Establish a process for reporting and addressing any concerns or issues.
5. The organization will make any adopted policy documents available to all job applicants, employees, contractors, vendors, and other stakeholders who are involved in the development and deployment of AI systems for job application processes.

Key Policy Considerations / Steps

1. **Develop clear and concise policy objectives**:
   The policy should clearly define the objectives of anti-bias and fairness in recruitment AI, which may include eliminating discriminatory practices, reducing bias, increasing diversity, and promoting fairness and transparency.
2. **Select appropriate data sources:**
   The AI used for recruitment should rely on data that is unbiased and fair, ensuring that it is not contaminated by historical bias or stereotypes. Therefore, the data sources used should be selected carefully to ensure that they are appropriate and reliable.
3. **Choose the right algorithm:**
   The algorithm used for recruitment should be chosen carefully to ensure that it is unbiased and fair. It should be designed to detect and mitigate biases in the data.
4. **Establish transparency and explainability**:
   The AI used should be transparent and explainable, meaning that the process of decision-making should be open and understandable to candidates. Candidates should be informed about how the algorithm works and how their data is being used.
5. **Conduct regular audits:**
   Regular audits should be conducted to ensure that the recruitment AI system is working as intended and is not causing any unintended negative consequences, such as perpetuating bias or unfairness.
6. **Include diversity and inclusion as a core value:**
   Diversity and inclusion should be integrated into the core values of the organization to ensure that they are prioritized in all aspects of the recruitment process.

7. **Ensure human oversight:**
   Human oversight should be included in the AI recruitment process, especially for final decision-making, to ensure that the decisions are fair and unbiased.

8. **Train staff:**
   Staff involved in the recruitment process, including hiring managers and HR personnel, should receive training on anti-bias and fairness in AI recruitment to ensure that they are aware of the potential biases and can make informed decisions.

9. **Collect feedback:**
   Candidates and employees should be encouraged to provide feedback on the recruitment process, including the AI used for recruitment, to ensure that the system is fair, transparent, and unbiased.

Ensuring multi-stakeholder involvement at all policy stages

Developing a policy that implements anti-bias and fairness principles in the AI sector is a complex process that requires careful consideration of the potential impact on individuals and society. This process itself should be regularly reviewed and updated to ensure that it remains effective and up-to-date with best practices and societal values.

The implementation of anti-bias and fairness policies in AI recruitment can be enhanced by the involvement of a multi-stakeholder community of practice, which should include job candidates. The community of practice brings together stakeholders with diverse perspectives and expertise to collaborate and share knowledge, experiences, and best practices for ensuring the development and deployment of fair and unbiased AI systems in recruitment.

It is suggested that an adapted version of the Communities of Practice Playbook (2021)[78] is used for this purpose especially for organizing and managing the relevant community. This playbook offers extensive guidelines as well as practical tools to assist in the operation of such a community.

Communities of practice are groups of stakeholders who come together to collaborate and share knowledge, experiences, and best practices for a particular topic or field of interest. These communities are characterized by active participation, collaboration, and mutual learning. Some of the main attributes of participatory communities of practice include:

- **Active participation:** Members of participatory communities of practice are actively engaged in the community and contribute to its activities and discussions. Participation can take many forms, such as attending meetings, sharing resources, or providing feedback.

- **Collaborative learning:** Members work together to learn from each other and to collectively develop solutions to problems. Collaboration can occur through sharing knowledge, discussing ideas, and working together on projects.

- **Shared purpose:** Participatory communities have a shared purpose, such as improving a particular practice, field, or system, in this case AI development and deployment. The

[78] Catana, G.C., Debremaeker, I., Szkola, S.S.E. and Williquet, F., The Communities of Practice Playbook, EUR 30466 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-26344-9, doi:10.2760/42416, JRC122830.

shared purpose provides a focus for the community's activities and helps to guide its decision-making.

- **Mutual benefit:** Members benefit from their participation by gaining knowledge, building relationships, and contributing to the community's work. The community also benefits from the knowledge and skills of its members.
- **Inclusivity**: Such communities are inclusive and encourage participation from diverse stakeholders. This inclusivity helps to ensure that the community is representative of the broader group of stakeholders it serves and that it considers a range of perspectives and experiences.
- **Openness**: Communities of practice are open to new ideas, perspectives, and feedback. Members are encouraged to share their thoughts and experiences openly and to listen to the ideas of others.
- **Informality**: They are also typically informal in structure and focus on building relationships and trust between members. Informality helps to foster a sense of community and encourages members to participate more freely.

By fostering active participation, collaborative learning, and inclusivity, communities of practice can promote knowledge sharing, build trust, and lead to more effective and sustainable solutions to problems. A multi-stakeholder community of practice can be instrumental in achieving anti-bias and fairness policies in AI recruitment via:

- **Representation of job candidates:** The community can ensure that job candidates are represented in the development and deployment of AI recruitment systems. This representation can be in the form of focus groups, surveys, or user testing, where job candidates can provide feedback on the system and ensure that it is fair and unbiased.
- **Participatory design:** diverse stakeholder groups can be involved in the design of AI recruitment systems. Participatory design involves engaging stakeholders in the design process to ensure that their needs, concerns, and perspectives are incorporated into the final product. Involving job candidates in participatory design can ensure that the AI system is designed to meet their needs and is fair and unbiased.
- **Collaboration on data collection:** stakeholders - members of the community can collaborate with organizations to collect and analyze data for AI recruitment systems. Collaboration on data collection can ensure that the data used for AI recruitment is unbiased and representative of the job market.
- **Feedback and monitoring:** stakeholders including job candidates can provide feedback on AI recruitment systems and monitor their use to ensure that they are fair and unbiased. This can include reporting incidents of bias or unfair treatment, and providing feedback on the user experience of the system.

6 Conclusions

D3.1 performed a thorough review on AI fairness policies and methodologies, considering both the legal and the algorithmic viewpoints. In particular, it provided an overview of AI legislation worldwide, and well as non-discrimination law in the EU and the US. Furthermore, it summarized a set of prominent algorithmic fairness definitions and methods for bias detection and enumerated a set of assessment criteria for these. Then, it presented a series of use cases, legal precedents, policies and frameworks related to AI fairness and ethical AI. Finally, it presented a set of suggestions regarding best practices towards achieving AI fairness, as well as a proposed template for developing AI fairness policies.

Some of the major findings of the performed work, which should be taken into account in policy making for fairness AI are summarized next:

- Discrimination by proxy variables, intersectional fairness and feedback loops comprise major issues when pursuing fairness in real world applications. Ongoing work is being performed in the algorithmic literature, although no one-size-fits-all solutions exist yet. The current EU legal framework needs to be adapted and extended, since it is widely recognized that it has not been able to appropriately handle such issues up until now.

- No one-size-fits-all fairness definitions or bias detection methods exist. Fairness is highly application-, scenario- and context-specific, since different real world applications of AI decision making systems and different social circumstances highly affect what is considered as "fair". Since the law cannot specialize on a case by case basis, this needs to be done by domain experts in collaboration with governmental and independent supervising and auditing authorities.

- Cross-sectorial collaboration is a necessity in practically every step of building both fair-by-design systems and methodologies, and AI fairness policies. There exists a large gap between law/ethics and data/algorithms and only such collaboration can bridge this gap and produce meaningful policies and best practices.

- Some specific fairness definitions have been distinguished by a handful of prominent studies on the intersection of law and algorithms [42, 15]: Conditional Independence, Separation (e.g. equal opportunity, equalized odds), Sufficiency (e.g. Calibration) can be considered suitable in different the application settings and contexts, while Counterfactual Fairness is considered a sufficiently expressive and adaptable definition that allows it to generalize in different cases and optimally represent substantial equality, in the spirit of the EU law.

The deliverable addresses a wide audience, comprising both law and ICT researchers, as well as policy makers and AI stakeholders. We make an effort to exemplify definitions and methods, while also identifying examples (precedents, use cases, proposed methodologies) from the literature that contribute to addressing existing gaps, risks and challenges in achieving AI fairness. The work done in this deliverable already provides guidance to the consortium, enabling us to explore promising methodologies such as counterfactuals and optimal transport in the conducted research.

References

[1] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[3] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

[4] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

[5] Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness, 2022.

[6] Diptarka Chakraborty, Syamantak Das, Arindam Khan, and Aditya Subramanian. Fair rank aggregation. *Advances in Neural Information Processing Systems*, 35:23965–23978, 2022.

[7] Farah Cherfaoui, Hachem Kadri, Sandrine Anthoine, and Liva Ralaivola. A discrete rkhs standpoint for Nyström mmd. *HAL open science*, 2022. hal-03651849.

[8] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163, 2016.

[9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery.

[10] B. Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *SafeAI@AAAI*, 2020.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[12] Janneke Gerards and Raphaele Xenidis. *Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-Discrimination Law*. European Commission, 2021.

[13] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

[14] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[15] Marc P Hauer, Johannes Kevekordes, and Maryam Amir Haeri. Legal perspective on possible fairness measures–a legal discussion using the example of hiring decisions. *Computer Law & Security Review*, 42:105583, 2021.

[16] F. Kamiran, I. Zliobaite, and T.G.K. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644, 2013.

[17] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 895–905. PMLR, 2020.

[18] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

[19] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *FAccT*, pages 353–362. ACM, 2021.

[20] Loukas Kavouras, Konstantinos Tsopelas, Giorgos Giannopoulos, Dimitris Sacharidis, Eleni Psaroudaki, Nikolaos Theologitis, Dimitrios Rontogiannis, Dimitris Fotakis, and Ioannis Emiris. Fairness aware counterfactuals for subgroups, 2023.

[21] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2017.

[22] Pauline T. Kim. Data-driven discrimination at work. *William and Mary law review*, 58:857, 2017.

[23] Pauline T Kim. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *Cal. L. Rev.*, 110:1539, 2022.

[24] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Information Technology Convergence and Services*, 2016.

[25] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26:1481–1496, 06 1997.

[26] Alejandro Kuratomi, Evaggelia Pitoura, Panagiotis Papapetrou, Tony Lindgren, and Panayiotis Tsaparas. Measuring the burden of (un) fairness using counterfactuals. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part I*, pages 402–417. Springer, 2023.

[27] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. NIPS'17, page 4069–4079, Red Hook, NY, USA, 2017. Curran Associates Inc.

[28] Council of Europe, European Court of Human Rights, and European Union Agency for Fundamental Rights. *Handbook on European non-discrimination law – 2018 edition*. Publications Office of the European Union, 2018.

[29] European Court of Human Rights. European convention on human rights, 2010.

[30] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 560–568, New York, NY, USA, 2008. Association for Computing Machinery.

[31] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommenders: Models, methods and research directions. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 2358–2361. IEEE, 2021.

[32] Arjun Roy, Jan Horstmann, and Eirini Ntoutsi. Multi-dimensional discrimination in law and machine learning - a comparative overview. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 89–100, New York, NY, USA, 2023. Association for Computing Machinery.

[33] Mark E Rushefsky. *Public policy in the United States*. ME Sharpe, 1996.

[34] Dimitris Sacharidis, Giorgos Giannopoulos, George Papastefanatos, and Kostas Stefanidis. Auditing for spatial fairness. In *Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023*, pages 485–491. OpenProceedings.org, 2023.

[35] Bernhard Schölkopf, Ilya Tolstikhin, and Bharath Sriperumbudur. Minimax estimation of maximum mean discrepancy with radial kernels. *NIPS*, 2016.

[36] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *AIES*, pages 166–172. ACM, 2020.

[37] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

[38] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[39] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *AAAI*, pages 9584–9594. AAAI Press, 2022.

[40] Sandra Wachter. Affinity profiling and discrimination by association in online behavioural advertising. *SSRN Electronic Journal*, 2019.

[41] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[42] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

[43] Raphaële Xenidis. Tuning eu equality law to algorithmic discrimination: Three pathways to resilience. *Maastricht Journal of European and Comparative Law*, 27(6):736–758, 2020.

[44] Yiqun Xie, Erhu He, Xiaowei Jia, Weiye Chen, Sergii Skakun, Han Bao, Zhe Jiang, Rahul Ghosh, and Praveen Ravirathinam. Fairness by "where": A statistically-robust and model-agnostic bi-level learning framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12208–12216, Jun. 2022.

[45] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pages 1–6, 2017.

[46] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000*, 2021.