

---

Electronic Thesis and Dissertation Repository

---

2-16-2022 11:30 AM

## Design of a Capability and Maturity Model for the Development of Trustworthy ADM Systems Based on Principled AI

Daniel Varona Cordero, *The University of Western Ontario*

Supervisor: Suárez, Juan L., *The University of Western Ontario*

Co-Supervisor: Pierre-Gerlier, Forest, *University of Calgary*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Hispanic Studies

© Daniel Varona Cordero 2022

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>



Part of the [Digital Humanities Commons](#), and the [Other Engineering Commons](#)

---

### Recommended Citation

Varona Cordero, Daniel, "Design of a Capability and Maturity Model for the Development of Trustworthy ADM Systems Based on Principled AI" (2022). *Electronic Thesis and Dissertation Repository*. 8391. <https://ir.lib.uwo.ca/etd/8391>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact [wlsadmin@uwo.ca](mailto:wlsadmin@uwo.ca).

## Abstract

Automatic decision-making (ADM) systems have permeated every sphere of society where a large amount of data is managed to fulfill prediction/classification needs. The enhanced capabilities ADM systems have brought into their applied sciences conditioned their evolution to more complex and less transparent machine learning algorithms and models (MLA & M). Nowadays, dissimilar predictions, or suggested decisions supported by MLA & M are found to be misleading, or discriminatory resulting in heated academic and public debates since these MLA & M are being applied in socially and politically sensitive areas such as crime prevention, justice management, among others. Thus, there exists an increasing concern among scholars and regulators regarding biased decisions when using complex non-transparent MLA & M, leading to the pursuit of an ethical development process to create ADM solutions. Available approaches lean towards the regulatory aspects of this problem, with a focus on the Human Rights International Law, to define the supposed trustworthiness of trustworthy artificial intelligence (AI). There is still a need to explore how this approach both intersects and harmonizes with the design-based engineering pursuit to achieve fairer decisions. This dissertation proposes a capability and maturity model for trustworthy ADM solutions to help reduce the social gap experienced by social minorities, such as the Hispanic community, because of discriminatory automated decisions. First, the specialized literature on bias in ADM systems is analyzed to identify current limitations of ML in fairness achievement. Also, the so-called international regulatory framework on “principled AI” is studied to determine which elements may be influenced to achieve design-based trustworthy ADM solutions. Variables like Discrimination, Bias, Fairness, and Trustworthiness, relevant within the principled AI context, are explored and incorporated within the model. The findings of this research project highlight the limitations of ML which 1) amplify and perpetuate bias and 2) stress the constraints of the AI international regulatory framework as a complementary methodological support for ADM solutions engineering. This reinforces the need for policy and software developers to join efforts to assure fairer outcomes produced by ADM systems.

## Keywords

Trustworthy Artificial Intelligence, Trustworthy AI, Principled Artificial Intelligence, Principled AI, Bias in Machine Learning, Machine Learning, Bias Automation, Biased Artificial intelligence, Biased Data Sets, Artificial Intelligent Systems, Discrimination, Discriminatory Decisions in Machine Learning, Discriminatory Artificial Intelligence, Fairness, Fairness from Design, Artificial Intelligence Design, Trustworthy AI Capability, and Maturity Model.

## Summary for Lay Audience

This thesis dissertation is the culmination of a research project aiming to reduce the discriminatory outcomes of ADM tools using AI and ML, by articulating a software engineering methodological model to ensure fairer decisions from trustworthy ADM solutions. The proposed model uses a structure similar to a known popular quality assurance model called CMMI, which develops a series of quality characteristics across different process areas organized in capability and maturity levels. The trustworthy-related variables are like quality variables already available in the software industry, which currently exhibits a functional dimension. Consequently, the model redefines these variables and integrates them using their ethical perspective, enhancing the available quality assurance approach in the software industry. To do so, exploratory studies of the current engineering methodological approach to ADM solutions, and of the principled AI international framework (a set of regulatory mechanisms seeking to reduce discriminatory outcomes produced by ADM technology with a focus on the International Law of Human Rights) were conducted. The resulting capability and maturity model for trustworthy ADM solutions proposed in this thesis is important as it helps reduce the social gap experienced by minorities, including the Hispanic community, as result of discriminatory automated decisions influencing the design of trustworthy ADM solutions early in the development process.

## Co-Authorship Statement

Chapter 1 is an extended version of an article already published in a peer-reviewed journal.

Chapter 2 is already published as a chapter for a specialized volume in computational social science.

Chapters 3, 4, and 5 are manuscripts which will be submitted to peer-reviewed journals.

The contribution of each author is stated below:

### **Chapter 1: Machine Learning's Limitations in Avoiding Automation of Bias**

Authors: Daniel Varona, Yadira Lizama-Mue, Juan-Luis Suárez

Status: Published in AI & Society

Analysis and writing were performed by Daniel Varona. Visualizations were made by Yadira Lizama-Mué. Juan Luis Suárez provided the theoretical background as well as copy-editing and consultation regarding interpretation of results. The manuscript was proof-read by Zeina Dghaim.

### **Chapter 2: Analysis of the Principled-AI Framework's Constraints to Become a Methodological Reference for Trustworthy-AI Design**

Authors: Daniel Varona, Juan-Luis Suárez

Status: Published as a chapter in Volume 1: Theory, Case Studies and Ethics, Handbook of Computational Social Science.

Experimental work, data analysis, visualizations and writing were performed by Daniel Varona. Juan-Luis Suárez provided consultation regarding experimental work, interpretation of results, general background, and methodological framework.

### **Chapter 3: Principled AI Engineering Challenges Towards Trustworthy AI**

Authors: Daniel Varona, Juan-Luis Suárez

Status: To be submitted to Target Journal (still to be defined)

Analysis and writing was performed by Daniel Varona. Juan-Luis Suárez provided the theoretical background as well as copy-editing and consultation regarding interpretation of results.

#### **Chapter 4: Discrimination, Bias, Fairness, and Trustworthy AI**

Authors: Daniel Varona, Juan-Luis Suárez

Status: To be submitted to Target Journal (still to be defined)

Analysis and writing were performed by Daniel Varona. Juan-Luis Suárez provided the theoretical background as well as copy-editing and consultation regarding interpretation of results.

#### **Chapter 5: Proposal of a Capability and Maturity Model for Trustworthy ADM Systems**

Authors: Daniel Varona, Juan-Luis Suárez

Status: To be submitted to Target Journal (still to be defined)

Analysis, modeling, and writing were performed by Daniel Varona. Juan-Luis Suárez provided the theoretical background as well as copy-editing and consultation regarding interpretation of results.

## Dedication

To my sons Diago, Fabian, and Lucas, who endured the sacrifices associated with Graduate school. I hope my effort inspires them to pursue their dreams, whatever they are.

## Acknowledgments

There were many people who helped shape this thesis. I would like to thank:

My supervisor, Professor Juan-Luis Suárez. I have been very fortunate to have his guidance and inspiration. Under his supervision I found that I have outgrown my engineering background, shaped by the software engineering industry and software engineering empirical studies, and transitioned into a professional with the ability to think outside of the rigid margins of formal education. I also found that I have improved several aspects of my personal and professional life, and that he has motivated me to pursue a better version of myself.

The three Graduate Chairs during my time in the Department of Languages and Cultures: Prof. Yasaman Rafat, Prof. Alena Robin, and particularly Prof. Constanza Burucua, whose guidance during the most crucial moments of my program was extremely valuable and for anchoring my feet to the ground when needed.

The Graduate Assistant Sylvia Kontra. I remember when we were first introduced the Department Chair's words were "She is the person who makes everything works". Sylvia made it possible to focus on my work while taking care of everything else. I always felt supported, which is so appreciated.

Professor Ana García-Allen who played a significant role as part of my emotional support system away from home. I have always felt her like a family member and a friend. I admire her professionally and personally. She has helped me stay focused and sane during stressful and challenging times, always with a smile and positivity.

My colleagues from CulturePlex, especially Bárbara Romero Ferrón and Zeina Dghaim, two members of my emotional support system. They brought laughs and filled empty spaces usually reserved for family. I thank them for reading multiple versions of my thesis and providing valuable feedback.

Lastly, and most importantly, I want to thank my wife Yadira Lizama-Mué, who is also a Ph.D. Candidate in our program. During this time, we have overcome several challenges such as starting a new life, a new language, increasing our family, the challenges of graduate



school, and a global pandemic. We maintained sanity and moved forward as a team. It was your support that made this thesis project possible, and I will always be thankful for you.

# Table of Contents

Abstract.....	ii
Summary for Lay Audience.....	iv
Co-Authorship Statement.....	v
Dedication.....	vii
Acknowledgments.....	viii
Table of Contents.....	x
List of Tables.....	xiii
List of Figures.....	xiv
List of Appendices.....	xvi
Chapter 1.....	1
Introduction.....	1
1.1 Definition of the Thesis’s Research Problem.....	1
1.2 Social Context of the Research Problem.....	4
1.2.1 Notions of Justice within Political Philosophy from Ancient World to Modern Era.....	5
1.2.2 Contemporary Theories of Justice.....	12
1.3 Engineering Context of the Research Problem.....	19
1.4 Methodological Approach to the Research Project.....	31
1.5 Works Cited in this Chapter.....	33
Chapter 2.....	39
Analysis of the Principled-AI Framework’s Constraints to Become a Methodological Reference for Trustworthy-AI Design.....	39
2    39	
2.1 Introduction.....	39
2.2 Related Research.....	42
2.3 Method.....	44
2.4 Results and Discussions.....	47
2.4.1 Semantic Similarity, Lexical Diversity, and Other Length Measures of the Documents in the Principled AI International Framework.....	52
2.4.2 Principled AI International Framework’s Text Analysis.....	55
2.4.3 Analysis of Proposed Principles.....	56
2.4.4 Analysis of Guidelines to Implement the Principles.....	60
2.4.5 Topic Modeling on the Principle’s Enunciation.....	63
2.5 Conclusions of the Chapter.....	66
2.6 Works Cited in the Chapter.....	68
Chapter 3.....	71
Principled AI Engineering Challenges Towards Trustworthy AI.....	71
3    71	
3.1 Introduction.....	71
3.2 Related Research.....	73
3.3 Method.....	77
3.4 Principled AI International Framework Modularity. Document Level.....	78
3.5 Principled AI International Framework’s Modularity and Derived Centrality Measures: Principle Level.....	79
3.6 Analysis of Relevance Across Principles’ Networks.....	86
3.7 Principles as Methodological Reference for Software Engineers.....	91

3.7.1	Analysis of the Principles Represented by the Single-Node Classes in the Degree Centrality Network .....	92
3.7.2	Analysis of the Principles Represented by the Page Rank’s Top Scored Nodes .....	94
3.8	Conclusions of the Chapter .....	104
3.9	Works Cited in the Chapter.....	106
Chapter 4	.....	109
Discrimination, Bias, Fairness, and Trustworthy AI	.....	109
4	109	
4.1	Introduction.....	109
4.2	Related Research.....	111
4.3	Analysis of the Variable Discrimination.....	112
4.4	Analysis of the Variable Bias .....	115
4.4.1	The First Point of Interest is Project Conceptualization .....	119
4.4.2	The Second Point of Interest is Project Design .....	119
4.4.3	The Third Point of Interest is Project Verification and Validation.....	121
4.5	Analysis of Variable Fairness .....	122
4.6	Analysis of the Variable Trustworthiness.....	124
4.7	Conclusions of the Chapter .....	128
4.8	Works Cited in the Chapter.....	130
Chapter 5	.....	134
Proposal of a Capability and Maturity Model for Trustworthy ADM Systems	.....	134
5	134	
5.1	Introduction.....	134
5.2	Related Research.....	136
5.2.1	Heavyweight Software Development Models .....	136
5.2.2	Agile Software Development Models.....	137
5.2.3	Feature-Oriented Software Development Models .....	138
5.2.4	Algorithmic Decision-Making (ADM) Systems’ Development Models	140
5.2.5	Quality Assurance of Algorithmic Decision-Making (ADM) Systems..	143
5.3	Method .....	154
5.4	Capability and Maturity Model for Trustworthy Algorithmic Decision-Making (ADM) Systems .....	156
5.4.1	Overview of the Model .....	156
5.4.2	Model’s Principles .....	157
5.4.3	Model’s Approach .....	158
5.4.4	Model’s Quality and Premise.....	159
5.4.5	Model’s Structure.....	159
5.4.6	Model’s Inputs and Outputs.....	161
5.4.7	Models’ Features, and Capability and Maturity levels .....	162
5.5	Examples of Domains in Which ADM Systems Have Proven Discriminatory Outcomes .....	178
5.5.1	Example 1: Health Care System .....	178
5.5.2	Example 2: Hiring Processes .....	179
5.5.3	Example 3: Recidivism Risk Assessment During Pretrial, Sentencing, and Paroling Assessment .....	180
5.6	A Brief Description of How the Proposed Capability and Maturity Model for Trustworthy AI Helps Mitigate ADM Systems’ Discriminatory Decisions.....	183

5.6.1	Transparency Feature.....	184
5.6.2	Security Feature .....	186
5.6.3	Bias Management Feature.....	187
5.7	Thesis Project’s Knowledge Mobilization Plan.....	189
5.8	Conclusions of the Chapter .....	192
5.9	Works Cited in the Chapter.....	194
	Appendices.....	201
	Curriculum Vitae .....	207

## List of Tables

Table 2-1: Ten most frequently used n-grams across all documents.....	55
Table 2-2: Ten most frequently used n-grams across principles. ....	57
Table 2-3: Ten most frequently used n-grams across the principles' declarations. ....	58
Table 3-1: Summary of principles based on the subnetworks' Weighted Degree distribution. .....	82
Table 3-2: Harmonic, Closeness, and Betweenness centrality measures. ....	87
Table 3-3: Page Rank and Eigen Vector centrality measures.....	89

## List of Figures

Figure 1-1: General Overview of Fairness Achievement in Artificial Intelligence [Own Elaboration, Based on (Chouldechova, 2017; Feldman et al., 2015; Fish et al., 2016; Hardt et al., 2016; Pedreschi et al., 2007; Solon and Selbst, 2016; Zafar et al., 2015)]. .....	22
Figure 1-2: Conceptualization of the Demographic Parity Technique [Own Elaboration, Based on (Chouldechova, 2017)]......	23
Figure 1-3: Conceptualization of the Supervised Learning Technique [Own Elaboration, Based on REF (Chouldechova, 2017; Feldman et al., 2015)]. .....	25
Figure 1-4: Conceptualization of the Calibration Checks Technique [Own Elaboration, Based on (Chouldechova, 2017; Hardt et al., 2016; Solon & Selbst, 2016)]. .....	26
Figure 2-1: Document and Author Types Distribution per Country [Own Elaboration]. .....	48
Figure 2-2: Author Type General Distribution [Own Elaboration]. .....	48
Figure 2-3: Document Type General Distribution [Own Elaboration]. .....	49
Figure 2-4: Document's Word Count and Lexical Diversity Relation [Own Elaboration]. ...	53
Figure 3-1: Communities Network of Documents Sharing Principles' Goals [Own Elaboration]. .....	78
Figure 3-2: Community Network of Principles Sharing Common Goals [Own Elaboration]. .....	80
Figure 3-3: Harmonic Centrality Network [Own Elaboration]. .....	87
Figure 3-4: Closeness Centrality Network [Own Elaboration]. .....	87
Figure 3-5: Betweenness Centrality Network [Own Elaboration]. .....	87
Figure 3-6: Page Rank Centrality Network [Own Elaboration]. .....	89

Figure 3-7: Eigen Vector Centrality Network [Own Elaboration]. .....	89
Figure 5-1: Simplified Model's Domain Conceptual Map [Own Elaboration].....	157
Figure 5-2: Model's Structure [Own Elaboration].....	159
Figure 5-3: Simplified Representation of the Model's Certification Process [Own Elaboration].....	160

## List of Appendices

Appendix A: List of documents included in the analyzed corpus to be referenced from chapter 2 as Principled AI International Framework.....	201
Appendix B: List of principles (Summarized based on the page rank's importance score).	203



# Chapter 1

## Introduction

This chapter describes the thesis' main research problem, which frames discriminatory decisions made or proposed by AI decision-making systems; conditioned through machine learning's (ML) limitations in avoiding the automation, amplification, and perpetuation of bias affecting historically marginalized population groups. This chapter also presents an analysis of the mechanisms ML uses to evaluate fairness in its algorithms, while exposes their weaknesses related to bias both in the data used to train algorithms, and in the actual algorithms. Lastly, it describes the thesis' general structure and research methods, which are detailed in the subsequent chapters.

### 1.1 Definition of the Thesis' s Research Problem

The sustained development AI solutions have exhibited in the last few years has made it clear that automated decision support systems (ADM) can no longer be conceived as a set of transparent techniques and methods consuming certain input parameters to be later processed in arriving to certain estimations. Nowadays, because of the high specialization ADM solutions can achieve, thanks to the advances in the machine learning field and according to the domains they are used in, ADM software can be perceived as complex systems, able to function through a self-learned undecipherable network of rules, usually called "black boxes" by engineers. As mentioned, the evolution of ML techniques may be responsible for this. After all, it is because ML algorithms were designed to find viable solutions on their own by identifying patterns in a training dataset which could be used for future implementation, that artificial intelligent systems (AIS) were able to be more "effective" in finding viable solutions given large data banks. However, ADM software's ability to learn and the responsibility that humans have outsourced to them does not assure their accountability.

In a sense, the impossibility of allocating accountability in ML algorithms and models, and the lack of consciousness in their reasoning processes are the main elements differentiating the problem of human-produced and machine-produced discrimination.

On the one hand, humans are able to reflect on their own biases to adjust their future decisions, acknowledging their will to do so, and are accountable for their actions. While on the other hand, ADM systems perpetuate their learned biases for as long as we use them before their produced discriminatory outcomes generate a social discomfort strong enough to prompt appropriate amendment efforts and cannot be held accountable for those outcomes. Yet, human-produced and machine-produced discrimination may be found to create an endless mutually dependent cycle in which the cycle's ultimate rupture could be determined by a mindful change in the human's set of shared values oriented to prevent discrimination to others. A change in the reality represented by the training datasets will permeate into the ADM learned patterns, and gradually support back the referred ideal human's set of shared values to prevent further discrimination.

ADM systems' learning abilities –as already stated–are based on their capability to spot patterns that invariably reflect pre-existing biases and discriminatory trends. Rather than benefiting and empowering as many people as possible, these abilities increase social injustices, or reflect only those represented in training dataset logs. Additionally, the fact that we outsource decision making to these software solutions erroneously diverts the accountability away from humans. As a result, people who have been negatively affected by automated decisions are unable to obtain explanations because the ADM software operator is unable to provide clarifications, nor receive remedies given that there is no specific information regarding the role of the variables included in the decision at hand. While the use of ADM is intended to benefit society, it is amplifying the injustices and social gaps vulnerable communities have historically experienced.

The unstoppable permeation of AIS in our society–increasingly supported by the automated processes–and the risk that entails its different communities, has led us to define the following as the research question for this thesis: how to reduce the discrimination against disadvantaged minorities that are a direct result of biased ADM systems?

Our goal is to design a capability and maturity model for trustworthy ADM solutions, incorporating currently available principles for fairer AI, from the regulatory field, while embedding them into a set of good practices across the software's lifecycle. This model will help reduce the discriminatory outcomes produced by ADM systems against equity seeking groups and support the construction of a fairer social contract.

Given the number of definitions that are available for the main working variables of the present research project we wanted to point out that we are using McCarthy's (2004) definition of AI as "*...the science and engineering of making intelligent machines, specially intelligent computer programs.*"; and Samuel's (1959) definition of Machine Learning as "*... the branch of artificial intelligence which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.*" Both definitions provides a simple and clear idea of the scope of AI and ML, useful in the context of this research project. Additionally, we use Castelluccia and Le Métayer's (2019) definition of Algorithmic Decision-Making Systems as "*systems that rely on the analysis of large amounts of personal data to infer correlations or, more generally, to derive information deemed useful to make decisions. Human intervention in the decision-making may vary, and may even be completely out of the loop in entirely automated systems.*" We decided to use Castelluccia and Le Métayer's definition as it is the working definition used by the European Parliament and, consequently, used by other government and inter-governmental organizations used for understanding the opportunities and challenges that those kind of systems presents.

Sections 1.2 and 1.3 below frame the social and the engineering contexts of the research problem. While the social context of the research problem explores the notions of justice in political philosophy as an approximation to fairness, the engineering context presented in section 1.3 describes the current approaches software engineers and ADM system developers are undertaking to build fairer solutions. By exploring context of the thesis's research problem, we can delve the elements that can be incorporated in the proposed model presented in chapter five.

## 1.2 Social Context of the Research Problem

The discriminatory decisions produced by ADM systems affect the perception of justice in societies using these solutions. Some AI and ML bibliography still argue that ADM systems only amplify existing discriminatory patterns yet insist accountability for the outcomes should be relegated to their source, apart from the ADM solution itself. Thus, there is insufficient recognition that the amplification of existent discriminatory patterns by ADM solutions demands a major revision of the mechanisms available to cope with such discriminations.

Certainly, there are dissimilar mechanisms in law and public policies which societies uses to avoid and mitigate discrimination. However, these must adjust to a new scenario, created by ADM systems, which are not audited, reproducible, or properly explained and yet used in decision making with a significant impact on society . It cannot be stressed enough the disadvantageous situation these systems create for the target population with respect to the fraction of people that profit from them by acting on the ADM systems' outcome or selling it to third parties.

Consequently, the research efforts to bring fairness to discriminatory ADM solutions must be seen as a matter of justice, where the trust among the actors is ensured through the technology design.

In the present research study, this justice is considered through regulatory norms created to provide a methodological reference for ADM software development and also for providing remedies for the discriminatory outcomes which have already affected people, that are analyzed in chapters two and three. It is also particularly considered from a fairness-based design perspective during ADM software development and deployment. Together, these efforts converge towards trustworthy ADM solutions through the systematization of the studied variables such as justice, fairness, and others explained in chapter four.

The following section shows the evolution of the notions of justice according to remarkable thinkers in political philosophy.

### 1.2.1 Notions of Justice within Political Philosophy from Ancient World to Modern Era.

As mentioned earlier, this section presents an analysis of the idea of justice, specifically within the political philosophy field of study. This section shows the notions of justice from Confucius to Rawls, furthermore Rawls's critiques from feminism and Black movements, seeking to identify elements that can be incorporated into the proposed solution.

Confucius was first attributed the act of relating ethics and political order (Clements, 2008). Confucius's doctrine can be summarized into a series of mandates among which "*Love the people, improve it morally and provide the needed means for daily life,*" "*Cultivate personal virtue and constantly seek perfection,*" and "*Having universal peace and general harmony as ultimate goal,*" separate themselves from the remaining four mandates (Confucius, 2003). The study of variables related to justice and fairness in the context of ADM systems exemplifies efforts to empower people, by creating mechanisms to improve available technologies and their development process from an ethical perspective, what aligns with Confucius's mandates and ultimate goal. The moral teachings of Confucius (Yu, 2009) are based on ren (*jen*) which is the human virtue founded on benevolence, loyalty, respect, and reciprocity between a dominant and dominated party, where the main characteristic of the dominant-dominated dichotomy rises as the former's obligation for protection towards the latter. It particularly important for this study, as it highlights a disfavored population sector discriminated at times unknowingly by ADM systems used by a dominant party without the means to identify or remedy the negative outcomes.

Socrates is considered the founder of the western political philosophy. His teachings explore the relation between knowledge and a just society and supports most of what is denominated systemic philosophical analysis (Platón, 2003) and the pursuit of truth. Socrates' notions of justice were intrinsically embedded in adherence to the Law being the instrument to maintain harmony between all actors of society (Jenofonte, 1993), establishing the guidelines and the proper framework for individuals and institutions, and

their interaction. According to Plato (Dodds, 1959), Socrates' dialogue with Calicles,<sup>1</sup> defends that every intervention in the public life must be done for good. This is also a shared goal within the Principled AI International Framework "AI for good" and it is the motivation for this thesis research project when aiming to influence the design of fairer ADM systems by improving the engineering approach and calling for an update of the quality features the software products and development processes are measured and assured with, based on the revision of emerging regulatory and standardization mechanisms.

In his work "Republic" (Platón, 2003a) Plato divides the ideal society into classes, for everyone to adhere to their social contract when building justice. Every social class has a specific responsibility in Plato's social model: the bottom class formed by farmers and trade class workers, the middle class formed by guardians and warriors, and the upper class gathering a superior type of guardians "the philosophers." Both the middle and the upper classes commonalities, with the upper only distinguished for its wisdom, better fitted to govern. Plato expands on "Socrates' class divisions for justice" on his dialog with the sophist (Platón, 2003b) stating that knowledge and science are above the Law and defending an authoritarian figure of government. This contrasts with Socrates' democracy and the recognition of the governed to think and vote for themselves. Plato's revision of "Republic" with the book "Laws" (Platón, 1999) posed no change to his idea of justice through a society well-governed by a -in Plato's words- "*soft, wise tyrant*".

Plato's ideas of justice transcend the ancient world into medieval Christianity through St. Augustine (Schall, 1998). Saint Augustine's definition of justice has two dimensions according to Chambers (Chambers, 2018): (1) a political notion of justice, similar or influenced by the ancient Greeks, understood as giving other human beings their due of social and political goods, and (2) the notion of theological justice referred to as "*iustitia*" in "*Augustine on Justice: A reconsideration of City of God, Book 19,*" and

---

<sup>1</sup> There is no certainty that Calicles was an actual parson. It is thought that he was a fictional character invented by Plato to embody all the injustices and opposition committed against Socrates.

which portrayed justice as the individual responsibility to be in the right relationship with God. The theological notion of Saint Augustine's justice served him to identify those illegally insisting that Rome was a pagan republic. The former notion of St. Augustine's justice was subdued to the latter under his belief that there was not the one without the other, what he called righteousness.

Perhaps the most influent political philosopher of medieval Europe was St. Thomas Aquinas, with his scholastic teachings, influenced by Aristoteles, Plato, and St. Augustine (Aquinas, 1993). Like St. Augustine, St. Thomas Aquinas provides a Christian superiority perspective to the notion of justice, as a superior justice. He defines justice (Aquinas, n.d.) as a virtue of rational creatures, men's habit of rendering to each his due by will, founding the complete structure of good works. The Aquinas notion of justice therefore arises as an intrinsic principle guiding good action and is therefore subjective. Then the quality of such justice remains in the object of it, which is the will.

Aquinas' defines two forms of justice, commutative justice in individual-individual scenarios and distributive justice in individual-community scenarios. Particularly, the distributive justice is based on the proportional distribution of common goods, and according to Cullen (Patrick, 2015), this no longer represents a contemporary problem. However, the challenges that emergent technologies represent to our societies' current notions of justice might show otherwise. In his report, Cullen concludes by pointing that human law may enter in conflict with justice, and those administrators enacting the Law may therefore contradict Aquinas's principle of proportional justice for an orderly society. Within the limits of the present research project, that can be noticed through the inability of the current legal framework to mitigate, avoid, or remedy discriminatory decisions produced by ADM solutions, thus making it a contemporary distributive justice issue. Laws are allowing big technology companies to harvest an individual's personal data for their own political benefit and profit, while subjecting people to unconsented clustering and classification given their personal characteristics and habits. Thus, the laws are no longer abiding the citizenry's conscience and the pre-existing order of society has been compromised.

In Europe during the Renaissance, Machiavelli's "The Prince" stands among the more influential secular political philosophy works (Johnston, 2002). In "The Prince," Machiavelli presents a manifest for a strong central power as the only means to avoid chaos in the social order. He takes a slightly different path from his predecessors as he does not entirely believe in divine justice (Copleston, 1999). His ideas of justice are found in the premise of law and order, and every new prince has the capacity and has an obligation to bring order where there is chaos. It is notable that the figure of power in Machiavelli's work, like his predecessor's, also stands above the law, this time, not using superior knowledge, or respecting a superior being, but by the implementation of brute force. Using an Aristotelian distinction per Parel (1990), it can be said that Machiavelli was not concerned with an idea of justice as it pertains to internal virtue or disposition of the soul, but with the notion of justice as it pertains to external acts of local law administration, defensive and offensive necessary wars, and the expansion of the empire. In the fifth part of "Allocutions" (Machiavelli, 1990, 1531) Machiavelli acknowledges that both an action and its intention must be just. Considering this theory, societies must therefore commit to develop trustworthy AI so that its actions are also just.

Thomas Hobbes shared Machiavelli's idea of the need of a robust and strong central power for maintaining order, and therefore justice (Copleston, 1999) in the early XVII century English Renaissance. Hobbes believed there was not such things as justice and injustice in a natural state (Green, 2008), and supported Machiavelli in viewing the human impulse for good being representative of the need for self-preservation. Hence, according to Hobbes (1980), injustice takes place when there is a discrepancy within an artificial order which was inexistent in the natural state formerly settled to rule human/human and governed interactions, guarded by power. This is the foundation of his theory of the "Social Contract." In Hobbes's social contract, everyone has rights, and to those, they also have proportional duties.

Aquinas's theory of commutative and distributive forms of justice can be seen in Hobbes's social contract based on what each part has agreed to upon entry (Olsthoorn, 2015). Both aspects of justice, commutative and distributive, are met because from the commutative perspective both parties exchange what was initially agreed upon and from



the distributive perspective as both parties comply with an established contract there is no injustice even if there is a difference in value between the goods exchanged. The righteousness and virtue are manifested by entering the contract. When transferring these ideas from the market to the social order, it can be inferred that governors play a referee role in the commutative expression of justice where citizens promise to abide by the rule of the sovereign for their interactions to be just. Then, trust and virtue are gained by the ruler when they adequately apply distributive justice. Therefore, in Hobbes's views, commutative and distributed expressions of justice correspond to individuals and governors, respectively. Hobbes shifted the existent notion of justice, from being dependent upon a divine source towards being an independent concept, turning justice and fairness into an agreed upon abstraction expressed by willingly entering into a contract.

In Hobbes's times, social justice was more in tune with equitable merit-based distribution of an individual's common properties and goods according to their value in a market-based society. Market forces have drastically evolved since Hobbes's times, and now some actors are now profiting from goods such as personal data and habitual information. In the context of emerging technologies, particularly ADM systems, it could be argued that the "use agreements," "privacy policies" etc. stand as just contracts, however, justice is compromised when unilateral contracts are imposed in order to receive a service.

John Locke and Jean-Jacques Rousseau, during the European Enlightenment period, had different opinions on Hobbes's social contract theories, and therefore different approaches to justice. Locke considers the scenarios in which no supra-authoritative figure exists, where a community must fill the gap in power; he defines two processes in the construction of the social contract (Powell, 1996): (1) contract for the society creation, where the community is built superseding the state of nature; and (2) contract of the government creation, where the relation between governor and governed is determined. Alternatively, Rousseau (1762) defends the argument that men willingly surrender their liberties in exchange for greater benefits inherent to life in the community.

Locke's second treaty of government (1690) presents one of the fundamental principles of political liberalism, which highlights the need for consent to be governed by an upper power. He explains that per foundation of a political society, citizens are obligated to accept the decisions of a majority. Legislature, chosen by the people, is the mechanism proposed by Locke to express and enact the decisions of the majority. Legislature also coexists with other powers like the executive and federative, so its own power is not absolute to maintain the law of nature as a permanent standard and to provide protection against arbitrary authority.

Rousseau (1959) similarly defends that all man are born equal and opposes the right of the strongest stating that men are only obligated to obey legit powers.<sup>2</sup> This poses an interesting point in contrast with the actual power of big tech companies like Google, IBM, and Facebook and their role in contemporary society. Rousseau's social contract underlines the notion of general will to solve common issues<sup>3</sup> of the social fabric, like what can be noted in the principled AI international framework mapped (Fjeld et al., 2020).

While Locke's justice is inconceivable without considering the right to property, Rousseau's justice is determined through the function of general will being defined as societal common interest. Both notions of justice, within the scope of the current research project may find a practical extension: the former, in the necessary recognition of individual's right over their personal data and digital information as private property, and the latter, in the determination of legal mechanisms to avoid big tech companies that benefit violate that ownership.

In capitalism, it is believed that every person gets no more, and no less, than what he or she gains via voluntary association with others. In a capitalist society justice is founded on the premise that all individuals are considered equal under the law (Reisman,

---

<sup>2</sup> First book of social contract page 8.

<sup>3</sup> Fourth book of social contract page 107.

2019). However, it is known this premise is not totally achieved in practice as the hollowness of the equality claim is vastly criticized in capitalist societies like the United States of America (USA) (Riley, 1989) where citizen characteristics like socioeconomic background, race, and gender influence their access to certain occupations, just to provide an example. The economic gap, apparently inherent to capitalist societies, constitutes the main element according to Isbister (2001), used to criticize this capitalist notion of justice.

Socialist ideals in Marx's and Engels's communist manifesto (1848) sustain that the abolition of bourgeois property and family structure is a fundamental requirement for building a society which accords with the political ideal of economic equality, hence social justice. In socialism, social justice can be achieved, first, by ensuring the distribution of social goods conform to the principle "*from each according to his/her ability, to each according to his/her contribution,*" and then, "*from each according to his/her ability, to each according to his/her need.*" Once society has achieved a higher level denominated "communism," where the need to work acquires a whole new dimension in a context characterized by democratized workplaces and social ownership upon the means of production.

The fascist justice of Hitler's Germany, Mussolini's Italy, Stalin's Russia, and Pinochet's Chile portrayed by Tewari<sup>4</sup> (2019) shared the subjugation of justice to the nation's interest. The nation's interest being the totalitarian idea of the leader's will. In contrast, the liberal ideal for justice stands for respect of individuals and associations. According to Barnett<sup>5</sup> (2000), the liberal notion of justice is comprised of three elemental rights: (1) the right to acquire, possess, use, and dispose of scarce physical resources, (including one's own body as while most property rights are freely alienable, the right to

---

<sup>4</sup> Manish Tewari is the former Union Minister of the Government of India, is also a lawyer and member of parliament.

<sup>5</sup> Part One: The Problem of Knowledge, sections four and five The Liberal Conception of Justice and Communicating Justice: The Second-Order Problem of Knowledge, respectively.

one's person is inalienable); (2) the right of first possession over unowned resources acquired by being the first to establish control of them; and (3) the right for freedom of contract specifying that a rightsholder's consent is both necessary (freedom from contract) and sufficient (freedom to contract) when transferring alienable property rights.

It is important, before exploring the contemporary notions of justice, to summarize the different notions already presented. For the ancient world, the notion of justice is linked with adherence to the laws, and it is founded on the domain of knowledge. Medieval Europe added a Christian approach by means of the acceptance of divine justice as an expression of the individual's will. The Renaissance framed justice by means of law and order, justifying the use of force given the changes of political and demographic divisions within the resultant societal evolutions. This was before European Enlightenment linked justice to rights over properties, and general agreement upon matters affecting the community. The notion of justice lastly evolved, during the Modern Era into a more complex idea through renovated notions of private and social property, including rights and liberties being representative of its principles.

### 1.2.2 Contemporary Theories of Justice

The most prominent contemporary notions of justice were defined by Rawls. Rawls's theory of justice has experienced several revisions since its initial publication in 1971. In his book "Theory of Justice" (1971) Rawls describes a "veil of ignorance," necessary for decision-makers to be unbiased and fair, as they would have no information on race, gender, religion, or any other variables when determining what is just and fair when judging others. The veil of ignorance is like the principle of unawareness that is criticized later in this chapter as one of the limitations of ML to cope with algorithmic bias. The implementation of the unawareness principle in ML shows that rather than conditioning fairness it enables other kinds of discriminatory decisions that are explained in the second part of the present chapter.

Rawls's theory of justice focuses on distributive justice, departing the traditional approach that he denominates "allocative justice." He centers his attention to organizing the basic structure of society, in his understanding of true distributive justice. He explores

the instrument of social contract, influenced by Locke and Rousseau, with the difference being “original position” achieved through the “veil of ignorance.” The original position” is a hypothetical situation embodying a mental experiment that can be described with two principles: (1) each individual has the right to an equal set of liberties to the same extent in regard of other individuals; and (2) social and economic inequities must be solved in a way that benefit most the least favored members of society, denominated “the difference principle.”

A first revision, or expansion, of Rawls’s theory of justice can be found in his book “Political liberalism” (1993). Rawls uses this first revision to explain why his theory aligns with a liberal definition of justice. The two principles supporting his notion of justice, he argues, conform a theory of legitimacy and stability in what he defines as a functional “overlapping consensus.”

The idea of overlapping consensus is needed to explain how a plural society, with diverse world viewpoints may achieve an agreement on what is fair. Additionally, the theories of legitimacy and stability behind the desirable overlapping consensus help to visualize the role of the state as a distant mediator in pluralist ideal societies. The overlapping consensus, resultant of a healthy public debate, also exposes the grounds for his doctrine of public reason by means of the reciprocity principle. Rawls’s doctrine of public reason broadens the reciprocity requirement for a plural society seeking overlapping consensus, clarifying that when citizens need to explain their political decisions to one another they must be able to justify their political decisions using publicly available agreed upon values and standards (1999).

A last review of Rawls’s theories of justice can be found in his work “Justice as Fairness” (2001). Rawls understood legitimacy as a mere standard of moral acceptability without a just political order. In Rawls’s reasoning, justice sets the optimal standard: the arrangement of social institutions that is morally best, while he argues that justice as fairness is superior to utilitarianism as the dominant tradition in modern political thought. However, his theories consider an ideal society without distinctions of race, gender, religion, and others and with birth and death as the sole means of entry and exit. This

ideal environment imposes still several challenges before considering Rawls's conception of justice.

The intention behind addressing unfairness and discriminatory decisions produced by ADM systems from the public policy with support of the International Human Rights Law described in chapters two and three, can be interpreted as an illustration of an attempted implementation of Rawls's ideas of justice. These chapters highlight the challenges both policy designers, and software engineers are currently facing and are further articulated in the proposed model described in chapter five.

Other justice theories coexisting with Rawls's include naturalist, feminist, and anti-colonialist conceptions of justice. The naturalist conception studied by Machan (1975) exposes a view of justice connected with human nature and behavior, and its relationship with human rights. Machan's naturalist conception of justice follows a similar approach to the one described earlier. The feminist and anti-colonialist perceptions of justice are explored later as part of the criticism to Rawls's justice theory.

The criticism to Rawls's theories of justice can be explored from three angles (Daniels, 1989): (1) the interpersonal character of Rawls's social contract, which leaves aside the bargain interests of other actors like future generations and nature itself which are not direct signatories of the current contract; (2) the divorce between Rawls' ideas of desirable equitable distribution and the self-optimization tendency to maximize efficiency of any government scheme, either capitalist or liberal (Kelly, 2013); and (3) the principle of difference derived in premises like "the original position" and "the veil of ignorance," which impacts the ability of individuals to establish historical principles when bargaining, therefore compromising the redistributive perception of justice (Nozick, 1974).

Other studies like Walzer and Wolff (1983; 1977) criticize Rawls's ideal societal structure which could only be possible if such society could be built from scratch. Additionally, feminist theories (Okin, 1989) emphasize Rawls's work limitations to include unfairness inherent to familiar relations as he focuses strictly on the basic structure of societies, failing to acknowledge unjust patriarchal social relations and a marked gender segregation in the labour market.

On the one hand, Mills' racial contract (1997) presents an anti-colonialist conception of justice exposing racism on the social contract of Locke, Rousseau, and Hobbes, and establishing that contemporary philosophers such as Rawls himself took their own white privilege for granted.<sup>6</sup> According to Cohen (1999) Mills' racial contract criticizes Rawls's social contracts as a model of relations among white individuals. Additionally, Mills' racial contract highlighted the gap between interactions of white and non-white individuals as a cause for systematic oppression of the former over the latter.<sup>7</sup> Overall, Mills' racial contract stresses the white supremacism of Rawls's ideal society based on the hypothetical utopian removal of all root causes of contemporary injustices for it to function.

Alternatively, the feminist Carole Pateman (1988) criticizes Rousseau's social contract's patriarchal conception of justice due to its proposed class reorganization within modernity. For Pateman, Rousseau's social contract's true pact consists of an agreement among men to distribute their access to female fertile bodies, accentuating inequalities between genders through lower salaries, gender violence, sexual harassment, etc. Like Mills, Pateman argues that Rawls's theory of justice is unable to recognize gender gaps and include them in his perception of justice.

In developed societies, like Western Europe, USA, or Canada, the different analyzed contemporary social contracts manifest themselves depending on the observer's angle, with several imperfections. Some transatlantic studies like those of The Economic Commission for Latin America and the Caribbean (ECLAC) (2021) and the Department of Economic and Social Affairs (2013; Engerman et al., 2002) identify that the Hispanic population within the studied societies are among their least favored sectors, whose inequalities deepen with added layers of race and gender. The next sections explore

---

<sup>6</sup> Mills, Charles W. (1997). *The Racial Contract*. Cornell University Press Ithaca and London. pp. 3–4.

<sup>7</sup> Mills, Charles W. (1997). *The Racial Contract*. Cornell University Press Ithaca and London. pp. 1–2, 5, and 7.

additional elements that feminism and the Black Movement contribute to the discussions of justice.

### 1.2.2.1 Justice From the Perspective of Feminism

According to Haslanger, Tuana and O'Connor (2012) there exist multiple dimensions within feminism ideologies,<sup>8</sup> which are focused mainly on the equality among men and women. Depending on the chosen bibliography, feminism can be chronologized in three or four waves. Studies using the three-wave classification understand the third wave as a continuous effort to fill gaps from the previous wave and includes the third and fourth waves as part of its counterpart classification. In this study, the analysis will be focused on the motors that moved the feminist need for justice regardless of any particular chronology.

During the Modern Era (XV-XVIII) and the European Illustration (XVIII- early XIX), women were excluded from the notion of citizenship (Amorós, 1990) so their first movement towards feminine justice was to revindicate their role in society, in terms of equality along men. It was only partially achieved in late 1800s and early 1900s when women started to conquer their right to knowledge and education, and consequently to vote and to hold property (Freedman, 2003). The different paces at which feminism has retaken women's rights around the world is in conjunction with the universal declaration human rights gave women to defy patriarchy and demand their inalienable rights as human beings to life, education, self- determination, equalitarian family roles, access to political and social life, access to jobs and equative salaries, etc. (Gillis, et al., 2007; Tong, 2009;). Incomprehensibly, these rights are yet being denied, on many different levels to women in many cultures around the globe (Duggan & Hunter, 1995; UN Women, 2019).

---

<sup>8</sup> According to Haslanger there exists multiple modalities of feminism including but not exclusive to cultural feminism, liberal feminism, radical feminism, ecofeminism, anarcho-feminism, feminism of difference, gender feminism, equality feminism, Marxist feminism, socialist feminism, separatist feminism, philosophical feminism, Islamic feminism, and lesbian feminism.



### 1.2.2.2 Justice From the Perspective of Black Movements

With the conclusion of the U.S. Civil war that put an end to slavery, a period of struggle started for the Black-American population in matters of civil rights. The social context within the 1950s and the 1960s for the U.S. black lives conditioned the surge of the Movement for Civil Rights, with the purpose of advocating and demanding black individuals be equal to white individuals before the law. In response, the 14th Amendment to the U.S. constitution (Amendment XIV, n.d.) gave black individuals their right by birth to obtain citizenship. This was previously out of reach for them given the jurisprudence in Dredd Scott's case (Dred, 1857). The 15th Amendment to the U.S. constitution (Amendment XIV, n.d.) granted them the right to vote, in response to the pressure made by the Black Power movement. However, the Voting Rights Act of 1965, a century later, was needed in order for black voters to freely exercise that right.

The Black Power movement in the USA challenged the agreed upon social structures described in traditional and historically settled social contracts by embracing pride for Black identity and culture (African American Heritage, n.d.). The movement created spaces within cultural and political institutions so Black lives could be empowered socially and economically. The Black Power movement assisted the individuals in their pursuit for social justice by claiming the right to life without fear; and shortly after U.S. black individuals became an example for other minorities and disfavored populations (such as non-English speaker immigrants), and other groups marginalized due to their gender identity and sexual orientation.

Parallely, the Black Feminism movement in US brought the struggles of sexism into the racism struggle, fighting Martin Luther King, Jr's "thingification" of Black women's humanity (McGuire, 2010), alongside the notion of "Black Matriarchy" as the root of all sorts of social ills<sup>9</sup> within the black community while forgetting the actual

---

<sup>9</sup> Retardment of the community and its consequences like the racism appreciation of rape and homosexuality, for example.

reasons why Black women were more independent in contrast to their white counterpart (Coontz, 2011; Friedan, 1964; Taylor, 1998).

Lastly, the Black Lives Matter (BLM) movement, stands—as the movement itself states in (Black Lives Matter, n.d.)—beyond police extrajudicial executions of Black people and adopt a necessary intersectionality as the movement include Black homosexuals, transsexuals, people with disabilities, undocumented persons, people with felonies and conviction history, women, and other lives along the gender identity range, when demanding social equality and justice on their basic human rights and their dignity. The BLM movement takes a step further from the racial contract when it criticizes Rawls’s and Rousseau’s social contract by listing the examples<sup>10</sup> of why Black lives constitute one of the disfavored sections of American society, which they describe as a State of Violence (Garza, n.d.). It would be interesting to research how many of the injustices the BLM movement have been fighting are present in the data logs ADM systems like COMPAS,<sup>11</sup> typically use to be trained, and how many of them permeated as a bias when predicting the likelihood of a defendant’s recidivism while determining a courthouse sentence.

It is interesting to note the common element in Feminism and the Black movement struggles with the lack of access to citizenship being the main cause for all the problems they were originally facing. This suggests that citizenship might be embody the social contract. Hence, the premise for fairness can be understood as follows: humans are all equals when they are citizens. Regardless, it evidences the sensitive situation of

---

<sup>10</sup> For the BLM movement USA is a State of Violence against the Black community because of the following facts: the majority of inmate population of the country are Black, the systematic racist police assaults against Black community members is founded in historical prejudice, Black homosexuals and transgenders are targeted by undervaluing fetichism from the heteropatriarchal generalized and protected standard of living, around 500, 000 undocumented immigrants are Black, several Darwinian experiments have targeted Black women, and other elements exposed by Garza in “A Herstory of the #BlackLivesMatter Movement.”

<sup>11</sup> Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software is a case management and decision support tool developed and owned by Northpointe (now Equivant) that is used by U.S. courts to predict the likelihood of a defendant becoming a recidivist.

immigrants, in contrast to the general population, before the agreed-upon social construct of justice.

### 1.3 Engineering Context of the Research Problem

The social and the engineering context of the stated research problem is mainly determined by the impact ADM systems have had in exacerbating the historical and systemic discrimination within our social fabric and pointing out how unjust the social order we still abide stands. The elements used to criticize Rawls's contemporary definition of justice as fairness (2001), like racial and gender inequities, for example, examined in the previous section, are amplified by present-day ADM's discriminatory outcomes. ADM's discriminatory outcomes are, in most of the cases, linked to human rights violations, therefore it is necessary to explore related regulatory available approaches along with the criticism of the ML's limitation to deal with bias and discrimination as disruptive elements to fairness, and per to Rawls: justice.

The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in ML systems (AccessNow Conference Declaration, 2018), the Montreal Declaration for a Responsible Development of AI (Université de Montréal, 2018), and the Statement on Artificial Intelligence, Robotics and Autonomous' Systems (European Group on Ethics in Science and New Technologies, 2018) all stressed the acknowledgment that ML discriminates individuals or groups and questioned who should be held accountable. Other studies identify the lack of target when considering moral, ethical and responsibility related issues (Varona, 2018). Matters like ADM system's failure to treat all individuals equally (fairness), the inability of ADM system's operators to explain the role of variables in the decision, and the lack of logs showing the variables permutations and the inability to interpret them if any (Explicability); the impossibility of having such systems audited by a third party (Auditability); and the violations to human basic rights manifested through the ADM system's discriminatory outcomes (Safety), to provide some examples, have diminished the user and general population's trust in ADM software-aided decisions.

According to specialized literature, the most frequent application areas in which predictive algorithms are used includes crime prevention, justice management, emotions analysis, crowd management, classifiers, and selection processes (Hardt et al., 2016; Zemel et al., 2013; Varona, 2018;), with the person being the object of analysis as the common factor. Some examples include flagging individuals with suicidal tendencies (Ayat et al., 2013), tailoring marketing strategies based on the estimation of people's sexual orientation (Walker, 2017), and automated assessment of a defendant's criminal potential (Sait Vural & Gök, 2017). Predictive algorithms are found to produce both false positives and false negatives which have misled decision makers in sensitive matters such as the freedom privation of an innocent individual, as described in next sections of this chapter. Current approaches to this matter are presented with a "reactive character." Dissimilar solutions and techniques are focusing on evaluation if a given predictive algorithm is balanced regarding its false positive/negative rates. This is a way to evaluate accuracy, rather than focusing on being more proactive at the design stage to achieve the desired fairness.

Fairness, according to the scope of the present research and Rawls, leads to justice. In Rawls's social contract an overlapping consensus is needed for an orderly society to function. Then, it is appropriate to believe that people need to trust in the social contract and the institutions for there to be justice. It can be seen, as evidenced in the previous paragraph, that ADM systems embody tools and methods that are permeating every sphere of the political, economic, and social fabric in modern society. Therefore, in the context of the present research, people's trust of ADM can be influenced through a set of related features which need to be proactively, and consciously managed during the ADM software project's lifecycle, to be deemed trustworthy.

To this, the pursuit of fairness should not be reactive using a given assessment technique but assured by orchestrating a process embedded—perhaps as part of the quality assurance process—through all development stages. It is important to highlight that there also exist some conceptual gaps imported from the AIS fundamentals regarding the individual inferences that are made based on the available information from a group with similar characteristics. It would be opportune to stress, in response to an argument

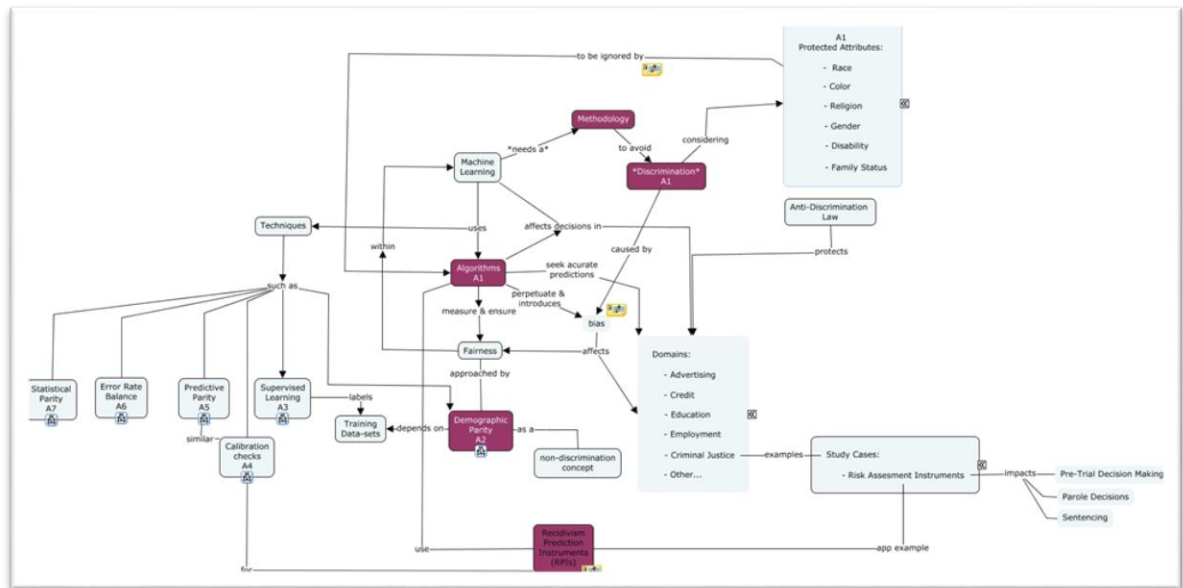
placing the focus on the group with similar characteristics to the target subject as a notion of fairness, that fairness should be dealt with from an individual case standpoint.

Research reviewed for this thesis (Chouldechova, 2017; Feldman et al., 2015; Fish et al., 2016; Hardt et al., 2016; Pedreschi et al., 2007; Solon & Selbst, 2016; Zafar et al., 2015) tend to analyze fairness in terms of “measure” (as per evaluation) and “assurance.” However, it must be identified that the concept of “assurance” is nothing else but “evaluation” when looking at the analyzed literature. This is evidence of the lack of proactiveness in the described methods. In other words, among the reviewed literature in the domain of related ML methods, the assurance of fairness will be achieved by means of evaluating the balance of a given algorithm in terms of the false positives and false negatives ratios produced. Therefore, the fairness assurance, in this case, is a reactive approach, showing a clear contradiction with the intention behind the “assurance” terminology. In these cases, assurance will not be found as the systemic organization of techniques and resources in an orchestra of processes and subprocesses or work stages pursuing to achieve a result with as much embedded “fairness” as possible.

Figure 1-1 shows a conceptual map that represents the general approach the studied research has towards achieving fairness in the context of AIS. As can be seen from the figure, all the techniques used in ML—among those studied in the references—to measure and ensure fairness by means of designed intelligent algorithms are reactive. The most used is the Demographic Parity technique, based on making decisions without considering protected attributes; this is closely followed by the Supervised Learning technique, which employs tagged datasets to train the algorithm. However, the Supervised Training Dataset technique has faced several criticisms among researchers due to two risks: being over-trained and introducing human prejudice when tagging the data logs.

Alternatively, “fairness assurance techniques” such as Calibration Checks, Predictive Parity, Error Rate Balance and Statistical Parity have been used much less, as evidenced in the studied bibliography. Perhaps these techniques are less popular as they focus on leveling the false positive and negative production rates in the pursuit of

balanced algorithms. Yet, a balanced false positive and negative production rate means the calibrated algorithm is equally biased, in favor and against each element within the context of the decision domain, not that the algorithm is fair. And finally, there are a scarce number of studies referring to unsupervised learning, and deep learning techniques.

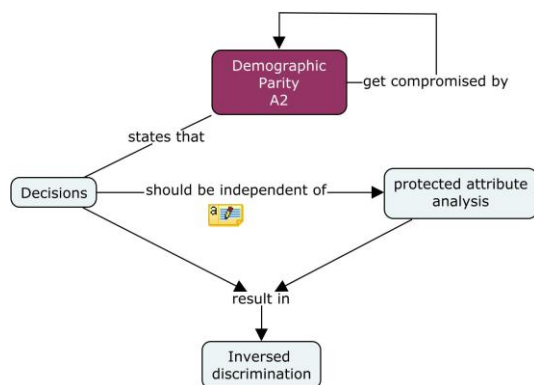


**Figure 1-1: General Overview of Fairness Achievement in Artificial Intelligence [Own Elaboration, Based on (Chouldechova, 2017; Feldman et al., 2015; Fish et al., 2016; Hardt et al., 2016; Pedreschi et al., 2007; Solon and Selbst, 2016; Zafar et al., 2015)].**

It is important to note that unsupervised learning and deep learning techniques can be considered variations of training techniques designed to overcome the risk of designing over-trained algorithms with supervised learning procedures. Figures 1-2 and 1-3, hereafter, exhibit a conceptual representation of the techniques more commonly used than the ones showed in Figure 1-1.

As can be seen in Figure 1-2, the main issue with the Demographic Parity technique is that it may lead to inverse discrimination, also known as white discrimination (Chouldechova, 2017; Dwork et al., 2011). This technique focuses on

leaving the protected attributes out of the analysis during the decision-making process, an idea that finds support in the literature (Feldman et al., 2015; Fish et al., 2016). As an example of how protecting given attributes might backfire and elevate the discriminatory gap among two population sectors, we would like to highlight the “Ban the Box” experiment detailed in (Cofone, 2019), where employers were banned from asking about criminal records on the application forms before the interview in the hiring process. In this case, the balance significantly favored white applicants over Black candidates because the former was perceived as less likely to have criminal records.



**Figure 1-2: Conceptualization of the Demographic Parity Technique [Own Elaboration, Based on (Chouldechova, 2017)].**

A similar approach to protecting attributes can result out of the use of Statistical Parity. According to Cofone (2019), when employers do not have all the information to measure individual performance, they tend to lean on knowledge learned from a—statistically tested—group of top performers to determine useful benchmarks to help them establish comparisons. This leads to erroneously overseeing the individual markers and, therefore, employers end up basing their decisions on discriminatory methods, compromising the employee’s self-development—in contexts such as promotions, and wages etc.—and contributes to institutionalizing and perpetuating bias in such contexts.

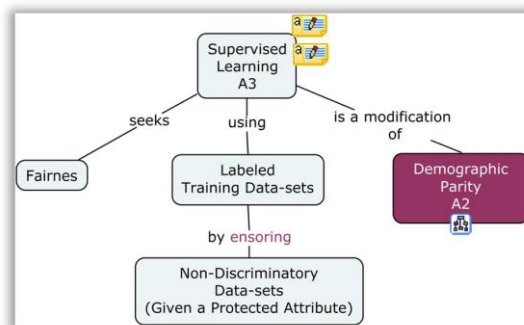
It seems the use of protected attributes defines if there will be a discriminatory decision as we know it or as inverse discrimination. Similarly, several studies (Mhasawade & Chunara, 2021; Mancuhan & Clifton, 2014; Žliobaitė & Custers, 2016) have demonstrated that some kind of discriminatory decision is made when attributes are

exposed to any deference, and even the slightest deference among them will affect the expected result. In consequence, we believe that having protected attributes should be avoided and that all attributes should be used equally, without any rank or distinguished weight. In addition, we disagree on the current idea of “protected attribute” regarding which attributes should be protected and in which context. We agree with the referenced studies when they identify the existence of certain—not so obvious—relationships between sensible and less sensible attributes, that come to light when protecting sensible and less sensible attributes triggers an equally discriminatory decision. Such is the case of attributes like “race” and “postal code.” While the postal code is not regulated as sensible by law because isn’t obvious when it triggers a discriminatory decision, it has been shown that it has similar outcomes to those conditioned by the race attribute in scenarios like financial loan applications and insurance premium calculations (Bathae, 2018).

We can all agree that the idea of having “protected” attributes came after—and as a response to— some discriminatory decisions made based on attributes. It is still arguable if the idea of having protected attributes is appropriate or not; or if it is just the execution of an idea still leading to discriminatory or inverse discriminatory decisions. The technique of Demographic Parity shows how a protected attribute-dependent decision may lead to a white discriminatory decision as well. In that same line of thoughts, it could be interesting to explore the effect that has for a given decision protecting all attributes that makes a person identifiable. Doing that might prevent the system to reach a decision at all, or else, it might allow the system to learn a set of completely different patterns to the ones resulting of the use of any attribute that makes a person identifiable. It would be interesting to explore the feasibility of those theses, and their results.

The Supervised Learning technique, shown in Figure 1-3, exhibits a variant to Demographic Parity which as stated in the literature (Chouldechova, 2017) will help overcome the risk of an inversed discriminatory decision. Chouldechova comes up with the idea of using a labeled training dataset to ensure nondiscriminatory (regular or inverse) decisions. One of the features of this technique is that it will cut in half the decision parameters to be considered.



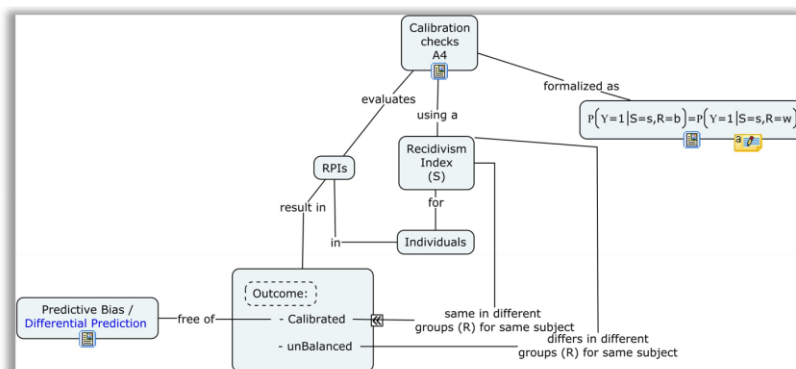


**Figure 1-3: Conceptualization of the Supervised Learning Technique [Own Elaboration, Based on REF (Chouldechova, 2017; Feldman et al., 2015)].**

When comparing the decision matrixes of both techniques (Demographic Parity, and Supervised Learning) it can be noted that the decision parameter somewhat grows instead of decreasing. The decision matrix for Demographic Parity can be found as a subset of the Supervised Learning’s decision matrix, which shows some unnecessary processing. And finally, it is our opinion that training the dataset as proposed in (Chouldechova, 2017) will refrain the algorithm from learning from some scenarios uncovered by the labeling procedures due to richness of real-life events. Our concerns find support on the following trepidation: All valid scenarios known up to the date (labeled as “I” in the decision matrix) will get affected somehow when any of the environmental condition changes.

The Calibration Checks technique can be seen in Figure 1-4; it aims at determining whether an algorithm is being discriminatory. The concept consists of a formalization of the balance of recidivism indexes as per Chouldechova (2017). However, what the calibration does is to determine when the algorithm is equally mistaken, in favor and against, the object of the decision. All these previously described techniques demonstrate that fairness should be attempted from the perspective of individual cases instead of using a group approximation because such an approach invariably leads to discrimination as it replaces the individual context with pre-established inferences, ignoring the new context. The accuracy percentages exhibited by predictive algorithms, usually oscillating between values from high 60s to low 80s

evidence the volume of predictions that can be explained by the error margin instead of the confidence interval used in the prediction.



**Figure 1-4: Conceptualization of the Calibration Checks Technique [Own Elaboration, Based on (Chouldechova, 2017; Hardt et al., 2016; Solon & Selbst, 2016)].**

It can be understood that both notions (individual and group approximations to the decision) are yet not discrimination free, but nevertheless, an individual approach provides a closer look at the subjects and contexts being targeted in the decision-making process. The Calibration Checks technique shows a reactive measure to verify whereas a given algorithm is free of bias in its predictions. The reactive character of the described techniques can also be found among less used techniques and methods such as Error Rate Balance and Statistical Analysis. It is important to highlight that evaluating algorithms as being non-discriminatory, or their degree of discrimination, is not the same thing as assuring algorithms are being fair.

A similar approach can be found in the literature (Pierson and Corbett-Davies, 2018), where the authors propose a method to evaluate whether discriminatory decisions based on a threshold tests exist, with applications in lending, hiring decisions, and policing in New York, USA. In this case, the method's focus of attention is not an algorithm, but the universe of decisions produced by that same algorithm. The method does not focus on evaluating algorithmic balance, but on assessing the discriminatory threshold exhibited among the algorithm's resulting decisions. It also adopts a reactive approach.

In contrast, there is some research which aiming to diminish, not eliminate, the intrinsic bias when algorithms operate with protected attributes. An example can be found in the literature (Cem Geyik et al., 2019), where the authors target in the hiring context the decreasing bias linked to a given set of attributes by having the distances between the candidate values and the mean value conformed by the universe of candidates as inputs for the ranking model. This method helps to deescalate the gaps among candidates, but it does not eliminate bias, nor does it provide a clear notion of how biased/unbiased an algorithm currently is.

Other authors like Gao and Shah (2020) show examples where different methods are used to assess the impact of dissimilar parameters on the resulting bias. In this case, the scenario is framed by ranking algorithms from web search engines, and the methods involved are Statistical Parity, and Disparate Impact Fairness. What we wanted to highlight from this study is that the experiment demonstrates our notions of bias differ depending on parameters such as urgency for, and context of the decision among others. Therefore, the notion of bias becomes even more subjective in such cases.

Unfortunately, the studied techniques and methods evidence the ML's limitations, over the past decade, to avoid bias and discrimination; and that the engineering approach followed during that period can be included, along with pre-existing bias in data, as a source for discriminatory ADM outcomes. That is why we think it is important to include in our study other approaches that might help to address the issue of trustworthy ADM.

It is our belief that AIS specialists should engage in a fairness-oriented development process, so that the final product of their processes may skip the calibration part *a posteriori*. More importantly, the AIS specialist's goal should be to deliver a product capable of not only proposing non-discriminatory decisions but also capable of adjusting its learning curve when conditions demand it.

While delving into the literature the rising concern in the last decade about human-related issues from AI systems' outcomes becomes evident. Every research resulting from these social, political, and technical concerns aim to provide partial or original solutions to achieve fairness when building algorithms (AccessNow Conference

Declaration, 2018; European Group on Ethics in Science and New Technologies, 2018). Many of these studies have made it clear that ethics, equity-seeking disciplines, and philosophy are becoming closer to the design field of AI solutions. Consequently, some researchers are pioneering proposals where philosophical constructions and analyses and artificial reasoning may converge. To this, the neutrosophic logic (Kharal, 2014; Mondal, 2015) may be a valuable tool to remain rooted within philosophical aspects of the AI solutions involving learning, fuzzy logic, statistical analysis, and the consequent decision making.

Neutrosophic logic is a discipline derived from neutrosophy as a branch of philosophy that studies the origin, nature, and scope of neutralities.<sup>12</sup> For the scope of this thesis, we can understand the notion of fairness (non-discrimination, unbiased, trustworthy) as the entity to explore neutralities. Neutrosophic logic uses fuzzy logic when dealing with entities that has no well-defined edges, such as the mentioned concepts, or like the degree in which the micro-environment of a single individual being target by an ADM system corresponds to the macro environment described by what that same ADM system learns from a group with similar characteristics. In that regard, neutrosophic logic may be valuable when determining neutralities in datasets and algorithms as a complement to the previously criticized algorithm calibration approach.

Other approaches are seeking support on the International Human Rights Law as exhibited in Fjeld et al. (2020). In the study, Fjeld maps a set of regulatory mechanisms, that numerous actor-like governments, intergovernmental organizations, and other stakeholders have proposed with the goal of establishing the foundations for the future of AI development within a framework of universally agreed-upon social, political, and

---

<sup>12</sup> Within the context of neutrosophy neutrality is defined as follows: considering an entity, which can be a concept “C”, and its opposite “!C”, neutrality refers to the entity that is neither the concept “C” nor its opposite “!C”. i.e., Let’s consider the following expression “It is discriminatory to say the subject “U” belongs to the array “A”, and its opposite “It is not discriminatory to say the subject “U” does not belong to the array “A”, then a neutral expression could state “It’s either discriminatory or not discriminatory to consider the subject “U” part of the array “A””.

moral principles, based on human rights. The study refers to that framework as “Principled AI.”

The Principled AI International Framework arises, as a complement—from the public policy area—for the software engineering attempts to mitigate ADM systems’ discriminatory outcomes. The idea of fairness, as criticized earlier, is conceived as the sum of techniques and methods oriented to measure and analyze false negative and false positive rates of intelligent algorithms by excluding properties from a given object of analysis when using those same techniques and methods. This poses a trend portraying a reactive approach towards the studied issue in the form of evaluation and measures applied to already designed algorithms. It is our opinion that assuring fairness in the context of ADM solutions, as in other contexts, should rather embrace a proactive approach; consequently, it becomes necessary to explore how trustworthiness might be planned from early stages of the development workflow.

In this final line of thought, we believe that, in the context of software development, fairness might be treated as a quality characteristic, a non-functional requirement.<sup>13</sup> By considering fairness as a quality characteristic or a non-functional requirement, it would be easier to incorporate assurance related activities into the software engineering process, in the initial stages. We also believe the proposed approach may help integrate the principles from the Principled AI International Framework mapped by Fjeld into the Software Development Process, especially when developing AIS. By doing so, we will be taking a step towards transforming the current reactive approach, exhibited in the analyzed literature, into a more proactive set of good practices, methods, and techniques in the hands of the AIS specialist in the form of a capability and maturity model for trustworthy AI.

---

<sup>13</sup> The software engineering term “non-functional requirement (NFR)” is used to categorize different constraints the software must comply with and that are not business functionalities and are usually managed as quality characteristics. An example of a non-functional requirement could be “The system must provide a response in less than 15seconds.” Some of the most popular categories NFR are sorted include usability, serviceability, security, scalability, interoperability, reliability, and maintainability among others.

Adopting a similar approach to quality assurance for trustworthiness assurance also facilitates the developers to embrace the idea of trustworthiness assurance more easily as they can establish parallels that help them overcome any possible initial resistance against the implementation of the model. Although desirable, it would not be possible to think that a given ADMS produces discrimination-free outcomes just because the development team followed the proposed model as part of the development process. Similar to the quality assurance, where the development team can only state that the software has the least possible number of defects so it produces the least number of errors (not that the software is error-free), it will only be possible to think that the ADMS is as free of bias as possible to produce the least number of decisions resulting in discrimination when following the proposed model on its development. This is helpful for the development team as its members will not find themselves in an endless and unsuccessful pursuit of perfection, when the variables of business they model determines it so, as the team can move forward to subsequent development stages with the least faulty ADM component amongst the available. These statements were integrated within the proposed model in form of specific practices focused on the re-use of ML components that had proven to produce less or non-discriminatory outcomes, or to use the dataset triggering the least number of discriminatory decisions for training the models, after those datasets have been assessed in the data pipeline.

A general analysis of the Principled AI International Framework, presented in chapter two, along with the ML limitations to deal with bias and discrimination presented in the present chapter, allow us to have a better comprehension of the thesis's research question's environment that was used to conceptualize the model's architecture. A deeper analysis of the principles proposed within the Principled AI International Framework, shown in chapter three, made it possible the determination the variables through which the goals of each capability level of the proposed model are orchestrated. The exploration of the variables like bias, discrimination, fairness, and trustworthy, exposed in chapter four and motivated from the analysis of the principles in chapters two and three, helped us defining the multi-dimensional scope of the proposed model, expressed in the definition of data, algorithm, and practice-oriented general objectives.

Consequently, we were able to propose an engineering model, exhibited in chapter five, that integrates both studied approaches, which were previously dichotomic up to reduce (short term) and eliminate (long term) social and ethical problems rooted in ADM systems. A brief illustration of the model application presented in the second half of chapter five, with the use of three different examples, helps visualize the model's implementation at different levels of capability and maturity.

Additionally, a parallel research<sup>14</sup> (Suárez & Varona, 2021) allowed us to know that Canadian and other university across 16 countries are not teaching the so needed ethical skills for future workers in the AI sector to safely, productively, and effectively engage with ADMS. This lack in training hinders the efforts to extend values of equity, diversity, and inclusion within the digital realm, and will constitute an element of resistance for ADMS development teams to implement the model. Accordingly, we designed a knowledge mobilization plan to extend the reach of the proposed model as a possible solution for the studied discrimination problem, and to support its assimilation and implementation by any interested ADMS development teams.

This thesis project has been conducted with a marked emphasis on the analysis of language. We replicated language processing methodology already used in another context (Suarez & Lizama, 2020) —transitional justice in the Colombian's peace agreements—given the proven benefits of this type of analysis when studying the social, ethical, and philosophical edges of a sensible problems involving policies and procedures affecting human beings. A description of the research methods employed during the execution of the present thesis project can be found in the following section.

## 1.4 Methodological Approach to the Research Project

The performed research methods included the execution of natural language processing (NLP) techniques like lexical diversity, semantic similarity, n-gram

---

<sup>14</sup> The ethical skills we are not teaching: An evaluation of university level courses on artificial intelligence, ethics, and society is co-funded by the Social Sciences And Humanities Research Council (SSHRC) and the Government Of Canada's Future Skills Program.

extraction, and topic modeling to identify elements within the principled AI international framework that can be used to influence trustworthiness from early stages of ADM systems development process. The method is further detailed in chapter two.

The use of network's theories maps the relation among regulatory documents in the Principled AI International Framework through their proposed principles, while performing dissimilar distance and degree-based measures to gain deeper insight into the elements that could be incorporated later into the model. The method is further detailed in chapter three.

The close reading method was used indistinctively to achieve all research questions addressed on each chapter. Specifically in chapter two, the method helped to further comprehend the intricacies of the modelled topics; in chapter three, it allowed us to delve into identified relations among principles in the Principled AI International Framework, while delving into the meaning of such modelled relations and deciding how to incorporate the principles—in form of specific practices—on the proposed model; and in chapter four, the method was particularly helpful as a tool for surveying the literature when determining how variables like bias, discrimination, fairness and trustworthiness could redefine the scope of functional quality variables, and later be incorporated as features of the proposed model. The close reading helped across all chapters in the definition of their theoretical frameworks.

Finally, the method when designing the proposed model uses a systemic approach while ensuring that the different pieces explored and described along chapters one to five were harmoniously connected as part of the model.



## 1.5 Works Cited in this Chapter

- AccessNow Conference Declaration. (2018). *The Toronto declaration: Protecting the rights to equality and non-discrimination in machine learning systems*. Retrieved from Access Now: <https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/>
- National Archives. (n.d.). *African American Heritage*. Retrieved August 2021, from <https://www.archives.gov/research/african-americans/black-power>
- Aquinas, Thomas. (1993). *Selected philosophical writings*. Oxford University Press.
- Aquinas, Thomas. (n.d.). *Summa theologiae II-II*. Benziger Brothers.
- Constitution Center. (n.d.). *Amendment XIV*. Retrieved from <https://constitutioncenter.org/interactive-constitution/amendment/amendment-xiv>. July 2021.
- Constitution Center. (n.d.). *Amendment XV*. Retrieved from <https://constitutioncenter.org/interactive-constitution/amendment/amendment-xv>. July 2021
- Amorós, C. (1990). El feminismo: Senda no transitada de la Ilustración. *Isegoria* (1): 139-150.
- Ayat, S., Farahani, H. A., Aghamohamadi, M., Alian, M., Aghamohamadi, S., & Kazemi, Z. (2013). A comparison of artificial neural networks learning algorithms in predicting tendency for suicide. *Neural Computing and Applications*, 23(5), 1381-1386. doi: <https://doi.org/10.1007/s00521-012-1086-z>
- Barnett, R. E. (2000). *The Structure of liberty: Justice and the rule of law*. Oxford University Press.
- Bathae, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, 31 (2), 889-938.
- Black Lives Matter. (n.d.). Retrieved from <https://blacklivesmatter.com/about/> . July 2021.
- Castelluccia, C. & Le Métayer, D. (2019). *Understanding algorithmic decision-making: Opportunities and challenges*. Panel for the Future of Science and Technology (STOA) of the Scientific Foresight Unit at Directorate-General for Parliamentary Research Services (DG EPRS) of Secretariat of the European Parliament doi:10.2861/536131
- Cem Geyik, S., Ambler, S., & Kenthap, K. (2019). Fairness-aware ranking in search and recommendation systems with application to LinkedIn talent search. *ArXiv*. <https://arXiv.org/abs/1905.01989v3>
- Chambers, K. (2018). Augustine on justice: A reconsideration of city of God, book 19. *Political Theory*, 19(5) 382-396 <https://doi.org/10.1080/1462317X.2018.1438781>

- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153-163. doi: 10.1089/big.2016.0047
- Clements, J. (2008). *Confucius: A biography*. Sutton Publishing.
- Cofone, I. (2019). Antidiscriminatory privacy. *SMU Law Rev*, 72(1):139–176. Retrieved from <https://scholar.smu.edu/smulr/vol72/iss1/11>. July 2021.
- Cohen, P. N. (1999). Book review: The social contract. *Review of Radical Political Economics*. 31 (2), 102–105. doi:10.1177/048661349903100208
- Confucius. (2003). *Confucius: Analects – With selections from traditional commentaries*. Translated by E. Slingerland. Hackett Publishing.
- Coontz, S. (2011). *A strange stirring: The feminine mystique and American women at the dawn of the 1960s*. Basic Books.
- Copleston, F. (1999). *A history of philosophy 3*. Continuum International Publishing Group.
- Daniels, N. (1989). *Reading Rawls: Critical studies on Rawls' "a theory of justice"*. Stanford University: Press.
- Department of Economic and Social Affairs. (2013). *Inequality matters: Report of the world social situation 2013*. United Nations.
- Dred, S. (1857). *Missouri state archives: Missouri's case, 1846-1857*. Missouri Digital Heritage. Retrieved from <https://www.sos.mo.gov/archives/resources/africanamerican/scott/scott.asp>. July 2021.
- Dodds, E. R. (1959). *Plato gorgias. A revised text with introduction and commentary*. Oxford University Press.
- Duggan, L. & Hunter, N. D. (1995). *Sex wars: Sexual dissent and political culture*. Routledge.
- Dwork, C., Hardt, M., Pitassiz, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. *ArXiv*, eprint arXiv:1104.3913, 1-24.
- Engerman, S. L., Sokoloff, K. L., Urquiola, M., & Acemoglu, D. (2002). Factor endowments, inequality, and paths of development among new world economies. *Economía*, 3(1), 41-109.
- European Group on Ethics in Science and New Technologies. (2018). *Statement on artificial intelligence, robotics and 'autonomous' systems*. Publications Office of the European Union. doi:10.2777/531856
- Feldman, M., Friedler, S. A., Moelle, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259-268.
- Freedman, E. B. (2003). *No turning back: The history of feminism and the future of women*, 464. Ballantine Books.

- Fish, B., Kun, J., & Lelkes, Á. D. (2016). A confidence-based approach for balancing fairness and accuracy. *Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics*, 144-152.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society, Harvard Law School.
- Friedan, B. (1964). *The feminine mystique*. W.W. Norton.
- Gao, R. & Shah, C. (2020). Toward creating a fairer ranking in search engine results. *Inf Process Manage* 57,1–19. <https://doi.org/10.1016/j.ipm.2019.102138>
- Garza, A. (n.d.). *A herstory of the #BlackLivesMatter movement*. The Feminist Wire.
- Gillis, S., Howie, G., & Munford, R. (eds). (2007). *Third wave feminism: A critical exploration*. Palgrave Macmillan.
- Green, M. (2008). Hobbes on justice. *Social & Political Philosophy*, 33.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *ArXiv*. doi:arXiv:1610.02413
- Haslanger, S., Tuana, N., & O'Connor, P. (2012). In E. N. Zalta (Ed.), *Topics in feminism*. The Stanford Encyclopedia of Philosophy.
- Hobbes, T. (1980). In C. Moya & A. Escotado (Eds.), *Leviatán*. Editora Nacional.
- Isbister, J. (2001). *Capitalism and justice: Envisioning social and economic fairness*. Lynne Rienner Publishers.
- Jenofonte. (1993). *Recuerdos de Sócrates; económico; banquete; apología de Sócrates*. Gredos.
- Johnston, I. (2002). *Lecture on Machiavelli's The Prince*. Malaspina University College.
- Kelly, P. (2013). *The politics book*. Dorling Kindersley.
- Kharal, A. (2014). A neutrosophic multi-criteria decision-making method. *New Mathematics and Natural Computation*, 10(2), 143-162.
- Lange, A. R. (2015). *Digital decisions: Policy tools in automated decision-making*. Center for Democracy and Technology.
- Locke, J. (1690). *Two treatises of government: An essay concerning the true original, extent, and end of civil government*. Black Swan.
- Machan, T. R. (1975). Law, justice, and natural rights. *Western Ontario Law Review*, 14, 119-130.
- Machiavelli, N. (1990). Allocution made to a magistrate. *Political Theory*, 18(4) 525-527. doi:10.1177/0090591790018004002
- Machiavelli, N. (1531). *The discourses*. Translated by Leslie, J. & Walker, S.J., revisions by Brian Richardson (2003). Penguin Books.

- Mancuhan, K. & Clifton, C. (2014). Combating discrimination using bayesian networks. *Artificial Intelligence Law*, 22, 211–238 DOI 10.1007/s10506-014-9156-4
- Marx, K. & Engels, F. (1848). *The manifesto of the Communist Party*.
- McCarthy, J. (2004). *What is artificial intelligence?* Computer Science Department, Stanford University retrieved from [https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems\\_2008\\_2009/Old/IntelligentSystems\\_2005\\_2006/Documents/Symbolic/04\\_McCarthy\\_whatissai.pdf](https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatissai.pdf)
- McGuire, D. L. (2010). *At the dark end of the street: Black women, rape, and resistance—A new history of the civil rights movement from Rosa Parks to the rise of black power*. Random House.
- Mhasawade, V., & Chunara, R. (2021). Causal multi-level fairness. *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society (AIES '21)* July 2021, 784–794 <https://doi.org/10.1145/3461702.3462587>
- Mills, C. W. (1997). *The racial contract*. Cornell University Press.
- Mondal, K. A. (2015). Rough neutrosophic multi-attribute decision-making based on grey relational analysis. *Neutrosophic Sets and Systems*, 7, 8-17.
- Nozick, R. (1974). *Anarchy, state and utopia*. Basic Books.
- Office of Probation and Pretrial Services. (2011). *An overview of the federal post conviction risk assessment*. Administrative Office of the United States Courts.
- Okin, S. M. (1989). *Justice, gender, and the family*. Basic Books.
- Olsthoorn, J. (2015). Hobbes on justice, property rights and self-ownership. *History of Political Thought*, 36(3) 471–498 <http://www.jstor.org/stable/26228629>
- Parel, A. (1990). Machiavelli's notions of justice: Text and analysis. *Political Theory* 18(4), 528-544. Sage Publications. <http://www.jstor.org/stable/191540>
- Paterman, C. (1988). *The sexual contract*. Polity Press.
- Patrick, C. (2015). *Aquinas on law and justice, conflict of human law and justice in the orderly society*. Southern New Hampshire University.
- Pedreshi, D., Ruggieri, S., & Tur, F. (2008). Discrimination-aware data mining. In: *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, 560–568.
- Pierson, E., Corbett-Davies, S., & Goel, S. (2018). Fast threshold tests for detecting discrimination. In *Proceedings of the 21st international conference on artificial intelligence and statistics*. ArXiv. arXiv:1702.08536v3
- Platón. (1999). *Diálogos. Obra completa. Volumen VIII: Leyes (Libros I-VI)*. introducción, traducción y notas de Francisco Lisi. Gredos.
- Platón. (2003). *Diálogos. Obra completa*. Gredos.
- Platón. (2003a). *Diálogos. Obra completa en 9 volúmenes. Volumen IV: La República*. Gredos.

- Platón. (2003b). *Diálogos. Obra completa en 9 volúmenes. Volumen V: Parménides. Teeteto. Sofista. Político*. Gredos.
- Powell, J. (1996). John Locke: Natural rights to life, liberty, and property. In *The Freeman*. The Foundation for Economic Education.
- Rawls, J. (1971). *A theory of justice*. Harvard University Press.
- Rawls, J. (1993). *Political liberalism*. Columbia University Press.
- Rawls, J. (1999). *The law of peoples*. Harvard University Press.
- Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.
- Reisman, G. (2019). Classical economics versus the exploitation theory. *Mises Daily Articles*. Mises Institute. Retrieved from <https://mises.org/library/classical-economics-vs-exploitation-theory>. July 2021.
- Riley, J. (1989). Justice under capitalism. *Nomos*, 31, 122-162. <https://www.jstor.org/stable/24219468>
- Rousseau, J.J. (1762). *Contrato social*. Espasa-Calpe.
- Rousseau, J.J. (1959). *Œuvres complètes*. Bibliothèque de la Pléiade. Gallimard.
- Sait Vural, M., & Gök, M. (2017). Criminal prediction using naive bayes theory. *Neural Computing and Applications*, 8(9), 2581-2592. doi:<https://doi.org/10.1007/s00521-016-2205-z>
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210-229.
- Schall, J. V. (1998). *At the limits of political philosophy*. CUA Press.
- Smarandache, F. (2015). *Symbolic neutrosophic theory*. Infinite Study.
- Solon, B., & Selbst, A. D. (2016). Big data's disparate impact. *CALIF. L. REV.*104, 671-732.
- Suarez, J. L. & Lizama-Mue, Y. (2020). Victims of language: Language as a pre-condition of transitional justice in Colombia's peace agreement. In *transitional justice in comparative perspective: Memory politics and transitional justice* 97-127. Palgrave Macmillan, Cham. [https://doi.org/10.1007/978-3-030-34917-2\\_5](https://doi.org/10.1007/978-3-030-34917-2_5)
- Suarez, J.L. & Varona, D. (2021). The ethical skills we are not teaching: An evaluation of university level courses on artificial intelligence, ethics, and society. *A Report to the Social Sciences and Humanities Research Council and the Knowledge Synthesis Grants Program*. Retrieved from [https://cultureplex.ca/wp-content/uploads/2022/01/Ethical\\_Skills\\_We\\_Are\\_Not\\_Teaching\\_Report.pdf](https://cultureplex.ca/wp-content/uploads/2022/01/Ethical_Skills_We_Are_Not_Teaching_Report.pdf)
- Taylor, U. (1998). The historical evolution of black feminist theory and praxis. *Journal of Black Studies*, 29(2), 234-253.
- Tewari, M. (2019). Oh justice! Raisina debates. Observer Research Foundation ORF blog Retrieved from <https://www.orfonline.org/expert-speak/justice-nazi-germany-54934/>. July 2021.

- The Economic Commission for Latin America and the Caribbean (ECLAC). (2021). *Afrodescendants and the matrix of social inequality in Latin America: Challenges for inclusion. Summary*. United Nations.
- Tong, R. (2009). *Feminist thought: A more comprehensive introduction (3rd edition)*. 289, 284-285. Westview Press (Perseus Books).
- Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Gonzalez Zelaya, C., & Van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 272 – 283.  
<https://doi.org/10.1145/3351095.3372834>
- UN Women. (2019). *Progress of the world's women 2019–2020: Families in a changing world*. United Nations. Retrieved from [progress.unwomen.org](http://progress.unwomen.org). July 2021.
- Université de Montréal. (2018). *Montréal declaration for a responsible development of artificial intelligence*. Université de Montréal.
- Varona, D. (2018). La responsabilidad ética del diseñador de sistemas en inteligencia artificial. *Revista de Occidente*, 446-447, 104-114.
- Walker, T. (2017, November 20). How much ...? The rise of dynamic and personalized pricing. *The Guardian*. Retrieved April 2018 from <https://www.theguardian.com/global/2017/nov/20/dynamic-personalised-pricing>
- Walzer, M. (1983). *Spheres of justice: A defense of pluralism and equality*. Basic Books.
- Wolff, R. P. (1977). *Understanding Rawls: A reconstruction and critique of "a theory of justice."* Princeton University Press.
- Yu, D. (2009). *Confucio para el alma o las claves milenarias para ser feliz*. Editorial Planeta.
- Zafar, M. B., Valera, I., Gomez Rodríguez, M., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv*. doi: arXiv preprint arXiv:1507.05259
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning* 28, 325-333.
- Žliobaitė, I. & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence Law*, 24, 183–201 <https://doi.org/10.1007/s10506-016-9182-5>

## Chapter 2

### Analysis of the Principled-AI Framework's Constraints to Become a Methodological Reference for Trustworthy-AI Design

As stated in the previous chapter, software stakeholders and other interested actors linked to ADM solutions development have been seeking to address the fairness issue from the regulatory and public policy perspective, in an attempt to overcome the technical limitations of machine learning already described in the second half of the referred chapter. Therefore, this chapter expands on the study of available efforts to solve the issue of discriminatory ADMS by analyzing the normative dimension of the thesis research problem framed by the Principled AI International Framework. This chapter shows the study of the documents defining the corpus of the referred framework identifying elements that may facilitate its adoption as a methodological reference by software engineers, especially, artificial decision-making solution developers. The chapter details an analysis of the mentioned corpus using natural language processing techniques and showcases language as the primary constraint for the framework assimilation as a methodological reference within the software development industry.

#### 2.1 Introduction

The engineering tactics seeking to solve social problems rooted in the use of technology, specifically in the use of AI and ML are as dissimilar as the problems they try to solve. Some examples of the mentioned tactics, detailed in the previous chapter, include algorithmic calibration from an engineering context, or the design of policies to normalize engineering procedures with the purpose of mitigating the resultant product's negative impact on society, from a regulatory context. These referred approaches also delineate an area of research that is increasingly attracting interest from the academic and professional communities.

In a short period of time, solutions to AI's social-related problems have evolved through different stages such as attempts to eliminate or diminish bias in data and algorithms (Mehrabi et al., 2019), efforts to achieve fairness (Mehrabi et al., 2019; Sahil & Rubin, 2018; Walker, 2017), and the most recently proposed set of principles for AI

design (Fjeld et al., 2020; Mittelstadt, 2019). The approach aiming to diminish bias and discrimination seeks to ensure that individuals or groups are treated equally in the context of a decision regardless of their attributes. In contrast, fairness focusses on managing a set of characteristics the software product, specifically AI and ML solutions, needs to comply with given the direct impact of those characteristics on individuals and groups targeted by the automated decision and in the implementation of amendments when needed. The latter approach seeks, through a set of dedicated principles, to explore the feasibility of using the International Human Rights Law<sup>15</sup> as a reference for the software development process, particularly when designing AI and ML solutions, in the pursuit of trustworthy AI. This recognized that the variables associated with most social problems stemming from the use of AI and ML are reflected in the corpus of law.

The importance of the principled AI approach, based on the International Human Rights Law, is found through the supplement it provides to overcome the current limitations of ML to deal with bias and discrimination criticized in chapter one. In her review (Fjeld et al., 2020), the author used methods like hand-coding and close equivalence to map the consensus in ethical and rights-based approaches to principles for AI to what can be denominated a Principled AI International Framework. Fjeld's proposal gathers international established policies whose authors have the agency and authority for implementation, in an abstract single regulatory mechanism. The selection criteria used by Fjeld for including AI regulatory initiatives, regarding the design and use of AI solutions in her study was mainly focused on collecting those policies proposing principles and guidelines for implementation. Her objective was to identify trends across the proposed principles and to uncover the hidden momentum in a fractured, global conversation around the future of AI. The analysis resulted in eight main themes: (1) Privacy; (2) Accountability; (3) Safety and Security; (4) Transparency and Explainability;

---

<sup>15</sup> The UN's Universal Declaration of Human Rights (UDHR), together with the International Covenant on Civil and Political Rights and its two Optional Protocols, and the International Covenant on Economic, Social and Cultural Rights, form the International Bill of Human Rights. For the purpose of the present thesis, when referring to the International Human Rights Law we include these already mentioned and the series of international human rights treaties and other instruments adopted since 1945 which have conferred legal form on inherent human rights and developed the body of international human rights.



(5) Fairness and Non-Discrimination; (6) Human Control of Technology; (7) Professional Responsibility; and (8) Promotion of Human Values. These themes can be used as categories under which a set of summarized principles can be arranged.

The mapping of the Principled AI International Framework constitutes an important step towards trustworthy AI from the field of policy making and regulations. However, it highlights a rupture in the dialogue of the principles across the involved actors. This perceived rupture motivated us to explore the original principles, using a different method, with the objective of identifying which elements within the Framework may be influenced to facilitate the principles adoption in software engineering and achieve trustworthy AI from the design stage.

To achieve that goal, the set of documents forming the Principled AI International Framework was analyzed integrating close reading and natural language processing (NLP) techniques, both detailed in the methods section.

The Principled AI International Framework, as mapped by Fjeld (Fjeld et al., 2020) consists of 35 documents published between 2016 and the last quarter of 2019. The list of documents in Fjeld's study was expanded to 41, including other documents published in 2020 which share the same scope. It is intended for policymakers, advocates, scholars, and others working to capture the benefits, and reduce the harms, of AI technology as it continues to be developed and deployed globally; among which her study includes AI developers.

Altogether, the Principled AI International Framework has different types of authors and signatories like government entities (n=16) representing 39.02% of authors, inter-government entities (n=3) for 7.31%, MultiStakeHolders (n=8) for 19.51%, civil society (n=5) for 12.19%, private sector (n=8) for 19.51%, and the Catholic Church (n=1) for 2.43%. The present study uses the author type classification provided by Fjeld in her study (Fjeld et al., 2020). Similarly, respecting each document author's statement about the purpose of the documents, these were categorized in action plan (n=1), commitment (n=1), general recommendations (n=1), guidelines for developers (n=1), policy-usage (n=1), principles and recommendations (n=1), standardization recommendations (n=1),

each representing 2.43 % of the documents in the Framework, considerations (n=2) for 4.86%, recommendations (n=4) for 9.72%, policy-principles (n=12) for 29.16%, and principles (n=16) for 39.02%. The Appendix I list the documents included in the analyzed corpus for Principled AI.

Although the public policies conforming the Principled International AI Framework seek to condition the development and use of AI, there was no evidence of international standards or any other auditing mechanism in the context of software development—established by IEEE<sup>16</sup> or ISO/IEC<sup>17</sup>—containing references to the proposed principles at the time of the study. This suggests that although the policies show compromise towards trustworthy AI, the principles have not been incorporated into the software industry-specialized regulatory mechanisms. This could significantly affect the principles adoption as a methodological reference for AI and ML developers.

## 2.2 Related Research

Like Fjeld, Jobin, and others (Jobin et al., 2019) mapped, by means of a scoping review, a global landscape of ethics guidelines. The studied documents included the ones in Fjeld's study, which expands to a total of 84 documents containing ethics-driven principles or guidelines for AI. In her study, Jobin also hand-coded the principles and categorized them into 11 themes: (1) Transparency, (2) Justice and Fairness, (3) Non-Maleficence, (4) Responsibility, (5) Privacy, (6) Beneficence, (7) Freedom and Autonomy, (8) Trust, (9) Sustainability, (10) Dignity, and (11) Solidarity. Both studies identify a rupture in the language across the authors without much explanation. This rupture is evident when noticing the different angles from which the defined themes are aborbed. It can be inferred that the differences in the defined themes to categorize the

---

<sup>16</sup> Institute of Electrical and Electronics Engineers is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity, with a subsidiary IEEE Standard Association in charge of designing software engineering specialized standards.

<sup>17</sup> International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). Its purpose is to develop, maintain and promote standards in the fields of information technology and Information and Communications Technology, has an AI specialized subsidiary.

principles are the result of the rupture. In the present research, we decided that we will analyze the documents included in Fjeld's study, as we were able to access them, and because not all the documents in Jobin's study explicitly propose principles for AI.

Mittelstadt (2019) argues that principles alone will not guarantee ethical AI. He compared the emerging principled AI approach for the software industry with the same approach in the field of health care, underlining an array of difficulties the software industry can find when attempting to implement a principle driven approach to develop ADM solutions. Among the obstacles for the principles' implementation in the software industry are a lack of common fiduciary duties, proven methods to translate principles into practice, and a robust legal and professional accountability mechanisms, found in the health care industry as emphasized by Mittelstadt's study. This is why it is important to identify ways of facilitating the principles adoption and their integration in software development methodologies or defining engineering mechanisms which will help adopt the principles as a methodological reference.

With the intention of investigating the intricacies of the principles in support of the stated goal for the present study, we applied some NLP techniques to the documents in the corpus, as detailed in the Methods section. This facilitated learning a set of specific traits from the texts, described in the Results and Discussion section, that otherwise would be more difficult and time-consuming to recognize. Although the amount of analyzed documents nor their length demanded complex NLP procedures, it was possible to take advantage of the use of non-complex methods within the NLP domain for text processing as explained in the next section.

Prior to 2016, other studies proposed principles for AI design. Those principles were mainly focused on conditioning AI's general theories application, and specific design engineering. In 2011, Bostrom (Bostrom & Yudkowsky, 2011) discussed three principles for AI design, specially about autonomous agents, which describe a functional standpoint for non-discriminatory AI. These principles were: the Principle of Substrate Non-Discrimination (concerning the understanding upon the moral status attributed to the autonomous agent); the Principle of Ontogeny Non-Discrimination (to defend the

autonomous agent functionality and experience regardless of how they came to exist); and the Principle of Subjective Rate of Time (to evaluate the agent experience's subjective duration as a basic measure of significance before ethical claims in a specific context like pain or deprivation of freedom). Evidently, Bostrom's non-discrimination notions were thought as favoring AI.

Similarly, Kitano and others (1997) evaluated different design engineering principles for a team of multiple fast-moving robots executing a non-coordinated real-time changing activity tested in a simulated soccer cup. The study aimed at challenging design principles of autonomous agents, multiagent collaboration, strategy acquisition, real-time reasoning, robotics, and sensor fusion. Like Bostrom's study, Kitano's had a product-driven interest, which is a pattern that can be found in the specialized bibliography prior to 2016. We added these two studies to our analysis, to identify the product-driven and performance-driven approach of the principled AI, and to describe examples of the use of those principles prior to 2016.

## 2.3 Method

In this study the empirical research method of close reading was combined with NLP techniques like lexical diversity, semantic similarity, n-gram extraction, and topic modeling to identify elements within the Principled AI International Framework that may be influenced to achieve a design-based trustworthy AI.

The documents shaping the Principled AI International Framework were collected from each author's website in PDF format, then converted to plain text to facilitate the data preparation procedures. For the cases in which the original documents were issued in a language different from the English, the English version, or an English translation approved by the document's author, was collected.

The data consists of three text-based sets. The first dataset consists of each document's text body, which was filtered by removing stop words like prepositions, articles, pronouns, among others, along with the repeated headings, footers, and margin notes resulting in a consolidated and semantically-robust corpus. The second dataset

consists of the principles and guidelines sections from each document in the corpus. Finally, the third dataset contains the principle's declarations that were manually filtered out from the second.

The data was processed using Python (Oliphant, 2007; Python Software Foundation, PSF), a generic and modern computing language, widely used for text analytics. Python's development environment is enriched with libraries like Gensim (Rehurek & Sojka, 2010) that was used for topic modelling, SciPy (Virtanen et al., 2020) tools including Pandas (McKinney, 2010) used for structuring the data, Matplotlib (Hunter, 2007) and Seaborn (Bisong, 2019) used for data visualization, IPython (Pérez & Granger, 2007) used for interactive computing and programming, and NLTK (Loper & Edward., 2009) used for word tokenization, entity recognition, length measurements across different sections of the documents, lexical diversity evaluation, semantic similarity comparisons, and analysis of the verb taxonomy.

First, the length measures by means of word counting and relative length measurements of the different sections of the documents reveal the priority given to the principles as per the portion of the text was reserved for them. These measurements were complemented with the metrics of lexical diversity of the documents and semantic similarity between them. The former was performed using the library LexicalRichness v0.1.3 for PyPi (McCarthy & Jarvis, 2010) using the "mld" method, helping to perform a first approximation to the discourse employed across the documents. The latter was conducted using the sklearn v0.24.1 module, specifically the cosine similarity described by Pedregosa and others (Pedregosa et al., 2011); which instead of using the Euclidean distance between the two vectors representing the words' frequencies uses the cosine of the angle formed by the two vectors. That is a practical technique to evaluate the closeness of different body texts and speeches regarding their content when different language structures are employed, particularly important when working with documents authored by people with different interests, backgrounds, languages, and geographies. No word embedding was considered in the lexical diversity or the semantic similarity analyses, on the contrary, both evaluations only took the word's stem into consideration regardless of its context.

All three datasets were processed using NLTK searching for n-grams ( $1 \leq n \leq 3$ ). This is a technique that weight and rank words combinations determining their value for the text based on their frequency. This analysis helped to understand from a macro perspective the content of the documents when applied to the first dataset, comprehend the elements conditioning the rupture of discourse—referred in the introduction section—when applied to the second dataset, and recognize the variables within the proposed principles, and their context, when applied to the third dataset.

Subsequently, the verbs from the principles and guidelines sections, contained in the second dataset, were extracted, and analyzed looking to review the taxonomy of the proposed actions along the corpus. The parts of the speech that matched with verbs were mined, and grouped using the lexeme similarity criteria, then the size of the lexeme was added to the lemma,<sup>18</sup> obtaining a summarized list of the verbs in the corpus. The resulting list of verbs was then contrasted with a verb taxonomy to identify their distribution within the taxonomy's levels and establish considerations regarding the passive or active character of the abilities demanded from the engineers through the verbs promoted by the corpus writing.

Lastly, topic modelling was performed over the principles' declarations in the third dataset, while exploring the apparent disconnection between the ideas behind the principles' declarations and the remaining text on each document. The Gensim's implementation (Blei et al., 2003) was applied for the Latent Dirichlet Allocation (LDA) to detect topics among the principles' declarations. The LDA is a generative probabilistic model in which each document is considered as a finite mix over an underlying set of topics. Each topic is represented as a set of words and their probability, which means that it is possible to rank topics on the corpus and the keywords in each topic. The technique of topic modelling helped to corroborate some inferences resulting from other tasks of the analysis.

---

<sup>18</sup> In English, for example, run, runs, ran and running are forms of the same lexeme, with run as the lemma by which they are indexed. Lexeme, in this context, refers to the set of all the forms that have the same meaning, and lemma refers to the particular form that is chosen by convention to represent the lexeme.

It is fair to state that extra care was dedicated to the topic modelling part of the analysis as, among the techniques used in the study, it is the one with an additional intrinsic uncertainty. One of the most important issues related to topic modelling with LDA is to know the optimal number of topics ( $k$ ) that should be examined. In consequence, different LDA models with variable values of  $k$  ( $5 \leq k \leq 25$ ) were built, computed the coherence for each topic, and selected the model with the highest coherence value. The best results were found with ten topics and ten keywords per topic, presented in the Results and Discussions section.

## 2.4 Results and Discussions

As mentioned in the Introduction, there are currently numerous efforts by several entities—i.e., governments, intergovernmental agencies, private enterprises, etc.—placed in designing fairer AI. The focus of the present study is on those efforts whose main emphasis seeks to standardize the responsible design and subsequent proper use of AI by means of principles-driven public policies supported by the International Humans Rights Law.

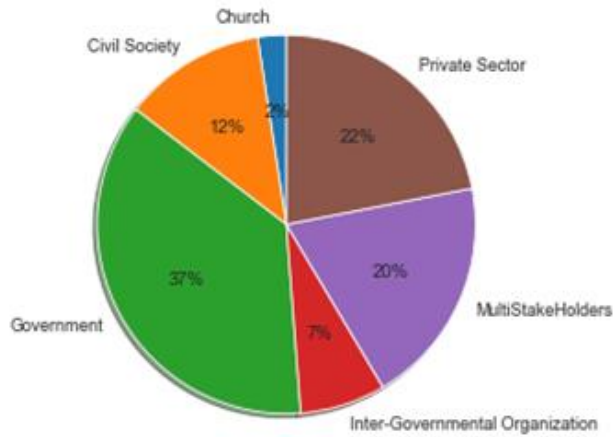
The referred policies designed for trustworthy AI represents an emerging research interest whose starting point can be located around the last quarter of 2016 and it has experienced an increase in the number of principles for AI related publications just two years later, likely due to the intensity of the geopolitical interest in IA of several nations (Suarez, 2018). In this respect, 2018 and 2019 are the years exhibiting a peak of publications of the documents forming the analyzed regulatory framework for principled AI with an average of 15.5 documents in each of the two years.

When exploring the most involved actors in producing or standing as signatories of the documents included in the regulatory Principled AI International Framework, the USA occupies a clear first position with the highest involvement (27.5%), followed by China (12.5%), and France (10%). The three of them combined are related to half of the documents analyzed in this study. Figure 2-1 shows more information about the origin and authorship of the documents. The left side of Figure 2-1 exhibits the distribution of

documents in the analyzed corpus according to their type, while the right side displays the document distribution according to their author type.

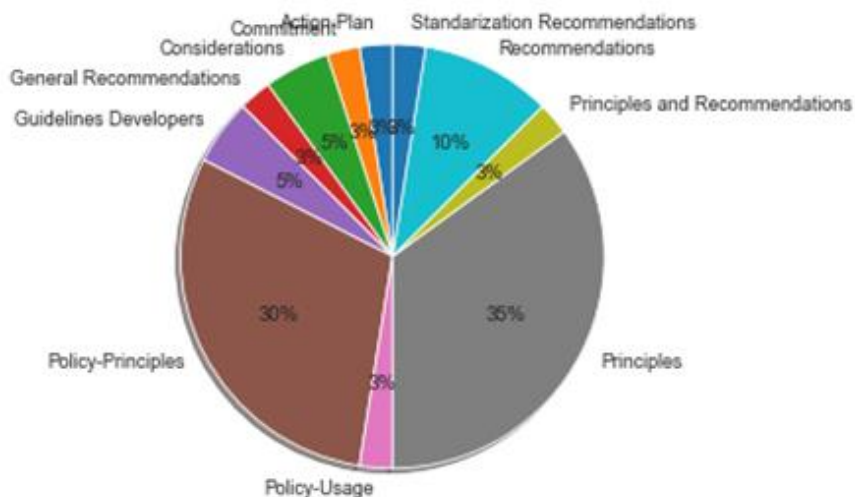


**Figure 2-1: Document and Author Types Distribution per Country [Own Elaboration].**



**Figure 2-2: Author Type General Distribution [Own Elaboration].**





**Figure 2-3: Document Type General Distribution [Own Elaboration].**

In Figure 2-1, it can be noted that in the USA, China, and France, the documents are produced mainly by authors listed as: MultiStakeHolders and Government entities, reusing the catalogue proposed in (Fjeld et al., 2020). There is a greater representation of the private sector in the USA when compared with the remaining countries in the studied corpus. The fact that governments are active entities in the production of these documents shows their commitment to solve the ethical and social problems that the current degree of penetration of AI in almost every aspect of daily life entails. The origins of these problems have been seen in the use of AI and in the early stages of its design, according to the themes highlighted by most of the principles proposed in the corpus.

The authors recognize that the list of countries presented in Figure 2-1 represents developed countries with high technological drive; and they believe it is necessary to stress that social problems as result of the use of technology are problems that transcend the digital gap between developed and developing countries. Hence, although the provisions of the studied documents should be considered equally valuable for and by all countries, it would be important to have a larger and more equitable presence of countries from different parts of the world so that a diversity of ethical and social problems can be studied in its connection with the use of AI in specific contexts.

In a similar line of thoughts, we wanted to explore the references between the analyzed documents as it would be interesting to comprehend the evolution of the principle's adoption in different regions and countries, and probable hierarchies amongst the documents, for example. Unfortunately, only nine documents provide information regarding what other documents, among the ones in the corpus, they reference. Additionally, only one document provide information about which members and non-members -of the inter-governmental organization (document's author)- adhere to the principles it proposes.

Figure 2-2 synthesizes the different types of documents. The leading role of governments in authoring the regulatory frameworks for the design and use of AI becomes clear throughout the corpus. The government-type author has authorship in more than a third (37.5%) of the documents, followed by MultiStakeHolders, and Private Sector type authors, with 20% each. In contrast, the values expressing the authorship of the Civil Society author type denote the need for greater activism on their part as these organizations would provide important input from affected by the decisions made by artificial intelligence systems (AIS).

The presence of Inter-Governmental Organization and Church<sup>19</sup>-type authors denote that the efforts to devise a regulatory framework transcend geographical borders, while connecting different entities, places, and people with a common idea which is then adapted to the particularities of each place.

When analyzing the classification types of the documents according to their purpose, Figure 2-3 shows predominantly those whose objective is to propose Principles (35%), and Policy-Principles (30%), that together with the Recommendations-type documents (10%) represent the 75% of the corpus. These classifications are self-issued by the authors in each document.

---

<sup>19</sup> Referring to the Catholic Church lead by the Vatican who produced the "Rome Call for AI Ethics".

It should be clarified that the "Action Plan" document-type reflects the document titled "AI for Europe" authors' criteria, which presents it as an action plan; however, the document's scope is limited to recommending principles for the design and use of AI, and the guidelines for the principle's implementation into a mechanism in service of designers and consumers. The same occurs for the document typified as "Guidelines Developers."

Regarding the ratio of authors according to their specialization in technical or non-technical backgrounds, in most cases, the analyzed documents lack data related to their authors' training. In other cases, it is stated that prior to their approval the documents were subjected to a public consultation exercise, without providing further details about the participants in the consult. Only seven documents vaguely declare information related to their authors' background.

It is relevant to recommend as part of the set of good practices during the documentation of future regulatory/standardization documents like the ones being analyzed, the inclusion of education background and the empirical experience of contributors; along with other demographical variables that can help determine AI's intrinsic social problems through the viewpoints of individuals from the different communities that are being currently discriminated by AI. Particularly the contributor's education background (i.e., software developers, policymaker, etc.) distribution ratio will allow deeper comprehension of possible differences in the specialized language used in the creation of the policies that might be hindering the implementation of the principles proposed by the regulatory initiatives. Analysis that could shed some light over elements that can be influenced for a better assimilation of the principles by software practitioners.

Based on the available information on the backgrounds of the authors, it can be said that there is a larger presence of authors with software related technical training among the documents produced by the private sector and stakeholders like universities, international standardization organizations, etc. In contrast, the opposite occurs for those documents authored by government organizations, whose author's backgrounds is predominantly on the regulation and policymaking side.

As mentioned before, registering the authors' background information would certainly help improve our understanding of the difficulties these documents face when they are used to establish a practical tool for designers and consumers of AI. Hence, it is essential to stress the need for both regulators (policy makers) and those who are regulated (developers and consumers) to collaborate in establishing a common language that would help the flow of information and knowledge across both domains, the policymaking and software industry.

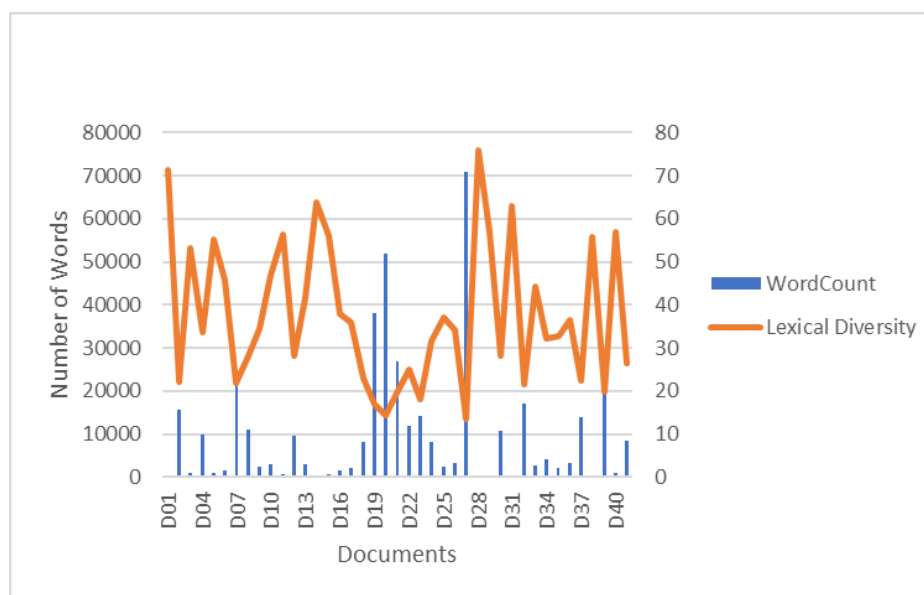
### 2.4.1 Semantic Similarity, Lexical Diversity, and Other Length Measures of the Documents in the Principled AI International Framework

The 41 documents delineating the Principled AI International Framework have a mean length of 10,246 words. The space devoted to the enunciation and description of the principles proposed in these documents exhibits an average length of 359 words, while for guidelines to help implement these principles the average length shows a value of 136 words. This represents a relative average length of 3.5% and 1.33% of the text in the documents, respectively. That is surprising, given that the portion of the documents dedicated to achieving their goals, which is the proposal of principles and guidelines/recommendations for implementing them, is limited to a significantly small share of the corpus.

These differences of relative length for the distinct parts of the body texts evidence that more efforts are dedicated to defining context and justifying the need for principles in most documents that to fully address the principles and explaining their implementation. Hence, it can be said that the analyzed documents are more tuned into providing a description of the problem they hope to solve, than to deepening in the solution they are supposed to outline.

Figure 2-4 shows the document's length and their lexical diversity. The left vertical axis expresses the number of words, while the right vertical axis shows the value for the lexical diversity for each document.

In general, the documents exhibit low values of the metric of lexical diversity. However, there is a peak within the dataset showing a maximum value of 76% in the document titled “Seeking Ground Rules for AI.” The mean lexical diversity<sup>20</sup> value showed by the documents in the corpus was 38%, with a standard deviation of 17 percentile points.



**Figure 2-4: Document’s Word Count and Lexical Diversity Relation [Own Elaboration].**

As can be seen in Figure 2-4, larger documents exhibit lower values of lexical diversity, and vice versa. Both series, word count, and lexical diversity were examined performing a bivariate correlation test<sup>21</sup> with the Pearson’s correlation coefficient, which resulted in (-0.616) significant at the 0.01 level, supporting an inverse correlation among these two variables. This may be denoting the use of a saturated and repetitive language

<sup>20</sup> Lexical diversity is one aspect of 'lexical richness' and refers to the ratio of different unique word stems to the total number of words. It was used to determine the richness of language on each document as an estimated measure of how redundant the language was.

<sup>21</sup> The SPSS v.23 statistical software was used to perform the bivariate correlation test.

in the larger documents of the corpus, in contrast with the use of a specialized language in a specific area of argumentation in shorter documents.

When establishing comparisons between the documents in the corpus, they appear to be semantically similar in the majority of the cases. The mean semantic similarity exhibited among the 309<sup>22</sup> compared unique pairs of documents reached a value of 81% with a standard deviation of 13 percentile points. The 25% of compared pairs reaches a 75.87% of semantic similarity, while the 50% of the pairs reaches an 83.15%, and the 75% reaches a value of 90.24%.

From a semantic perspective, the list of the closer ten pair of documents includes the pairs D07-D19, D08-D19, D20-D19, D07-D08, D21-D02, D02-D39, and D39-D19 with approximately 97% each. These are in addition to the pairs D02-D20, D41-D19, and D39-D21 each one exhibiting around 96% of semantic similarity. The list of the more dissimilar pair of documents includes the pairs D04-D26 with 40%; D04-D09, D04-D25, D04-D03 each with around 38%; D04-D11 with 37%; D04-D31 and D04-D01 with 36%; D04-D15 and D04-D16 with 30%; and D04-D14 with 27%.

The high semantic similarity values suggests the existence of certain relation among the initiatives. On the one hand, the high exhibited value among the sampled documents suggests that the regulatory initiatives are, at least semantically, disposed towards a common goal. Whilst on the other hand, the fact that it was possible to perform document pairings draw attention to possible interdependence between them. The referred interdependence could lead to a given hierarchical structure or any other organigram of some sort, amongst principles, that is further explored in chapter three.

The content of the documents is further explored by means of the NLP techniques of n-gram extraction and topic modeling, detailed hereafter.

---

<sup>22</sup> Number of all possible pairing combinations between the 41 documents excluding self-pairing and duplicates.

## 2.4.2 Principled AI International Framework’s Text Analysis

The text of the documents grouped as part of the Principled AI International Framework was compiled in a single dataset and subjected to the NLP technique of n-gram extraction, as mentioned in the methodology section. The Table 2-1 shows the ten most used n-grams in that corpus.

**Table 2-1: Ten most frequently used n-grams across all documents.**

	Unigrams	Absolute Freq.	Relative Freq.	Bi grams	Absolute Freq.	Relative Freq.	Trigrams	Absolute Freq.	Relative Freq.
1	ai	7476	1.94	artificial intelligence	1693	0.44	autonomous intelligent systems	416	0.11
2	data	4156	1.08	ai systems	653	0.17	ieee global initiative	351	0.09
3	systems	2502	0.65	machine learning	603	0.16	ethics autonomous intelligent	318	0.08
4	intelligence	2011	0.52	intelligent systems	446	0.12	global initiative ethics	314	0.08
5	artificial	1798	0.47	autonomous intelligent	429	0.11	initiative ethics autonomous	314	0.08
6	research	1728	0.45	human rights	368	0.10	algorithms artificial intelligence	126	0.03
7	human	1670	0.43	use ai	326	0.08	strategy artificial intelligence	112	0.03
8	use	1595	0.41	computer science	321	0.08	national strategy artificial	111	0.03
9	development	1515	0.39	ethics autonomous	319	0.08	discussion paper national	109	0.03
10	technology	1506	0.39	personal data	298	0.08	paper national design	109	0.03

As expected, the top five positions of the unigrams frame the context of the texts’ argument. Interestingly, the term "human" occupies position seven in the same column; perhaps this demonstrates the human approach that has the regulatory framework intended to be defined by these documents. More details on this specific aspect are provided below. Another interesting finding to highlight is that the scope of the regulatory framework delimited by the documents in the corpus can be distinguished in positions eight “use,” ninth “development,” and tenth “technology.” It should be stated that the table presented is an excerpt from a larger analysis, where for the case of the

unigrams terms that interest us as “ethics” and “ethical” ranked in the 13 and 25 positions, respectively, with relative frequencies of 0.30 and 0.23.

If examined closely, in the bigram’s column can be seen how the unigrams gain context. Consequently, the focus on "human" and "use" takes on a new nuance in human rights and the use of AI. Another element suggested through the bigrams is that that focus will be influenced by an ethical approach to autonomous or intelligent systems (row 9), with special consideration to personal data (row 10). We believe it is worth to mention that from the extended analysis of the 50 most relevant N-grams, in the bigrams column terms like "data protection" at position 25<sup>th</sup>, and "trustworthy ai," in position 38<sup>th</sup>, add sustainment to the previous statement about the corpus’ ethical approach to artificial and intelligent systems through how they handle personal data.

It can also be noted from the trigrams that the effort to achieve national strategies for an ethically aligned design is orchestrated in the context of AI, with support in rows 2, 4, 5, 7, and 8 from Table 2-1. The previous idea gains strength by including trigrams like: “ethically aligned design” (row 23), “ethical matters raised” (row 26), “matters raised algorithms” (row 27), and “intelligent systems law” (row 28) from the extended analysis.

Having presented an analysis of the terms in which the documents from the corpus are expressed and considering the small relative length dedicated to the approach of principles for a trustworthy AI, it is now opportune to narrow the focus of the analysis specifically on the principles section and evaluate its correspondence with the rest of the document.

### 2.4.3 Analysis of Proposed Principles

The portion of the text corresponding to the principles enunciation and description was manually separated and compiled into another dataset. This dataset was also subjected to the NLP technique of n-gram extraction. Table 2-2 shows the ten most frequent n-grams.

When the analysis was delimited to the proposed principle’s enunciation and argumentation portion of the text, as can be seen in Table 2-2, the context in which the



object of argument is demarcated remains the same, which is in complete consistent with the rest of the sections of the documents. However, verbs such as "must" in row 4, and "ensure" in row 8 convey a more normative character to the principles. This idea is further explored through an analysis of the verbs used in the principle's descriptions.

**Table 2-2: Ten most frequently used n-grams across principles.**

	Unigrams	Absolute Freq.	Relative Freq.	Bi grams	Absolute Freq.	Relative Freq.	Trigrams	Absolute Freq.	Relative Freq.
1	ai	532	3.91	ai systems	111	0.82	artificial intelligent systems	11	0.08
2	data	208	1.53	artificial intelligence	59	0.43	aida driven decisions	9	0.07
3	systems	154	1.13	use ai	25	0.18	context consistent state	8	0.06
4	must	131	0.96	ai system	24	0.18	consistent state art	8	0.06
5	use	112	0.82	personal data	18	0.13	ai systems must	8	0.06
6	human	97	0.71	ai technologies	18	0.13	states parties present	8	0.06
7	development	94	0.69	ai must	18	0.13	parties present covenant	8	0.06
8	ensure	92	0.68	ai development	18	0.13	ai system lifecycle	6	0.04
9	government	82	0.60	ai research	17	0.12	appropriate context consistent	6	0.04
10	people	74	0.54	machine learning	16	0.12	present covenants recognize	6	0.04

The bigram "ai must", in row 7 supports the idea mentioned in the previous paragraph on the normative dimension associated with the skills that should be attributed to the design and consumption of AIS. Other bigrams strengthen the context of the principles, recognizing their range of action from the academy (row 8), and industry (row 9), as well as the emphasis on personal data (row 5). The column that exhibits the trigrams in Table 2-2 provides no new or relevant information rather than expanding on the bigram "ai must" with the chain "ai systems must" (row 5) and recognizing the work throughout the life cycle of AIS (row 8).

A close reading of principles containing the chain represented by the tri-gram "ai systems must" are mostly oriented to the general normative intent of the initiatives for developing and using fair ("...AI systems must be fair..."), non-discriminatory ("...AI

systems must not be discriminatory...”, and “...AI systems must not discriminate based on a list of personal attributes...”), sustainable and environmentally friendly (“...AI systems must be sustainable/environmentally friendly/or both...”), to provide some examples. Undoubtedly, those principles showcase the principled AI international framework’s clear intention to condition the operationalization of those variables into a methodological framework that could facilitate the principles implementation.

Interestingly, unigrams such as "rights," "privacy", "security," "transparency," and "fairness" exhibit a relative frequency of 0.40, 0.39, 0.30, 0.26, and 0.25 units, respectively, occupying more distant positions from the top scored unigrams in the table, within the extended analysis (50 n-grams). They are historically among the terms that describe the ethical dilemmas rooted in the use of AI solutions. The same goes for Trigrams: "standards best practices," "privacy data protection," "equality diversity fairness," "diversity fairness social," and "fairness social justice," have relative frequencies that hold values of 0.04 units in the first, and 0.02 in the rest.

It is our opinion that although it appears that there is a common agreement among the involved actors on which issues need work to achieve a trustworthy AI, they address those issues differently. The absence of a common criterion could, among other difficulties, affect the definition of a methodological mechanism for software developers. Pursuing to prove whether such agreement over the fundamental issues, expressed in the form of principles, really exists, and despite the apparent disconnection with their description, the analysis to each of the principles was narrowed to their enunciation.

Table 2-3 shows the ten most common n-grams used in the enunciations of the principles proposed in each document across the corpus. Note that the description of the principles is excluded from the analysis. As can be seen in the table, n-grams change completely compared to Table 2-2.

**Table 2-3: Ten most frequently used n-grams across the principles' declarations.**

	Unigrams	Absolute Freq.	Relative Freq.	Bi grams	Absolute Freq.	Relative Freq.	Trigrams	Absolute Freq.	Relative Freq.
1	ai	57	4.29	ai systems	9	0.68	ai systems deployed	3	0.23

2	principle	32	241	artificial intelligence	9	0.68	inclusive growth sustainable	2	0.15
3	privacy	20	150	ensure bot	5	0.38	growth sustainable development	2	0.15
4	data	19	143	non discrimination	4	0.30	sustainable development well	2	0.15
5	transparency	18	1.35	accountability transparency	4	0.30	development well human	2	0.15
6	fairness	16	1.20	u government	4	0.30	values fairness transparency	2	0.15
7	human	16	1.20	principle respect	3	0.23	fairness transparency explainability	2	0.15
8	rights	15	1.13	sustainable development	3	0.23	transparency explainability robustness	2	0.15
9	ensure	15	1.13	transparency explainability	3	0.23	explainability robustness security	2	0.15
10	systems	15	1.13	privacy security	3	0.23	robustness security safety	2	0.15

The unigrams present in the principle's declaration include variables like "privacy," "transparency," and "fairness," while bigrams include "non-discrimination," "accountability transparency," "transparency explainability," and "privacy security." Lastly, trigrams enrich the idea of inclusive growth sustainable development with human values, in rows two to five, with "robustness," which also appears alongside other features like the one already mentioned. In the extended analysis (50 most frequently used n-grams) it can be noted how the variables related to the context and scope of the corpus are pushed to more distant positions on the list in favor of those variables linked to the objectives that are sought to achieve with the proposed principles.

When comparing Tables 2-2 and 2-3, the disconnection between the principle's declarations and their descriptions become clear. This is a common problem among the analyzed documents, and it could be an element that makes it difficult to build a standardization mechanism that serves as a methodological reference in the design of trustworthy AI. The description of the proposed principles has a more direct link to the documents in general than to its statement, whereas a more logical link between these two parts should be established through the statements of principles in a "justification-enunciation-argumentation" outline.

The general overview on the documents referred their focus on human, and human rights as the agreed basis for principled AI. The scope of the documents mainly encompasses the use and development of technologies from an ethical perspective, with a human rights-based foundation. In general, the documents seek to support the concept of trustworthy AI from variables such as data protection to arrive at the conception of an AI ethically aligned design.

It is not until the analysis is narrowed to the principles' statements that the variables associated with trustworthy AI are extended using terms such as: privacy, transparency, fairness, non-discrimination, accountability, explainability, and security. This denotes that the value of principles lies in their statements, in addition to pointing at the existing disconnection of the principles' statements with their description, and with the remaining body text.

To arrange a "justification-enunciation-argumentation" outline, it is adequate to identify the need to operationalize the variables involved with the idea behind Trustworthy AI. Our hypothesis at this respect, is that mapping the concepts around the idea of trustworthy AI can influence a common understanding of it as an object of study, therefore impacting the effectiveness and efficiency of the efforts dedicated to ensuring the develop of ADM systems with trustworthiness as a quality characteristic. The hypothesis its motivated by a premise stating that understanding a problem well makes up 50% of its solution.

A section that also has been reserved a place for, although small, in most documents is dedicated to proposing a set of guidelines and recommendations supporting the principles implementation. The following section shows an analysis of the guidelines and their aligning with the principles.

#### 2.4.4 Analysis of Guidelines to Implement the Principles

When considering the guidelines as the set of actions projected to support the implementation of the principles, they are expected to represent concrete actions attached to construction like taxonomies such as applying and creating. In this regard, it is curious

that verbs with presence among the 50 n-grams most frequently used are "consider," with a relative frequency of 0.77 units, "ensure," with 0.47, and "must" with 0.44. It should be clarified that also the verbs "use," "design," and "research" are within this set of n-grams, however serving as an explanatory function. For example, in sentences such as: "...operator organizations use...," "... ai design...," and "... ai research..." also present among the bigrams and trigrams identified as the 50 most frequent.

In an effort to expand upon the actions proposed for the implementation of the principles, the verbs were extracted and subjected to a lemmatizing process to consolidate them into a summarized list that helps a better understanding of the skills behind the proposed actions. The summarized list consists of 30 verbs. It can be declared that while there are verbs associated with Bloom's taxonomies such as "apply" (17.2%), and "create" (14.8%), which represent approximately one third of the list, the set of verbs includes mostly actions commonly related to other Bloom's taxonomies such as "understand" (20.6%), "analyze" (32.1%), and "evaluate" (15.3%).

Verb taxonomies used to describe the software development process within the specialized literature (Azuma, Coallier, & Garbajosa, 2003; Hernán-Losada, Lázaro-Carrascosa, & Velazquez-Iturbide, 2004; Lister et al., 2004; Whalley et al., 2006) were found to be adjusted versions of Bloom's verb taxonomy (Krathwohl, 2002) for the digital age. The contrast of the abilities described by the summarized list of verbs with the ones defined in available professional competency profiles for software development (Colomo-Palacios et al., 2010; Costa & Santos, 2017)—usually divided in generic and specific competences—suggests that these verbs describe abilities software developers perform both as an active ability, when executing activities who are intrinsic to the software development process, and as a passive ability, when referring to the modeled context.

We verified the context in which the extracted verbs, summarized in the list of verbs, were used, by performing a close reading of the principles' descriptions and the guidelines for their implementation. As a result of the close reading, it was possible to determine that verbs associated with the taxonomies "analyze" and "evaluate" were used

as common language, as can be seen in: "Technologists have a responsibility to ensure the safe design of AI systems," which is the second principle in AI Policy Principles (Appendix A-D06). In this example the verb "to ensure," associated with taxonomies like "analyze" and "evaluate," is not making direct reference to the "verification" and "validation" tasks within the analysis, design, and software implementation stages, but is used as a common expression, giving the verb a passive character on its context. Something similar occurs with "Workplace AI should be tested to ensure that it does not discriminate against vulnerable individuals or communities," which is the second guideline proposed to support the implementation of the principles in the future of work and education for the digital age (Appendix A-D10). In this case, the verbs "tested" and "ensure" could be interpreted as skills directly referencing the quality assurance tasks, however the context framed by the guideline link them more with the work environment, alluding to designer's own biases, and that measures are taken to ensure—now as an active skill — those non-discriminatory decisions are being conditioned from the design stage, hence, another passive ability. The verbs "does" and "discriminates," from the same example, are acknowledged and associated to the taxonomy "apply," however, in that context, they are more adhered to the business than to the development process.

Then, it is accurate to say that, within the context of the present corpus, the group of verbs from the summarized list associated to taxonomies like "understand," "analyze," and "evaluate" encompasses a set of skills that could be perceived as passive skills in a practical context such as software design, particularly when designing AI solutions. This inclination for passive skills in the language used to describe the proposed actions for the implementation of the principles can be perceived as another element elevating the difficulties for the principle's assimilation by AI designers. Consequently, this could prevent the framework from becoming a methodological reference during the AI project lifecycle. At the same time, it is understandable that this inclination can respond to an interest in maintaining the proposal as a general framework that can then be adapted to each context as needed, safeguarding its global and general character.

The same is true when performing this analysis on the principles, where there is also an inclination for verbs representing passive abilities such as the aforementioned.

This may be normal given the practical context in which the actions within the corpus are framed. A balance between effectiveness and generality of the proposals based on the cost and benefit linked to the use of certain language remains to be achieved in these types of documents.

#### 2.4.5 Topic Modeling on the Principle's Enunciation

The topic modeling was performed to triangulate some of the observations that have been presented earlier. The ten most represented topics in the text are listed below; it should be clarified that the topics are extracted based on the sections dedicated to the principle's declaration only<sup>23</sup>:

Topic 1. “system”(0.0000022) + “ai”(0.0000019) + “right”(0.0000019) + “must”(0.0000017) + “technology”(0.0000016) + “shall”(0.0000009) + “human”(0.0000009) + “people”(0.0000007) + “research”(0.0000004) + “decision”(0.0000004)<sup>24</sup>

Topic 2. “agency”(0.036) + “assessment”(0.017) + “even”(0.017) + “system”(0.017) + “accountability”(0.016) + “obligation”(0.015) + “automate”(0.015) + “assess”(0.014) + “individual”(0.013) + “decision”(0.013)

Topic 3. “system”(0.029) + “ai”(0.014) + “value”(0.013) + “human”(0.010) + “rapidly”(0.009) + “automation”(0.009) + “grow”(0.009) + “power”(0.008) + “people”(0.008) + “share”(0.008)

---

<sup>23</sup> The analysis of the topics modeling for the documents in the Principled AI International Framework, and for the portion of the text corresponding to the guidelines for the implementation of the principles, along with comparisons between these and the topics drawn in the present study are shown in another publication. This has been done in order to maintain the focus of this study on the exploration of the principles, and in correspondence with the communication strategy of the research project.

<sup>24</sup> The value of the frequency for the terms in topic one is so close to “0” that more than four significant digits are needed to show a digit different from “0.”

Topic 4. “right”(0.064) + “shall”(0.043) + “law”(0.017) + “freedom”(0.013) +  
 “education”(0.012) + “protection”(0.011) + “public”(0.011) +  
 “include”(0.008) + “religion”(0.008) + “family”(0.008)

Topic 5. “wide”(0.020) + “solution”(0.017) + “definition”(0.014) + “seek”(0.012) +  
 “practice”(0.012) + “way”(0.009) + “dialogue”(0.008) + “algorithm”(0.008)  
 + “explore”(0.008) + “research”(0.007)

Topic 6. “remedy”(0.009) + “diversity”(0.004) + “promote”(0.004) +  
 “inclusion”(0.004) + “effective”(0.003) + “equality”(0.011) + “non”(0.002)  
 + “discrimination”(0.001) + “right”(0.000) + “system”(0.000)

Topic 7. “must”(0.040) + “life”(0.013) + “development”(0.013) + “public”(0.012) +  
 “ais”(0.012) + “people”(0.012) + “individual”(0.011) + “decision”(0.009) +  
 “human”(0.009) + “personal”(0.009)

Topic 8. “constraint”(0.004) + “educate”(0.004) + “oppose”(0.004) +  
 “maximize”(0.004) + “openness”(0.004) + “scientist”(0.004) +  
 “listen”(0.004) + “interpretable”(0.004) + “engineering”(0.004) +  
 “socially”(0.001)

Topic 9. “system”(0.028) + “ai”(0.026) + “datum”(0.020) + “ensure”(0.012) +  
 “human”(0.012) + “technology”(0.010) + “design”(0.009) +  
 “development”(0.007) + “must”(0.007) + “people”(0.007)

Topic 10. “government”(0.039) + “ai”(0.015) + “public”(0.013) + “policy”(0.012) +  
 “ensure”(0.011) + “research”(0.011) + “sector”(0.010) + “must”(0.010) +  
 “recommend”(0.010) + “take”(0.010)

Among the topics listed above, topic nine is identified to be dominant, since it is the most representative topic among the principles section in 25 documents, which denotes 62.5% of the documents in the corpus. A topic’s dominance expresses the topic’s representativeness, through its terms, across the pieces of texts being analyzed when compared with other topic’s distribution in the same pieces of text. In that sense, the



representativeness of Topic 9 is followed by topic ten, being dominant in the principle's enunciation of 10% of the documents; and Topic 2, in 5%. In contrast, Topic 1 does not exhibit any dominance throughout the documents in the corpus. The rest of the topics happened to dominate the distribution of representativeness in the principle's enunciation section of a single document each.

It is interesting to see how these topics that have become dominant in the great majority of documents reflect the following: (1) the objects of discourse ("systems," "ai," and "technology" to name examples); (2) the action field influenced by these objects ("human," "people," and "decision" among others); and (3) the subjective methodological approach that was already criticized in previous sections (expressed in terms such as "must," "ensure," and "assess," for example).

This supports the idea that there is a clear notion of the problem being addressed with the regulatory framework for the AI based on the variables that are affected, but the consensus on the methodological approach to be followed has yet to mature. It is clear that all author types, throughout the period and space covered by the analyzed documents, face difficulties in providing a tool that can be used in practice, with a clearly defined set of tasks and guidelines for AI solution designers to follow.

At the time of this research, no International Standardization Office (ISO) standards were found, nor published by the Institute of Electrical and Electronics Engineers (IEEE) standards association, although it is known that the latter institution is taking into consideration some of the documents of the corpus in the design of standards related to the subject. Therefore, a contrast with the terminology used in the standards could not be established.

## 2.5 Conclusions of the Chapter

The analysis of the Principled AI International Framework presented in this chapter is a much-needed complement to the study of the technical and engineering limitations in the field of machine learning -presented in chapter one-, as it address the thesis's research problem from the standpoint of the mechanisms created to shield the human rights that are currently being violated as result of discriminatory ADMS' outcomes. It allows to expand the range of issues that must be taken into account when building fairer ADM systems that are not being considered as evidenced in chapter one. Additionally, it provides further elements, by means of the proposed principles, exposing less evident unfair ADMS' outcomes such as the ones linked to the right to be forgotten, for example.

The present study highlights several issues rooted in the language used in the redaction of the documents within the Principled AI International Framework that may be restraining the framework's adoption as a methodological reference for ADM system developers, while it distinguishes the operationalization of the variables on which the principles are based as the fundamental element to achieve a trustworthy ADM solution from the design stages. The identified linguistic differences, conditioned by the different professional backgrounds of the contributors to the Principled AI International Framework, might affect the assimilation of the principles in the framework and their proper implementation.

There exists general consensus on what variables need to be normalized in order to develop and use trustworthy ADM solutions, as evidenced in exhibited semantic similarity of the documents in the corpus, and the analysis of the principle's enunciation. Contrastingly, there is less consensus in how the enunciated principles need to be operationalized, when looking at the multiple edges aborded across the principle's description and related guidelines. That variety of definitions negatively affects an adequate implementation of the principles in the framework. In addition, the variables contained in the principles proposed by the Principled AI International Framework and their interrelation, needs to be further explored in order to map the trustworthiness's conceptual vicinity in the context of ADM solutions. Doing that will reduce the

ambiguity created by the multiple edges those variables are intended to be addressed from, according to the principle's description and implementation guidelines; and influence a better assimilation of the principles and their further incorporation in the design of ADM solutions. As a consequence, we find appropriate to conduct a more in-depth study of the principles within the Principle AI International Framework, exhibited in chapter three.

The examination of the proposed guidelines for the principle's implementation evidenced their lack of agency to overcome the principles' described ambiguities and capacity to facilitate the principle's assimilation by software developers as a methodological reference to be incorporated into the software development lifecycle, especially when building ADM solutions, proving that further analysis of the framework is needed. This is particularly important when considering that the topic modeling of the framework's corpus points the methodology related elements as a primary concern, along with the lack of mechanisms from international standardization and normative institutions (at the moment of the study) to implement the framework's principles.

The analysis presented in this chapter provides the basis for a better comprehension of the environment of the projected solution to the thesis's research problem and the fundamentals for the proposed model's architecture. Still, a further analysis of the relation among the Principled AI International Framework (Chapter three) and the main variables associated to trustworthiness like fairness, non-discrimination, etc. (Chapter four), identified by a first approach to the principles presented within this chapter, are needed for the model design.

## 2.6 Works Cited in the Chapter

- Lister, R., Adams, E. S., Fitzgerald, S., Fone, W., Hamer, J., Lindholm, M., McCartney, R., Mostrom, J.E., Sanders, K., Seppala, O., Simon, B. & Thomas, L. (2004). A multi-national study of reading and tracing skills in novice programmers. *Working group reports from ITiCSE on innovation and technology in computer science education*, 119-150.
- Azuma, M., Coallier, F., & Garbajosa, J. (2003). How to apply the Bloom taxonomy to software engineering. *Software technology and engineering practice: Eleventh annual international workshop*, 117-122. IEEE Computer Society Press.
- Bisong, E. (2019). Matplotlib and Seaborn. *Building machine learning and deep learning models on google cloud platform*. Apress. doi: [https://doi.org/10.1007/978-1-4842-4470-8\\_12](https://doi.org/10.1007/978-1-4842-4470-8_12)
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993–1022.
- Bostrom, N., & Yudkowsky, E. (2011). The ethics of artificial intelligence. In K. Frankish, & W. Ramsey, *Cambridge handbook of artificial intelligence*. Cambridge University Press.
- Colomo-Palacios, R., Tovar-Caro, E., García-Crespo, Á., & Gómez-Berbís, J. M. (2010). Identifying technical competences of IT professionals: The case of software engineers. *International Journal of Human Capital and Information Technology Professionals*, 1(1), 13. doi:10.4018/jhcitp.2010091103
- Costa, C., & Santos, M.Y. (2017). The data scientist profile and its representativeness in the European e-competence framework and the skills framework for the information age. *International Journal of Information Management*, 37(6), 726-734.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society.
- Hernán-Losada, I., Lázaro-Carrascosa, C., & Velazquez-Iturbide, J. A. (2004). On the use of Bloom's taxonomy as a basis to design educational software on programming. *Proceedings of World Conference on Engineering and Technology Education*, 351-355.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9, 90-95. doi:10.1109/MCSE.2007.55
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 389-399.
- Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E., & Matsubara, H. (1997). RoboCup: A challenge problem for AI. *AI Magazine*, 18(1), 73-85. doi: <https://doi.org/10.1609/aimag.v18i1.1276>

- Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212-218. doi: 10.1207/s15430421tip4104\_2
- Loper, S. B., & Edward., E. K. (2009). *Natural language processing with Python*. O'Reilly Media.
- McCarthy, P. M., & Jarvis, S. (2010). MTLN, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392.
- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv*. doi:arXiv:1908.09635v2 [cs.LG]
- Mittelstadt, B. (2019). *Principles alone cannot guarantee ethical AI*. Oxford Internet Institute, University of Oxford.
- Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, 9(3), 10-20.
- Pedregosa, F., Varoquaux, N., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in {P}ython. *Journal of Machine Learning Research*, 12, 2825-2830.
- Pérez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science & Engineering*, 9, 21-29. doi:DOI:10.1109/MCSE.2007.53
- Python Software Foundation. (n.d.). *Python*. (Python Org) Retrieved February 2020, from [www.python.org](http://www.python.org)
- Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.
- Sahil, V., & Rubin, J. (2018). Fairness definitions explained. *ACM/IEEE international workshop on software fairness*.
- Suarez, J. L. (2018). La nacionalización de la estrategia en torno a la inteligencia artificial estado, política y futuro. *Revista de Occidente*, 446-447 (Julio-Agosto), 5-18.
- Trewin, S. (n.d.). AI fairness for people with disabilities: Point of view. *arXiv*. doi:<https://arxiv.org/ftp/arxiv/papers/1811/1811.10670>
- Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., . . . van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, 261-272. doi:<https://doi.org/10.1038/s41592-019-0686-2>

Whalley, J. L., Lister, R., Thompson, E., Clear, T., Robbins, P., Kumar, P. A., & Prasad, C. (2006). An Australasian study of reading and comprehension skills in novice programmers, using the bloom and SOLO taxonomies. *Proceedings of the 8th Australasian Conference on Computing Education*, Volume 52, 243-252.

## Chapter 3

### Principled AI Engineering Challenges Towards Trustworthy AI

The efforts to design fairer ADM solutions have been increasingly migrating from the domain of software engineering, given the limitations exposed in chapter one, to other fields like public policy and regulation design, some of which were explored in chapter two. This chapter expands upon the study of the Principled AI International Framework deepening in the relations among the framework's proposed principles, with the objective of determining how they can be incorporated into a methodological tool for developers of ADM systems, particularly into the model proposed in chapter five. The analysis of the principles includes the application of methods of network theory and close reading, while exhibits further insights regarding the variables, according to the principles, used to frame trustworthy artificial decision-making solutions.

#### 3.1 Introduction

When studying the impact of artificial intelligent systems' (AIS) outcomes on their target population by means of the evaluation of the decisions they have been subject to, several recurrent issues linked to biased and discriminatory decisions frequently emerge (Miller, 2020; Varona, 2020). As a result, it can be concluded that the software industry has been developing and deploying unfair solutions, or at least that these solutions lack the necessary requirements to satisfy the ethical, social and political desires to have a principled, trustworthy AI for all individuals and social groups. On the one hand, can be agreed that most of the discriminatory and biased outcomes produced by ADM solutions are unintended, as they were not engineered into the system by the team of developers. On the other hand, it can also be agreed that from the available specialized literature, the existence of such discriminatory and biased outcomes and their source (data, algorithm, or introduced by the human) have been vastly researched and documented. Surprisingly, none of the reasons presented by the Chaos Report (The Standish Group, 2020) for describing the main causes for project failure in the software industry make reference to any element associated to the potential harmful impact of AIS

solutions, against disadvantage minorities, when conditioning, amplifying, or perpetuating bias.

At the time of this study there was no international standard from the International Standardization Organization (ISO) or from the Institute of Electrical and Electronics Engineers (IEEE) dedicated to address from a methodological point of view the ethical implications of AIS. There exist many efforts to cope with specific traits of the referred implications that have been found within machine learning related research (Yapo & Weiss, 2018) and other fields of study (Chouldechova, 2017; Fjeld et al., 2020).

The latest trends to overcome unfair AIS have been criticized in the previous chapter, which is evidence that the methodological limitations of the followed approaches as linked to faulty data collection and preparation processes, inefficient algorithmic fairness assessment, and lack of commitment to fairness from the design stages. The mentioned study also points out the emergent inclination of industry and governments to delegate in the policy making the fixing of the incomplete methodological models used in software industry.

These deficits have led to the description of a set of regulatory norms forming the international principled AI framework (Fjeld et al., 2020) to be adopted by software engineers as a complement to their methodology reference, especially when designing and developing AIS. The principled AI framework, however, through its determination to act as an international reference to be later adjusted to the particularities of any region, country, or enterprise kept a regulatory based language, divorced from the practices and habits of the software industry. This resulted in creating new obstacles for the framework adoption as a methodological reference that are discussed in the literature, and analyzed in chapter two.

The present chapter expands on the findings detailed in chapter two regarding to the language differences in the Principled AI International Framework by providing further discussion of examples illustrating the semantic significance of such differences. The goal of such discussion is to have all necessary elements for the design of a



capability and maturity model for trustworthy AI based on the international principled AI regulatory framework.

## 3.2 Related Research

Among the available specialized literature (Mittelstadt, 2019), is discussed some differences between high-level principles proposed to complement, from an ethical perspective, the software development methodologies. The study allocates the source of such differences to the lack of political and normative agreements, along with other factors more related to the software industry like the current state of standards and norms, scarcity of proven methods to translate principles into practice, and the inexistence of robust legal and professional accountability mechanisms.

An example of the differences referred in the previous paragraph is discussed in by Gong (2020). The study contrasts the variable privacy in relation to transparency from both a practical and a principled perspective. The conclusions establish transparent privacy as a dynamic component that improves data quality while underlining it as the only viable path towards accountability and public trust. This conclusion supports our recommendation to update the software quality features currently normed and standardized by the International Standardization Organization and International Electronics Commission (2014) and its derived subdivisions.

Likewise, other study (Buruk et al., 2020) can be seen as an additional example as it analyses existing conflicts among the principles proposed in the three sources: The Asilomar AI principles, The Montreal Declaration for Responsible Development of Artificial Intelligence, and Ethics Guidelines for Trustworthy AI which are all included in the present study. The performed method exposed the documents to a checklist designed to compare elements on the principle's enunciation and description and concluded that all three documents approached the very variables they propose from different points of view.

A possible reason for the aforementioned definitional ambiguity can be found in the literature (Krafft et al., 2020). The study surveyed experts and review public policy

documents to examine researcher and policy-maker conceptions of AI and found that AI researchers favor definitions of AI that emphasize technical functionality, while policymakers are more tuned with definitions that compare systems to human thinking and behavior instead. Additionally, the study asserts that functionality centered definitions are found to be more inclusive of technology's use today, whereas human focused definitions tend to speculate with hypothetical future technologies.

Parallely, Madaio (Madaio et al., 2020) identifies that the abstract nature of the multiple regulatory mechanisms, created to guide from an ethical standpoint the development of AIS, is the reason why is so difficult to operationalize the variables contained in the referred mechanisms. The study also explores the use of dedicated checklists designed to ethically assist the assurance and assessment of fairness during the process of software development. In this regard, the authors indicate that unless those checklists are grounded in practitioners' needs, they may be misused; therefore, they should be designed following the development team efforts towards the variable rather than placing the focus directly on the variables themselves. We found those elements particularly important as they reinforce the need for solving the language ambiguity described in the previous chapter. This is further explored in the present chapter, with the goal of gaining more understanding of the definition, scope, and interrelation of the variables in the principles, so the elements derived from the analysis of the principles presented in these two chapters can be efficiently incorporated into the proposed model described in Chapter five.

In summary, Gong and Bunk, by dissecting two apparently opposite variables (like privacy and transparency, that ended up being two dimensions of the same principle), and by contrasting the principles of three regulatory initiatives, have helped to frame the divorce, established by Mittelstadt, between the language used in the definition of ethical principles and the language utilized in the creation of standards and norms helping to implement those principles. While Krafft and Madaio point out the dichotomic character of the professional background of regulators and software developers, and the abstract nature of the principle driven initiatives, respectively, as two of the reasons for

such divorce. Therefore, it is appropriate to state that these studies are limited to defining the problem suggested by Mittelstadt and exploring its causes.

These studies set the theoretical foundations for the issue of ambiguity around the principles' variables that we have stressed in the previous chapter, allowing us to take it a step further and focus on exploring the principles in order to determine their multidimensional scope through their interdependence. We used network theory methods to delve principles sharing common goals, their interconnection, and strength as a cluster, and in the universe of principles per the Principled AI International Framework, by means of several centrality measures. The algorithms we used as part of our methods when applying the network's centrality measures are described next.

First, we must state that a heuristic approach (Blondel et al., 2008) to extract clusters was used to establish the network modularity. Blondel's method is quite popular, time saving, and produces results with high levels of accuracy. The network modularity would constitute at first view of the relation of the nodes within the network, a notion that it is supported by the evaluation of the degree centrality of the network. This is the simplest measure of node connectivity, especially in nondirected networks as the ones in this study. It assigns an importance score based simply on the number of links held by each node.

The harmonic centrality, other of the closeness applied measures, is a variant of closeness centrality, particularly efficient when dealing with graph containing disconnected nodes, as are some of ours. The method poses a different way of calculating the average distance using the inverse of the distances of one node to all other nodes, enabling the algorithm to deal with infinite values resulting from the islands nodes in the network as defined by Marchiori & Latora (2000).

The closeness centrality scores each node based on their 'closeness' to all other nodes in the network by evaluating the shortest paths between all nodes and assigning each node a score based on its sum of shortest paths. This provides an idea of which nodes are best placed to quickly influence the network, the main reason our closeness

centrality results are so like the betweenness centrality measure. We used closeness centrality as defined in by Bavelas (1950).

The betweenness centrality measure (Freeman, 1977) works with the number of times a node lies on the shortest path between other nodes. It shows the nodes acting as ‘bridges’ between other nodes in the network by identifying all the shortest paths and registering the number of times each node falls on one. For that reason, it is mostly used to understand the flow within the network. However, caution needs to be used as it can underline the most influential node in the cluster or a node on the frontier of several clusters. We try to avoid that misunderstanding by triangulating the results by means of the comparison with other centrality metrics.

The algorithms PageRank (Brin & Page, 1998) and Eiger Vector’s centrality (Bonacich, 2007) incorporate a notion of importance to the closeness measures already described. For the case of the present study, the PageRank algorithm constituted an objective way to measure closeness considering not only distance between nodes, as with the previous described algorithms, but by also considering the importance of the node’s vicinity. More relevant neighbors equals more relevance of the self, which brings certain objectivity to the notion of importance. By doing so we were able to identify the significantly most important nodes in the network, what directly translates into the most important principles on the corpus. The performed approach focused on random walks through the graph, which proved to be an asset as discussed in the section “Centrality measures in principle’s network.” In this respect, the Eiger Vector’s centrality, along with the PageRank are superior to the other previously described centrality measures, as they use vicinity degrees to calculate the relevance for each node.

The networks were visualized using Gephi (Bastian et al., 2009), which is a popular open-source software for network exploration, manipulation, and visualization. The networks are represented using the Fruchterman-Reingold layout (Fruchterman & Reingold, 1991). This algorithm simulates the graph as a system of mass particles. The nodes simulate the mass particles, and the edges are springs between the particles. Although it is slower than other methods for bigger networks the Fruchterman-Reingold

layout has become a standard as it offers a clearer representation for the identified clusters.

### 3.3 Method

The present study was primarily conducted using the theoretical research method of close reading, allowing us to gain specific and comprehensive understanding from the texts included in the principled AI international framework. We paid particular attention to the principles proposed within the corpus formed by the mentioned regulatory instruments. Our close reading was complemented with network analysis techniques which helped us comprehend the relation between the documents through the principles contained in each document, and between each principle in the corpus.

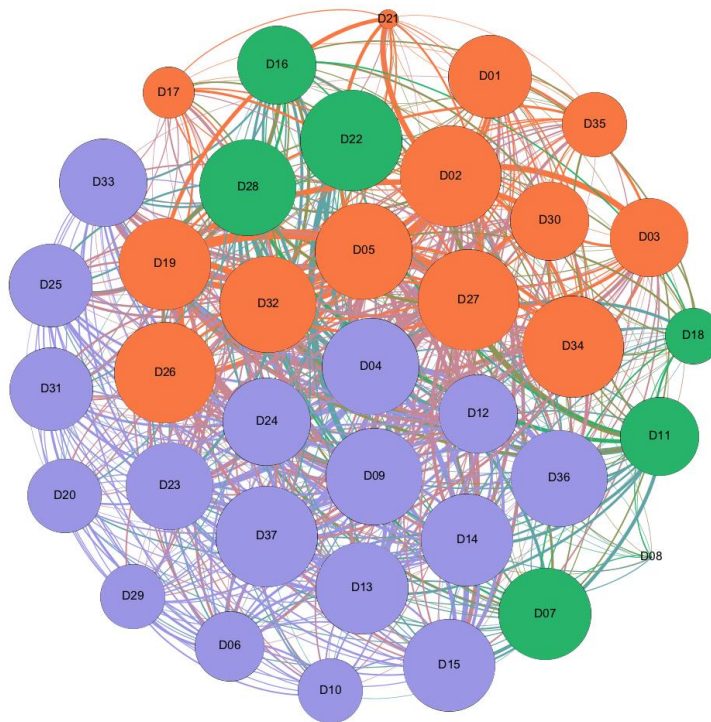
First, all proposed principles were listed. Second, two main networks were built using Gephi, as mentioned earlier, applying the Fruchterman-Reingold layout. The first exhibited the connections between the documents proposing the exact same principles, similar principles, or principles whose description were semantically close to each other even when the principles' enunciations were different. The second exhibited the relations between each principle, also attending to their semantic similarity. Other derived networks were created to help visualize related metrics providing deeper insights on the information contained in the main networks that are detailed in the following sections of the paper. The criteria being analyzed is the principle's message, and it is transversal to all networks.

Next, the principle's recurrence ratio among the documents in the corpus was calculated and the list of principles reduced, discriminating those outside the software engineering scope in order to provide further interpretation layers that might be needed for their assimilation by software engineers, as a proposal for methodological reference. In the section "Principles as methodological reference for software engineers" we used the PageRank-appointed most important nodes in the principle's network to exemplify the referred interpretation layers.

Last, the research method of triangulation was performed to contrast the results of different approaches of the centrality measure. Distance and degree-based measures were computed using degree, harmonic, closeness, and betweenness centrality, while the eigen vector and page rank centralities added a vicinity's relevance element to the previous measurements.

### 3.4 Principled AI International Framework Modularity. Document Level

Figure 3-1 represents the visualization of the identified clusters of documents sharing a common message through the principles proposed. The network exhibits the relation among the 37 regulatory instruments in the corpus, which are represented by an equal number of nodes, through 583 edges. The size of the nodes denotes the number of shared principles that a document has with the rest of the nodes in the network. As stated in the methods section, the criteria for establishing the relations among the nodes represent if the principles proposed by the analyzed documents target the same objective.



**Figure 3-1: Clusters Network of Documents Sharing Principles' Goals [Own Elaboration].**

As can be seen in Figure 3-1, the documents in the Principled AI International Framework can be outlined in three different clusters. Cluster one (the larger) is represented by purple, gathering 46% of the nodes, and includes documents D04, D06, D09, D10, D12-D15, D20, D23-D25, D29, D31, D33, D36, and D37. A second cluster represented by orange, gathering 41% of the nodes in the network and includes documents D01-D03, D05, D17, D19, D26, D27, D30, D32, and D34. And a third cluster (the smallest) is represented by green, gathering 13% of the nodes and includes documents: D07, D08, D11, D16, D18, D22, and D28.

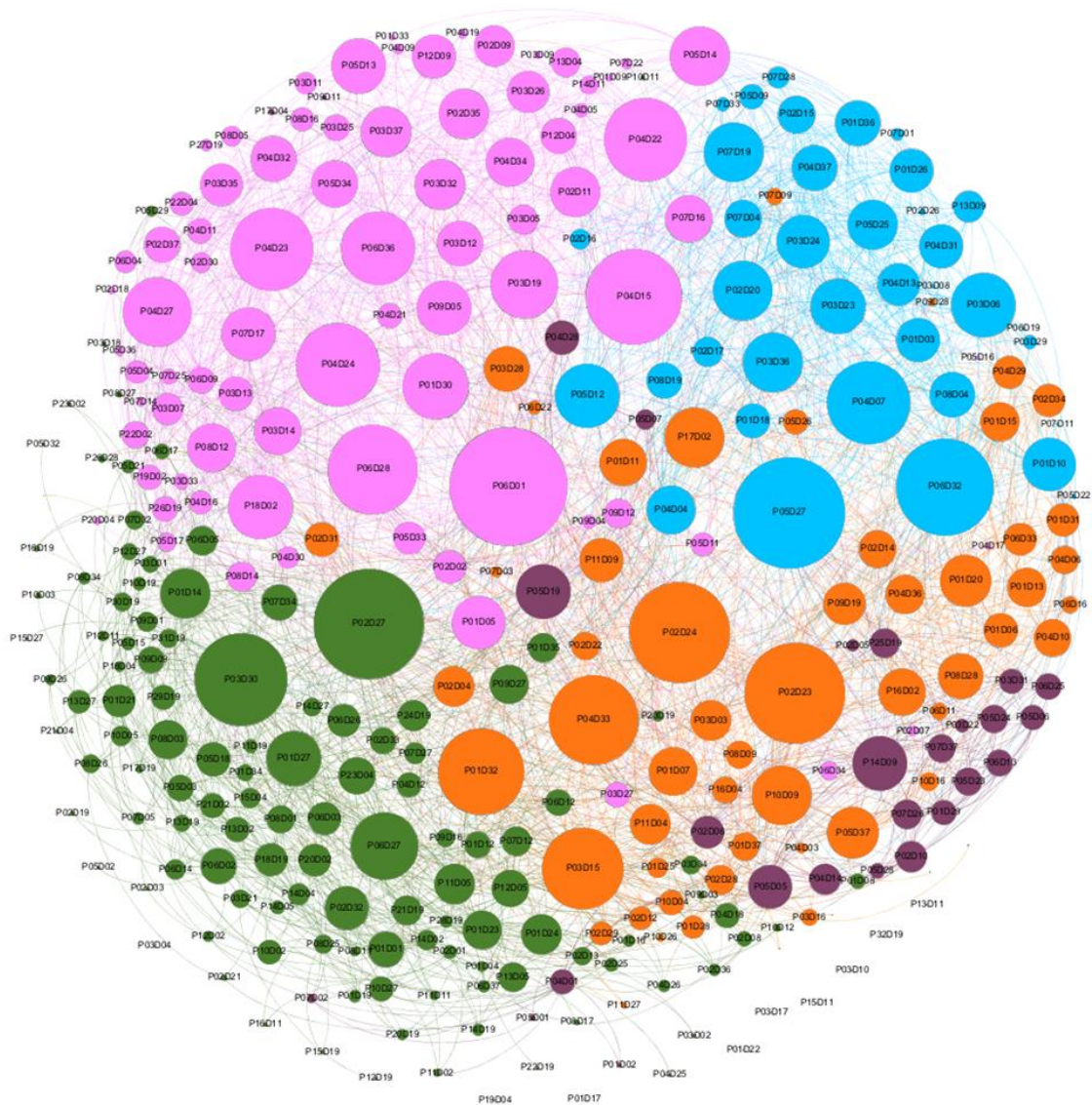
The similar sizes of the two modules of the network representing clusters one (46%) and two (41%) suggests that there are two predominant sets of pursued common goals among the represented documents in contrast with a third set of goals shared by a smaller number of documents delineating cluster number three (13%). Although the relations between the nodes in the network represents shared principles, we wanted to double-check the existence any particular distribution amongst other attributes within the dataset such as the document's authors or document's scope that once the clusters were delimited. In that regard, we found that the identified clusters exhibited a heterogeneous distribution of the author type, and the document type attributes (see Appendix A list of documents), and that neither are elements of distinction when comparing the clusters. Therefore, it can be stated that there is no particular relations, driven by the documents' authorship nor their scope, within clusters, that could propitiate further analysis parallel to the principles on their own.

### 3.5 Principled AI International Framework's Modularity and Derived Centrality Measures: Principle Level

When exploring the principles, it can be noticed from Figure 3-2 that the number of delimited clusters grouping the principles almost doubles the number of those modeled considering the documents. The fashioned network for principles, based on the degree metric, includes 323 nodes and 2858 edges divided in five well-defined clusters and nine other single-node modularity classes. The nodes' nomenclature responds to P##D## where the first pair of hashtags correspond to the (P) principle's number (01-32) and the

second pair to the number of the (D) document's number (01-37). The edges model the shared scope among connected principles.

The network visualizes the results of the degree centrality measure. We first use it for describing the clusters the principles are grouped by, and then we perform some comparisons with other centrality measures like harmonic centrality, closeness centrality, betweenness centrality, Page Rank, and the Eigen vector's centrality.



**Figure 3-2: Cluster Network of Principles Sharing Common Goals [Own Elaboration].**



The larger cluster embodies 36.22% of the network with 117 nodes and is represented with the color green. It is seconded by a cluster with 80 (24.77%) nodes represented with the color pink. These two subnetworks represent more than half of the connections in the principles' network, followed by other two clusters sizing 53 (16.41%) and 37 (11.46%) nodes represented with colors orange, and blue, respectively. The smallest sub-network gathers 27 (8.36%) nodes and is represented with the color purple. Let us refer to them as cluster one, two, three, four, and five, respectively, according to the order in which they were mentioned.

The different identified clusters described in the previous paragraph indicate that there are five focus areas the analyzed principles are covering. Cluster one covers an area that is mainly focused on the human aspect of AI, that is, by having the human being as the user, the developer, the researcher, the educator, or the object of study. It also primarily pursues AI for the benefit of all, and to empower as many people as possible, according to the most connected node. Cluster two focuses on functionality variables like privacy, security, robustness, reliability, etc. that in the software engineering field of study, can be treated as quality variables. Similarly, cluster three also focuses on a specific set of variables, this time linked to fairness and justice. It also focuses on the prevalence of human control over AI solutions. Cluster four focuses on transparency. Lastly, cluster five focuses on liability and accountability for trusted AI.

Differences in sizes and degrees among the identified subnetworks provide a notion of the dispersion of the focus in each cluster. The focus area delimited by cluster one stands out from the rest as it gathers the larger number of nodes. It is closely followed by the focus area framed by cluster two, which also separates itself from the remaining clusters. It is interesting to point that, as exhibited in Table 3-1, clusters three and four (being smaller) are more interconnected. Both clusters exhibit a clear scope through their principles by means of the variables they pursue to harness: ethical, fairness, impartiality, justice, accountability, and transparency.

It is noticeable that clusters one and five, being the largest and the smallest clusters, respectively, have similar interconnection ratios. While cluster one spreads

through the different factors the principles’ authors of the documents in this cluster believe must be achieved to benefit and empower humans, cluster five extends through the elements framing liability and accountability of AI solutions at times, and at others, liability and accountability of developers and users. Cluster five also includes other variables related to accountability, which are sufficiently independent to be spread in subcategories like no-subversion, and controllability, just to mention two examples.

In regard of the clusters’ interconnection ratios, their exhibited value might be suggesting that clusters three and four have more cohesion in contrast to clusters one and five. Consequently, the focus areas covered by the former are sharper than the area covered by the latter, which present themselves as fuzzier. The interconnection ratio was evaluated by dividing the number of nodes in a cluster by the total degree of that cluster. This suggests that principles oriented to normalize variables, such as the principles grouped in the clusters three and four, are more consistent, than those focused on the general purpose of the principled AI initiative, expressed in “AI good for all,” “Sustainable AI,” etc.

We want to point out that none of the clusters represented in Figure 3-2 follows a regular distribution regarding the nodes’ degree, what means that a few principles are more common either because of their frequency (described “as is”) on the corpus, because their specific goal is shared by other principles, or because they were specified using too general terms. The degree’s unbalanced distribution is illustrated by the distance between the particular degree of principles in a cluster and the average degree of that same cluster, expressed by the standard deviation as is shown in Table 3-1.

**Table 3-1: Summary of principles based on the subnetworks’ Weighted Degree distribution.**

Cluster Number	Total Degree	Average Degree	Standard Deviation	Nodes (Weighted Degree)
1	1439	12.30	14.03	P02D27 (69), P03D30 (58), P06D27 (42), P01D27 (34), P01D14 (31), P02D32 (27), P11D05 (25), P01D24 (25), P08D03 (24), P09D27 (24), P01D23 (24), P12D05 (23), P01D01 (22), P23D04 (22), P07D34 (22), P06D02 (22), P05D18 (21), P06D03 (20), P20D02 (20), P06D26 (20), P18D19 (20), P01D21 (20), P24D19 (19), P06D05 (18), P01D35 (18), P13D05 (18), P21D19 (18), P07D12 (18)
2	1743	21.79	13.75	P06D01 (74), P04D15 (60), P06D28 (56), P04D22 (52), P04D23 (52), P04D24 (52), P06D36 (46), P04D27 (43),

				P03D19 (42), P01D30 (41), P18D02 (40), P05D14 (37), P09D05 (34), P05D13 (34), P01D05 (33), P07D17 (33), P02D11 (31), P03D32 (31), P02D35 (31), P03D37 (31)
3	1204	22.72	14.04	P02D23 (63), P02D24 (63), P04D33 (55), P01D32 (54), P03D15 (51), P17D02 (37), P10D09 (37), P05D37 (32), P01D20 (32), P01D07 (30), P16D02 (29), P01D11 (29), P03D28 (28), P08D28 (28), P11D09 (27)
4	924	24.97	13.75	P05D27 (70), P06D32 (61), P04D07 (52), P05D12 (40), P03D06 (40), P07D19 (37), P02D20 (37), P03D36 (37), P03D23 (34), P03D24 (33), P01D10 (33)
5	404	14.96	14.06	P14D09 (34), P05D19 (34), P05D05 (27), P04D28 (21), P02D06 (20), P04D14 (19), P02D10 (19), P25D19 (18), P07D26 (18)

The referred unbalance is more accentuated in clusters one and five as their standard deviation values are greater than and closer to their average degree, respectively, which supports our previous statements about these two clusters of principles sweeping a less strongly-defined focus area. The principles listed in the Table represents a small portion of the principles, around 1/4 and 1/3 of the principles included in their clusters and yet they sum, in each case, more than half of the total degrees for the cluster. All of them are far above the value of the standard deviation for the cluster. That proportion turns them into the principles whose goals are more spread along their respective networks and, therefore, they become principles of special interest for the present study. The following principles are of note.

From cluster one, the principle represented by the node P02D27 (China), declared as “For Humanity,” is described as “The research and development of AI should serve humanity and conform to human values as well as the overall interests of humankind. Human privacy, dignity, freedom, autonomy, and rights should be sufficiently respected. AI should not be used to against, utilize or harm human beings.” The principle represented by the node P03D30 (United Arab Emirates) declared as “Humanity,” is described as:

AI should be beneficial to humans and aligned with human values, in both the long and short term; AI systems should be built to serve and inform, and not to deceive and manipulate. Nations should collaborate to avoid an arms race in lethal autonomous weapons, and such weapons should be tightly controlled. Active cooperation should be pursued to avoid corner-cutting on safety standards. Systems designed to inform significant decisions should do so impartially.

From cluster two, the principle represented by the node P06D01(USA), declared as “Work to maximize the benefits and address the potential challenges of AI technologies,” and described is as:

Working to protect the privacy and security of individuals. Striving to understand and respect the interests of all parties that may be impacted by AI advances. Working to ensure that AI research and engineering clusters remain socially responsible, sensitive, and engaged directly with the potential influences of AI technologies on wider society. Ensuring that AI research and technology is robust, reliable, trustworthy, and operates within secure constraints. Opposing development and use of AI technologies that would violate international conventions or human rights and promoting safeguards and technologies that do no harm.

From cluster three, the principle represented by the node P02D23 (Organization for Economic Co-operation and Development OECD) is declared as “Human-centered values and fairness,” and is described as:

AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, nondiscrimination and equality, diversity, fairness, social justice, and internationally recognized labour and equality, diversity, fairness, social justice, and internationally recognized labour rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art

This is along with the principle represented by the node P02D24 (G20), equally declared as “Human-centered values and fairness” and described exactly as P02D23.

From cluster four, the principle represented by the node P05D27(China), declared as “Be Ethical”, and is described as:

AI research and development should take ethical design approaches to make the system trustworthy. This may include, but not limited to making the system as fair as possible, reducing possible discrimination and biases, improving its transparency, explainability, and predictability, and making the system more traceable, auditable and accountable.

The principle represented by the node P06D32 (Japan), declared as “The Principle of Fairness, Accountability, and Transparency,” is described as:

The Principle of Fairness, Accountability, and Transparency: In an "AI-Ready Society", it is necessary to ensure fairness and transparency in decision-making, appropriate accountability for the results, and trust in the technology, so that people who use AI are not subject to undue discrimination with regard to personal background, or to unfair treatment in terms of human dignity.

The principle represented by the node P04D07 (European Commission for the Efficiency of Justice CEPEJ) is declared the “Principle of transparency, impartiality and fairness” and is described as:

Make data processing methods accessible and understandable, authorize external audits: a) a balance must be struck between the intellectual property of certain processing methods and the need for transparency (access to the design process), impartiality (absence of bias), fairness and intellectual integrity (prioritizing the interests of justice) when tools are used that may have legal consequences or may significantly affect people’s lives. It should be made clear that these measures apply to the whole design and operating chain as the selection process and the quality and organization of data directly influence the learning phase and b) the first option is complete technical transparency (for example, open-source code and documentation), which is sometimes restricted by the protection of trade secrets. The system could also be explained in clear and familiar language (to describe how results are produced) by communicating, for example, the nature of the services offered, the tools that have been developed, performance and the risks of error. Independent authorities or experts could be tasked with certifying and

auditing processing methods or providing advice beforehand. Public authorities could grant certification, to be regularly reviewed.

From cluster five, the principle represented by the node P14D09 (Argentine<sup>25</sup>), declared as “Accountability” is described as “People and corporations who design and deploy AI systems must be accountable for how their systems are designed and operated. The development of AI must be responsible, safe and useful. AI must maintain the legal status of tools, and legal persons need to retain control over, and responsibility for, these tools at all times.” The principle represented by the node P05D19 (UK) is both declared and described as “Strengthening access and control.” As such, they represent the most connected principles in their sub-networks.

As suggested earlier, we considered it opportune to contrast this analysis with other centrality measures to determine the principles that may be appraised over the rest of the analyzed set. This helped us evaluate those principles, unleashing a more effective operationalization of the rest when approached from a methodological perspective, and it allowed us to focus on the general aspects first and then to narrow the scope on to specific elements.

### 3.6 Analysis of Relevance Across Principles’ Networks

The harmonic, closeness, betweenness, Page Rank and Eigen vector’s centrality measures are shown in Tables 3-2 and 3-3. To determine the higher-scored nodes for each network we designed the following method. We calculated the average and standard deviation values for each centrality measure. Then, as the values in some cases did not represent peaks within the top scored, we divided the distance between the highest value and the frontier value (set by the sum of the average and the standard deviation results) in two halves. Last, we listed the principles on the upper half. By doing so we wanted to avoid arbitrariness as much as possible when determining the summarized list of

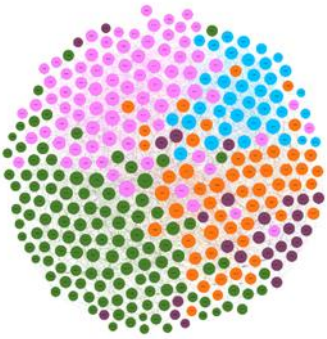
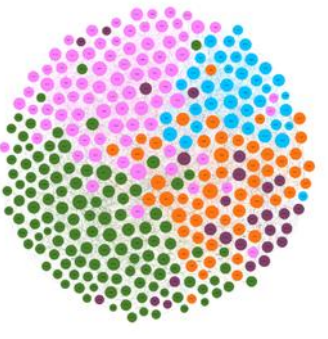
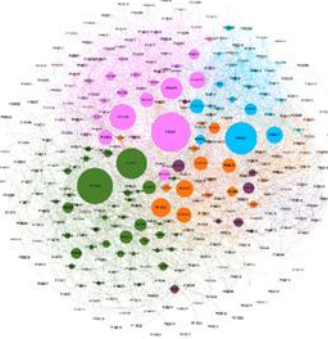
---

<sup>25</sup> Civil society Think20, commonly referred to as T20. They direct their work with direct emphasis on the G20 intergovernmental organization.

principles not only more connected on the network regardless of their respective clusters, but also attending to the importance of their vicinity.

It is not surprising the similarity of the networks represented in Figures 3-3 and 3-4 given that the approach followed for those metrics are similar. It is not surprising either that the measurements highlight the same 11 principles out of 15 unique nodes listed. The four contrasted networks exhibit, as per the top-rated nodes, only differences in their position on the ranking, which is understandable given the network structure. Also, the nodes denominated P05D19 and P14D19 from cluster five (referred in the previous section when listing the more connected nodes for each cluster) and the nodes P06D32 and P03D19 flagged with a star in Table 3-2 are nodes appointed by a single centrality measure. As the latter are not the last listed nodes of their corresponding rankings at Table 3-2, we feel more inclined to believe that their relevance has arisen given the metric's functioning rather than because of how we decided to make the cut for each metric. Therefore, they will be also included in the summarized list without further examination. In contrast, the principles P05D19 and P14D19 are excluded from the summarized list of principles as they are relevant to their clusters, but not in the same way to the network as a whole.

**Table 3-2: Harmonic, Closeness, and Betweenness centrality measures.**

 <p><b>Figure 3-3: Harmonic Centrality Network [Own Elaboration].</b></p>	 <p><b>Figure 3-4: Closeness Centrality Network [Own Elaboration].</b></p>	 <p><b>Figure 3-5: Betweenness Centrality Network [Own Elaboration].</b></p>
Higher Scored Nodes		
P06D01 (0.6009), P02D27 (0.5847), P05D27 (0.5706),	P06D01 (0.5350), P02D27 (0.5165), P06D28 (0.5073),	P06D01 (0.0586), P03D30 (0.0540), P05D27 (0.0481),

P02D23 (0.5698), P02D24 (0.5698), P06D28 (0.5644), P03D30 (0.5642), P04D33 (0.5623), P04D15 (0.5610), P01D32 (0.5602), P06D32 (0.5567),* P03D15 (0.5562)	P04D33 (0.5040), P03D30 (0.5032), P01D32 (0.5032), P02D23 (0.5032), P02D24 (0.5032), P03D15 (0.5016), P04D15 (0.4945), P05D27 (0.4929)	P02D27 (0.0461), P06D28 (0.0392), P03D19 (0.0319),* P01D32 (0.0282), P04D33 (0.0255)
--	---	--

We can say that the smallest subset of nodes from Table 3-2, corresponding to those with higher betweenness centrality values, are the ones working as bridges between the clusters within the network. When considering the structure of the network we can also state that higher values of betweenness centrality mean more general or recurrent principles that will connect with other whose descriptions will tend to specialize in a given direction. Let us exemplify with the description of principle P06D01, presented in the previous section. Undoubtedly, the broader specification of P06D01 finds specialization in P16D02 (USA), P03D03 (Switzerland), or P04D03, for example. This narrows the scope to take extra care to ensure fairness of AI-based systems when using them to make decisions about individuals; adopting a human-in-command approach retaining control over, and responsibility for, AI solutions at all times in the figure of the decision maker; and ensuring a genderless unbiased AI, to provide a few cases.

We can also state that the ranked nodes under the closeness centrality network are the nodes providing faster access to all other nodes in the network. Let us use P2D27 to illustrate this scenario as is the second ranked principled after P06D01. Principle P02D27's description captures the general goal of the Principled AI International Regulatory Framework by affirming that the research and development of AI solutions should serve humanity and conform to human values as well as the overall interests of humankind like privacy, dignity, freedom, autonomy, and respect for human rights; and it clarifies AI solutions should not be used against human beings or harm them.

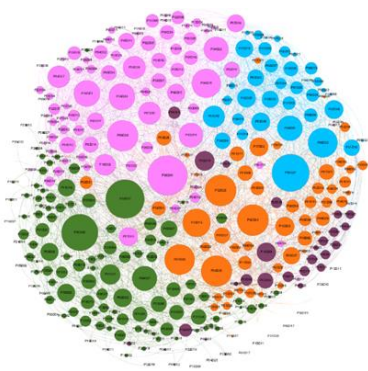
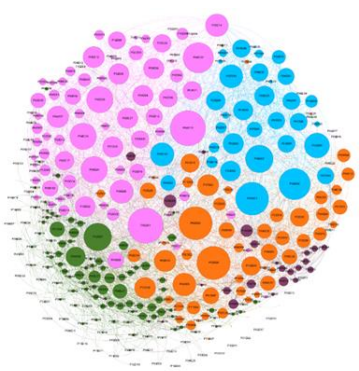
We must clarify that we used, as explained before, the harmonic closeness centrality measure as the network included nine single class nodes and this centrality measure is known to be effective in such cases. As can be seen in Table 3-2, the results exhibited for the mentioned metric are similar to the closeness measure, what leads us to believe that the principles represented by the single class nodes had not much influence



over the results when contrasting both sets of values. A probable reason could be the network’s composition, that shows the great majority of nodes being connected to one another. The nodes listed under this measure on Table 3-2 represent principles with the closest relation to any other principle in the framework.

When considering the importance of the vicinity the principles are connected to, Table 3-3 shows that Page rank and Eigen vector’s centrality support the results of the previously evaluated measures. As can be noted, Page rank doubles the number of and includes the principles appointed by Eigen vector’s centrality; the same (the inclusive relation) is true for the rest of the contrasted centrality metrics.

**Table 3-3: Page Rank and Eigen Vector centrality measures.**

			
<p><b>Figure 3-6: Page Rank Centrality Network [Own Elaboration].</b></p>		<p><b>Figure 3-7: Eigen Vector Centrality Network [Own Elaboration].</b></p>	
Higher Scored Nodes			
P06D01(0.0113), P03D30(0.0104), P06D28(0.0089), P06D32(0.0085), P04D23(0.0076), *P06D27(0.0072), P06D36(0.0067), *P05D12(0.0063),	P02D27(0.0110), P02D23(0.0090), P04D33(0.0089), P04D15(0.0084), P04D24(0.0076), P04D22(0.0072), *P01D30(0.0066), P18D02(0.0062)*	P05D27(0.0107), P02D24(0.0090), P01D32(0.0088), P03D15(0.0081), P04D07(0.0073), P03D19(0.0071), P04D27(0.0065)*,	P04D15 (1.0000), P06D01(0.9987), P02D23 (0.9443), P02D24(0.9443), P05D27 (0.9359), P06D32(0.8899), P04D22 (0.8605), P02D27(0.7831), P04D07 (0.7771), P04D23(0.7513), P04D24 (0.7513), P06D36(0.7483)

Both metrics, exhibited in Table 3-3 above, aim at the importance of each node on the network and they complement the analysis of closeness centrality with a sort of elite/prestige/importance/relevance score, vicinity-based, for a given node. A higher score means the node is more important, something that is achieved if the referred node is connected with other nodes that are relevant themselves as well. Several studies (De

Meo, 2017; Suzuki, 2017; Prateek, Pasala & Aracena, 2013) use Page Rank and Eigen Vector centrality measures for crossed validation purposes like we do, in this analysis. For the purpose of our study, we deploy them following different approaches: Eigen evaluate each principle's vicinity, including the distance consideration from every node to each other; and the page rank randomly travels the graph registering the frequency of hitting each node along the path.

Among the main differences between the results of the implementation of the Page rank and the Eigen vector's centrality measures, illustrated in the Table 3-3, stands out the one consisting of the relevance values exhibited by the node P04D15 (Spain) in both metrics. P04D15 represents a principle proposed by Telefónica<sup>26</sup> (Telefónica, 2018) in reference to assuring the variable of privacy and security from the design stages within the project lifecycle. The way Telefónica perceives the implementation of the mentioned principle (as can be noticed through the principle's implementation guidelines) bridges the principle with a set of other non-functional variables and business philosophies, which in turn relates it with other relevant principles. The relationship of the principle P04D15 with other principles, through the variables described in the guidelines for its implementation, creates a gap between the principle's argumentation and the principle's statement. This consequently fades the principle's scope. We believe this could be one of the reasons for the ambiguity around some of the principles, that was referred in previous sections of this chapter, and in chapter two.

The Page rank approach has been demonstrated to be less susceptible than Eigen vector's centrality to the issue posed by the disconnection between the principles' declaration and description. Therefore, we deepened our analysis of the nodes scored by the Page rank algorithm with higher relevance values. The referred analysis is presented

---

<sup>26</sup> Telefónica, S.A. is a Spanish multinational telecommunications company headquartered in Madrid, Spain. It is one of the largest telephone operators and mobile network providers in the world providing fixed and mobile telephony, broadband and subscription television to Europe and the Americas. Listed as Private Sector in the present study.

in the section 3.7.2, hereafter, and a summarized list of the principles can be found in Appendix B.

### 3.7 Principles as Methodological Reference for Software Engineers

In chapter two, and in previous sections of this chapter, we have conducted the analysis of the Principled AI International Framework following a distant reading approach. First, applying NLP techniques to assess the documents in the framework, and to explore the semantics of the principles to have an idea of their scope. Then, using techniques from network theories to explore the relations among the principles, so as to determine which were more relevant for their specific clusters and for the general network. Once we have determined the most relevant principles in the network, we performed a close reading so we could have a better understanding of their particularities, were able to operationalize them, and transform them in engineering practices that we could incorporate as part of the proposed model. The achieved complementation of both methods, distant and close reading, used to analyze the Principled AI International Framework exemplifies Underwood's (2019) discussion regarding the methods coexistence without one superseding nor conflicting the other. In the present thesis project both methods provided useful and complementing description scales for the studied framework at different depth levels.

In this section we use the method of close reading to analyze a sample of the studied principles, consisting of the principles represented by the single (disconnected from the network) nodes, and the principles represented by the nodes with higher Page rank's relevance score. Our aim is to discuss possible ways for them to be adopted from a software engineering point of view. By doing so, we seek to provide our interpretation of the principles while highlighting the elements posing a methodological challenge towards a development model focused on trustworthy AI. First, we present our observations of the single-node classes as we think it is important to delve into the uniqueness of those principles. And secondly, we present a subsection addressing the principles represented by the Page rank top scored nodes from the previous section as they can be perceived as

the most transversal principles among the studied principles as they link other relevant nodes together.

### 3.7.1 Analysis of the Principles Represented by the Single-Node Classes in the Degree Centrality Network

There are seven principles in the corpus with no relation to other principles, and they are represented by the IDs P03D04 (USA), P19D04, P03D10 (Belgium), P01D17 (China), P03D17, P32D19 (UK), and P01D22 (Canada). For the most part, these principles are out of scope within the software industry and are more aligned with policy design as is the following cases:

The principles P03D04 and P19D04 are proposed by Future of Life<sup>27</sup> (2017) and refer to the links between science and public policy, on the basis of the healthy and constructive exchange between AI researchers and policymakers. They also highlight the necessary cautions to be taken around speculative postures regarding the future AI capabilities. The authors classify principle P03D04 as a research-related issue, and the principle P19D04 as a long-term issue, both outside a clear methodological grip within the software engineering.

We were not surprised to see the principle represented by the node P32D19 isolated due to its scope. The principle proposed by the British House of Lords in (2018) settles the British vision for the United Kingdom in an AI world and what, in their judgment, the transformative potential for artificial intelligence on society requires.

Last, to conclude this first subdivision discussing policy-related principles of this section we present principle P01D22 proposed by the Treasury Board of Canada Secretariat in (Treasury Board of Canada Secretariat, 2018). It discusses the sentence of “People should always be governed – and perceive to be governed – by people.” This statement is included under the headline: “Policy, Ethical, and Legal Considerations of

---

<sup>27</sup> Non-profit research institute and outreach organization based in Boston, USA that works to mitigate existential risks facing humanity, particularly existential risk from advanced AI.

AI,” which intends to determine the retention of the responsibility of the made decisions supported or provided by AIS in the figure of the human decision maker. We agree with P01D22, mainly because there are no available mechanisms to hold automatic decision-maker solutions accountable for their outcomes yet, as criticized in chapter two and by Varona (2018).

In contrast, there is a subgroup of those principles listed at the beginning of the section that can be discussed by software engineers with a methodological prism. An example of this is principle P03D10, proposed by The Public Voice Coalition (2018) which highlights the unidirectional identification between humans and AI systems, and what they denominate as identification asymmetry. The coalition is troubled by AIS or the AIS’ operator knowing a big deal of people when people know little or nothing about the AIS they are exposed to. This principle can be easily adopted by software engineers and AI designers by including identification-related functional requirements during the requirements’ elicitation stage. This can be done in agreement with the client, or de facto. A disagreement with the client in this regard will lead to an economic hazard to the software development team, a matter of conscience to the development team, or an ethical default to the client needs to be further discussed.

Principles P01D17 and P03D17 proposed by the government of China, specifically by the Standards Administrations office (Collection of Institutional Authors, 2018), seek to define a set of variables founding AIS in the pursue of standardizing terminology research which academics and practitioners around the globe can agree upon and reach a better understanding. We agree with this need as stated the previous chapter. The present study, along with the one described in chapter four, aims to argue in favor of the idea. However, from a methodological standing within the IT industry, we find it necessary the creation of a standard dictionary for the emerging terms the international standards have failed to include so far.

In particular, the principle P03D17 focuses on measuring the level of intelligence of AIS and can be related to the efforts behind the idea of the Turin machine (Pinar, Cicekli & Akman, 2000; Marshall, 2021). The adoption of this principle might be

subjected to the conceptualization, operationalization, and modelling of variables like artificial understanding, artificial general intelligence, artificial comprehension, artificial conscience, etc., spurring the further creation of an authority figure to regulate AIS certification procedures and seals.

### 3.7.2 Analysis of the Principles Represented by the Page Rank's Top Scored Nodes

The most important nodes in the network represent principles with the AI beneficial scope, AI human and ethical dimensions, several variables contributing to fairness, and risk management obligation as the main factor for principled AI.

The highest scored principles, with 0.0113 and 0.0110, were principles P06D01 and P02D27, respectively. The former approaches the maximization of the benefits AIS use should bring to society and highlights AIS's development challenges, while the latter stresses the human dimension of the use and development of AIS. Both topics are central to the very scope of the principled AI international framework. For that reason, we believe their centrality on the network, responds to a generalization-specialization type of relation with the other nodes they are connected to on their clusters and the network in general. Specifically, P06D01 suggests understanding benefits of AIS and points at the related challenges in a set of variables that are crossed listed with other top scored principles as can be seen here below. Likewise, P02D27 shares the same variables and outlines a human focus to them. The general scope for both principles has gained them their relevance over the remaining principles and their description may be broken down as follows:

- 1). Protect the privacy and security, dignity, freedom, autonomy, and rights of individuals.
- 2). Understand and respect the interests of all parties that may be impacted by AI advances.

- 3). Ensure that AI research and engineering communities remain socially responsible, sensitive, and engaged directly with the potential influences of AI technologies on wider society.
- 4). Ensuring that AI research and technology are robust, reliable, trustworthy, and operate within secure constraints.
- 5). Opposing development and use of AI technologies that would violate international conventions or human rights.
- 6). Promoting safeguards and technologies that do no harm.

In the first place, for item 1), there are several mechanisms within the software development industry implemented to ensure privacy and security. The latest trends define both terms as quality variables (Bures et al., 2021); Tahaei, Frik & Vaniea, 2021) Therefore, privacy and security do not impose bigger challenges from a methodological perspective as they do not the protection of individual's dignity, freedoms, autonomy, and rights. Each of those variables needs to be operationalized so that they can be further measured, converted into parameters, and modelled in a way that software testers and auditors are able to verify and validate throughout the project lifecycle. In order to orchestrate an adequate symbiosis between their analysis, design, and implementation, and their verification and validation, several artifacts should be built including, but not limited to, dedicated checklists and metrics, the definition of the threshold conditioning formal changes requests in medium and large size projects, etc.

Second, items 2) and 3) entail bigger challenges. On the one hand, a letter of understanding might be necessary to evidence how the impact of AIS over the interests of all involved parties were discussed and agreed. This will determine the stakeholders, clients, and development team's responsibilities, as well as the remedy actions if required. As neither of the parties should be judge, there should be a third party with the authority to evaluate whereas the letter considers the impact upon the most affected population sector, usually left out in such considerations, and that in fact all interests are adequately addressed. This authority figure can be easily appointed from the external

service hired by the client to evaluate the quality of the software that is being developed in response to client's requirements. For the case of small projects, or projects that follow an agile methodology, this suggestion needs to be revised. Alternatively, item 3) can be addressed when quality assurance companies widen their efforts to certify, for example, that the software development (specially the AIS development) has being conducted in accordance with the listed variables in the item. There exist previous efforts (SEI, 2010) in which development practices have been contrasted and certified, signifying a prestigious quality measure. In this case, if a certified development team experiences a change on their member distribution, a new certification process should be required for the new team. Therefore, a threshold needs to be defined to determine how the change might affect losing a pre-existing certification in the proposed model. A scholarly facet is promoted by the principle P18D02 when highlighting that educational organizations should include ethics, and other topics like security, privacy, and safety, as integral parts of curricula on AI, machine learning, computer science, and data science. An initial exploratory study on the competences being currently taught on AI, and ethics and society can be found in the literature (Suarez & Varona, 2021), where it was found a gap in the training of needed ethical skills along 503 courses offered across 66 universities from 16 countries.

Also, there are some good practices for software development defined to ensure and validate robustness, reliability, and compliance with given constraints related to the business being modelled. This needs to be revised in correspondence to the research portion item 4) is referring to. Additionally, the mentioned variables need to be revised in the context of AIS. In relation to trustworthiness, this means being coherent with our idea of not dealing with trustworthiness as a single non-functional variable, but as part of the business model in software engineering. We highlight the conceptualization of the trustworthiness variable and explore to what extent it can be modeled as a model guiding the software development process, especially AI solutions or solutions with a component of AI.

Lastly, items 5) and 6) intend to ensure that even when failing to address the previous items, the produced system does not violate international conventions, human



rights, or harms people in any way. These components would be needed during the conceptualization phase of the project and reflected on the letter of understanding referred earlier, as it is at this moment that international conventions that might be contravened by the software being produced after deployed need to be specified and described. Having determined the international conventions related to the business being modelled by the project and the updated list of internationally agreed upon human rights, both parameters should be turned into checklists that auditors and software testers need to execute during the stages of requirements analysis, design, and testing. The software flaws identified from the referred checklist execution should be reported and trigger formal change petitions to the project base line. When the flaws are identified after the software has been deployed and client have performed remedy actions, the development team should develop fixes for the identified flaws. We are aware these suggestions have an economic impact on both the client and the development team, and that is the reason why we encourage that all responsibilities be defined from the project conceptualization stage. Additionally, we also encourage that the suggested steps modelled as part of the support processes be included in the software development process, so all actions, artifacts, and conditions are defined beforehand. The position “to do no harm” has an independent dedicated principle, P06D28 (Sweden), expressed through software features like safety and security, to which we are committed.

The human dimension posed in the previous principles is extended in the principles P03D30 (United Arab Emirates), P02D23 (Organization for Economic Cooperation and Development OECD), P02D24 (G20), P04D33 (European High Level Expert Group on AI), P01D32 (Japan), and P03D15 (Spain). Their descriptions can be broken down as:

- 7). Nations should collaborate to avoid an arms race in lethal autonomous weapons, and such weapons should be tightly controlled.
- 8). Active cooperation should be pursued to avoid corner-cutting on safety standards.

9). Ensure non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights.

10). The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice.

11). Humans interacting with AI systems must be able to keep full and effective self-determination over themselves and be able to partake in the democratic process.

12). AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans.

13). Develop, utilize, and implement AI in society to expand the abilities of people and allow diverse people to pursue their own well-being.

14). Introduce appropriate mechanisms for literacy education and for the promotion of proper use of AI so that people do not become over-dependent on AI or misuse AI to manipulate other people's decision-making.

We want to point out that we consider items 7), 8), and 14) to be more linked to the regulative and normative efforts of governments than being directly linked to the software development and thus we will not consider them here.

Also, the variables proposed by item 4) on the previous list to achieve trustworthiness are, this time, enriched by variables proposed by item 9). Similar to our suggestions on the previous variables, we encourage the identification of which international labor rights might be affected by the outcomes of the AIS being built, adding them to the aforementioned letter of understanding and to act accordingly. Additionally, non-discrimination, fairness, and social justice should be standardized so every party involved in the auditing, formal revision, or testing of the software development process and its artifacts have an agreed upon set of definitions to work on, similar to what the international norms (International Standardization Organization and

International Electronics Commission, 2014) do by dedicating one to the standardization of the language and notions the remaining related norms are based on. Equality and diversity need to be operationalized so that they can be incorporated to checklists focused on ensuring them as quality variables and modeled in metrics so they can be objectively measured.

Item 10) presents itself against complete AIS autonomy while making decisions. This is easier to achieve in information management systems as they are built as part of a model responding to a socio-technical process (Sanchez-Gordón, 2021) or set of processes. In contrast, AIS is commonly used within contexts where the understanding of the interaction of all business actors is limited, or complex. That is the reason why, in the AIS context, extracting information and applying knowledge out of examples and patterns make more sense. While socio-technical information management systems are process driven, AIS are data driven. That being said, it will likely require a log tracking service running parallelly to the main AIS, trained to determine if the main AIS have reached a checkpoint where the intervention of a human is desirable, and react accordingly. That will might create some concerns on the client's end and impose an obstacle on the implementation of this suggestion. This leads us to another concern as for who will determine when the AIS has reached a point where the human intervention is desired given that the training probably will only highlight the several points where human intervention can be introduced.

Item 11) might be seen as a complement to suggestions from the previous paragraph. While item 10) defends the human autonomy by limiting AIS's, item 11) emphasizes that responsibility must be retained by humans. We share the vision of AIS as a decision-making support system, and the human being the ultimate decision maker. This is more complicated than cataloging all software as socio-technical solutions, and IAS as two opposite poles of a single dichotomy. There are contexts in which human interaction to reach a decision is counterproductive when a decision is made based on parameters for example, when stabilizing balance in the context of walking rescue robots, or vessels' cargo loading/downloading automated system. Therefore, AIS aided decisions

and responsibility are issues that need further research and better understanding. Principle P05D12 (Canada) connects this item with the idea of fairness that is discussed below.

Variables like human subordination, coercion, deception, manipulation, condition, and other associated terms presented in item 12) must be formalized mathematically so a scale can be built. The adoption of the principle demands the determination of different scales to help assessing if an AIS is being unjustifiably subordinating, manipulating, etc. humans. Those scales are not a fixed one for all kind of scale, and must be adjusted for each business being modelled, as different contexts will have different thresholds. A way of building the mentioned scales might be achieved through the risk management subprocess and involving as many stakeholders as possible given the context of the solution being modeled, so the development team can design adequate risk response strategies, accordingly.

Finally, item 13) centers on the pursuit of well-being by means of the develop and use of AIS. Given the several meanings for well-being, and how closely related to the context they are, it would be helpful to have an agreed-upon definition to help software engineers evaluate when they are designing a solution that will improve the sense of well-being of the users or the target population, and when the deployed solution is positively impacting their well-being. Consequently, like the other variables, well-being needs to be managed as a dependent variable on the discriminatory or non-discriminatory nature of decisions based on decisions, predictions, and/or recommendations proposed by AIS, accordingly operationalized, and planned as part of the quality assurance activities in the development process. Again, that suggestion may impose an obstacle in projects where an agile methodological approach is followed.

In addition to the human dimension described above, P05D27 (China) and P01D30 (United Arab Emirates) establish an ethical feature showcasing ethical design approaches as the paths towards trustworthy AI, which is mainly framed, once their description is broken down, by the following elements:

- 15). Developing fairer systems.

- 16). Reducing possible discrimination and biases (including algorithm operational biases).
- 17). Improving the system's transparency, explainability, and predictability, and making them more traceable, auditable, and accountable.
- 18). Ensuring representative and biasfree datasets.
- 19). Mitigating the risks inherent in the systems being designed.

Items 15) and 16) underline the three main variables that can be used to group all other software quality features within the context of our research problem. Hence further exploration is required to gain more comprehension upon their role in the assurance of trustworthy AI, and the definition of the conceptual vicinity for other variables that are also related with these (discrimination, bias, and fairness) in the context of AIS. The mentioned conceptual map can serve as a frame of reference for software developers at every stage (including maintenance) of the development process. An effort in this direction can be found in Chapter four.

The recurrence of variables like the ones listed by item 17) implies that there exists a set of characteristics that need to be included in the quality assurance process of software projects, especially when developing AIS solutions. The current quality characteristics for software systems, standardized by ISO-IEC (International Standardization Organization and International Electronics Commission, 2014) as an update by the International Standardization Organization ISO and International Electronics Commission (2001) are still limited to the functional dimension of features such as functionality, liability, usability, efficiency, maintainability, and portability. All of these are strictly product centered. However, as can be inferred from recent concerns (Greene et al., 2019) and recent efforts to address those concerns (Fjeld et al., 2020; Hagendorff, 2019; Jobin et al., 2019) regarding the ethical implications of the software outcomes and their impact on human life, it is evident the need for a more human centered set of quality characteristics to complement the product-centered functional dimensions of the available software quality characteristics with their social dimension .

Other principles like P06D32 (Japan), P04D15 (Spain), P04D23 (Organization for Economic Co-operation and Development OECD), P04D24 (G20), P04D07 (European Commission for the Efficiency of Justice CEPEJ), P04D22 (Canada), P03D19 (UK), P06D36 (Vatican) can be considered specializations of this item and therefore are discussed.

Aside from our previous suggestion to include these features as quality characteristics for software products, we would like to clarify that supplementary work is also needed. Just to illustrate an example orchestrating most of the variables in item 17), we will refer to the definition of information management flows associated with the use of AI systems including the necessary elements (access policy to which piece of information, period of time the information will be available, for example) and moderating the communication between the stakeholder and the decision maker (regardless of the latter) to be incorporated (ex officio) in report modules, helping those adversely affected by an AI system supported decision to obtain relevant information and details of the decision they were object of.

In regard of item 18) we acknowledge the efficacy of the available mechanisms directed to ensure representativeness of data, along with other data management associated issues like noise, duplicity, etc. (Wachter & Mittelstadt, 2019). Hence, we would rather put our attention on the biasfree-portion of the item. To that purpose, we must point out that there are three main stages in which developers can influence datasets to be as biasfree as possible: the data collection period, the data cleaning and pre-processing (let us call it preparation) period, and the evaluation period, when the preliminary results show the first hints of bias. As per the scope of the present study we do not expand on the topic as there is plenty of specialized literature about it. We need to stress that, overall, the philosophy of a biasfree dataset with the current limitations, is reduced to building a system capable of bring neutrality to data, regardless of the existence of biases in the dataset, unless dealing with new reality, new data, and new untrained models.

Then, item 19) represents one of the easiest elements of the studied principles to be adopted as a methodological reference for software engineers, who are extensively trained in risk management. The principle P04D27 (China) includes on its idea of risk management all required actions directed to achieve the items previously discussed.

Finally, P06D27 fosters diversity and inclusiveness among the development team as a mean to benefit people who otherwise could be easily neglected or underrepresented in AI applications. We are certain that this approach enriches the currently available trends framing contemporary studies of the issue of software project staffing, especially those with particular interest in candidate's personal traits.

### 3.8 Conclusions of the Chapter

The present study expands on a previous exploration of the Principled AI International Framework, presented in chapter two, and indicates that -as suggested by the first exploration of the framework-, the language as a disruptive element in the adoption of the proposed principles. The analysis documented in this chapter also highlights several challenges software engineers might face when taking the mentioned regulatory framework as a methodological reference to produce software products, specially ADM systems, such as (1) the principle's ambiguity, and (2) the need for convincing the project's client of the need for developing technical components outside their direct interest that are strongly linked to fairness related requirements, to provide examples.

The principle represented by the P06D01 nomenclature is found to be the principle whose description, written in too general terms, have a generalization-specialization type relationship with other principles it is connected with, as the majority of them can be traced back to subsections of that one. Therefore, the principle could be emphasized as one of the cores of the regulatory framework being analyzed, and its associated challenges should be dealt with priority if a set of priorities are to be established by software engineers adopting the norms as a methodological reference when developing AIS. These suggestions are articulated along the different capabilities and maturity levels the proposed model in chapter five is built on.

Similarly, the principle represented by the P02D27 is found to the one that is closest to the purpose of the regulatory framework as a whole, given it is described using most of the top n-grams describing the principle's enunciation and description corpus. That also gives the principle certain universality within the studied norms. As a result, its associated challenges should be equally prioritized when determining a methodological mechanism to help implement the regulatory framework. Considerations that were articulated through the design of the specific goals of the proposed model in chapter five.

The discussion about the most relevant principles, according to the Page Rank's measurement, allowed us to draw a set of challenges for software engineers, that



included: the need for distinct variable operationalization; definition of new dedicated metrics; determination of several thresholds triggering formal change petitions, or providing a scale for artificial intelligence or human understanding; the creation of new structures to rule the development in respect with complying with these principles; and the revision of existing mechanisms within the software engineering management procedures that may be improved to include the norms' requirements. Suggestions that are incorporated as specific practices of the proposed model in chapter five.

The study critiques the outdated quality characteristics approach standardized by the IEC/ISO 9126, which place the focus of the characteristics solely on the product and ignores the impact of such systems on the people, specially to their human rights, values, and freedoms. It also proposes a revision of the norm considering the inclusion of some of the discussed variables as quality features.

The study recommends a further analysis of terms like discrimination, bias, and fairness. It was determined that these variables include others, which are more specific, so it can be said that they have a more general scope. Thus, discrimination, bias, and fairness can be treated as non-functional requirements in software development, specially AIS projects, addressed by the assurance of more specialized variables that are connected to them. Additionally, because of their more general scope, the study also recommends to research how these three variables fit in the idea of trustworthy AI as a business philosophy, suitable for software development. The suggested study is presented in chapter four.

### 3.9 Works Cited in the Chapter

- Access Now Organization. (2018). *Human rights in the age of AI*. AccessNowOrg.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1), 361-362. Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/13937>.
- Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, vol. 22, n° 6, 705-730.
- Blondel, V.D., Guillaume, J.-L. Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, n° October.
- Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, vol. 29, n° 4, 555-564.
- Brin S. & Page L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the Seventh International Conference on the World Wide Web (WWW1998)*.
- Bures, M., Klima, M., Rechtberger, V., Ahmed, B. S. Hindy, H., & Bellekens, X. (2021). Review of specific features and challenges in the current internet of things systems impacting their security and reliability. paper accepted at *WorldCist'21 - 9th World Conference on Information Systems and Technologies*.
- Buruk, B., Perihan, E. E. & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, vol. 23, 387-399.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, vol. 5(2), 153-163.
- Collection of Institutional Authors. (2018). *AI standardization white paper (CESI)*. National Standardization Management.
- De Meo, P., Musial-Gabrys, K., Rosaci, D., Sarne, G.M.L. & Aroyo, L. (2017). Using centrality measures to predict helpfulness-based reputation in trust networks. *Trans. Internet Technology.*, vol. 17, n° 1, 20.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society.
- Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, vol. 40, n° 1, 35-41.
- Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. *Software: Practice and Experience.*, vol. 21, n° 11, 1129-1164.
- Future of Life Institute. (2017). ASILOMAR AI Principles. *ASILOMAR Conference on Beneficial AI*.

- Gong, R. (2020). Transparent privacy is principled privacy. *arXiv*. arXiv:2006.08522
- Greene, D., Hoffmann, A. L. & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings 52nd hawaii international conference on system Sciences*.
- Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30, 99-120.
- House of Lords Select Committee on Artificial Intelligence. (2018). *AI in the UK: Ready, willing and able?*. Authority of the House of Lords.
- International Standardization Organization ICO and International Electronics Commission. (2001). *ISO/IEC standard 9126:2001 software engineering — Product quality*. ISO.
- International Standardization Organization and International Electronics Commission. (2014). *International standard ISO/IEC 25000: Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — Guide to SQuaRE (Second edition)*. ISO.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.*, vol. 1, 389–399.
- Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.
- Madaio, M.A., Stark, L., Wortman Vaughan J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
- Marchiori, M., & Latora, V. (2000). Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, vol. 285, n° 3-4, 539-546.
- Marshall, J. M. (2021). Technoevidence: the "Turing limit" 2020. *AI & Society*.
- Miller, K. (2020). A matter of perspective: Discrimination, bias, and inequality in AI. In *Legal regulations, implications, and issues surrounding digital data*, 182-202, IGI GLOBAL.
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, vol. 1, 501–507.
- Pinar, S., Cicekli, A.I. & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines*, vol. 10, 463-518.
- Prateek, S., Pasala, A. & Aracena, L.M. (2013). Evaluating performance of network metrics for bug prediction in software. *20th Asia-Pacific Software Engineering Conference (APSEC)*.
- Sánchez-Gordón, M. (2021). Connecting the dots between human factors and software engineering. *Latin American Women and Research Contributions to the IT Field* 17.

- Saner, M. & Wallach, W. (2015). Technological unemployment, AI, and workplace standardization: The convergence argument. *Journal of Evolution and Technology*, vol. 25, n° 1, 74-80.
- Software Engineering Institute SEI. (2010). *CMMI for development version 1.3*. Carnegie Mellon University.
- Suarez, J.L. & Varona, D. (2021). *The ethical skills we are not teaching: An evaluation of university level courses on artificial intelligence, ethics, and society*. Social Sciences and Humanities Research Council.
- Suzuki, S., Aman, H., Amasaki, S., Yokogawa, T., & Kawahara, M. (2017). An application of the pagerank algorithm to commit evaluation on git repository. *43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*.
- Tahaei, M., Frik, A., & Vaniea, K. (2021). Privacy champions in software teams: Understanding their motivations, strategies, and challenges. *CHI Conference on Human Factors in Computing Systems (CHI'21)*.
- Telefónica. (2018). *AI principles of Telefónica*. Telefónica.
- The Public Voice Coalition. (2018). *Universal guidelines for AI*. The Public Voice Coalition.
- The Standish Group. (2020). *CHAOS Report 2020*. The Standish Group.
- Treasury Board of Canada Secretariat. (2018). *Responsible artificial intelligence in the Government of Canada. Digital disruption white paper series*. Treasury Board of Canada Secretariat.
- Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. Chicago: University of Chicago Press.
- Varona, D. (2018). La responsabilidad ética del diseñador de sistemas en inteligencia artificial. *Revista de occidente*, n° 446-447, 104-114.
- Varona, D. (2020). AI systems are not racists just because. *T-13 hours: Building Community Online in CSDH/SCHN2020*.
- Wachter S. & Mittelstadt, B.D. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Bus. Law Rev*, 494–620.
- Yapo, A. & Weiss, J. (2018). Ethical implications of bias in machine learning. *Proceedings of the 51st Hawaii International Conference on System Sciences*.
- Yorks, L., Rotatori, D., Sung, S., & Justice, S. (2020). Workplace reflection in the age of AI: Materiality, technology, and machines. *Advances in Developing Human Resources*, vol. 22, n° 3, 308-319.

## Chapter 4

### Discrimination, Bias, Fairness, and Trustworthy AI

In this chapter we describe the analysis of the variable discrimination, bias, fairness, and trustworthiness. As identified in chapter three, there exists a set of specialized variables, such as security, privacy, responsibility, etc., that are used to operationalize the principles in the Principled AI International Framework. These variables are defined in such a way that they contribute to others, of a more general scope, like the ones studied in this chapter, in what was defined in chapter three as a generalization-specialization relationship. Our aim in this study is to comprehend how we can use bias, discrimination, and fairness as an intermediate layer between the thesis research project's main goal (Trustworthy ADM solutions) and the variables that will be assured during software project's lifecycle (security, privacy, responsibility, etc.). Bias, discrimination, and fairness are mainly approached with an operational interest by the Principled AI International Framework, so we included sources from outside the framework to complement (from a conceptual standpoint) their study and their relationship with each other.

#### 4.1 Introduction

The negative implications associated to the evolution of ML, and by extension to AI systems, and the fact that algorithms and models are increasingly complex and less explainable, make it difficult for users/auditors/developers/researchers to identify if AI systems produce outcomes with negative consequences for humans. However, the most disturbing factor in this evolution of AIS is to learn that we keep outsourcing our responsibility over our decisions to the software we use.

There exists a palpable need for auditing black box algorithms, not only from the verification and validation processes staged as part of the software project lifecycle, but also from other areas like policymaking. Both the engineering approach and the stipulation of the regulatory approach needs to be incorporated into an integrative mechanism oriented to reduce and mitigate ADM systems-produced discriminatory

outcomes, analyzed in chapters two and three. The traditional approach conducted to manage discrimination, prejudice or bias, and algorithmic unfairness historically exhibits a reactive character that must be overcome, as criticized in chapter one. Additionally, the needed proactive approach must incorporate the determination of possible remedy actions due to discriminatory ADM systems' outcomes. Then it is not only necessary to coordinate efforts for mobilizing professionals from multiple disciplines of the technical and humanistic fields to reach a better understanding of the problem and its solution, but also for standardizing their language to achieve a more effective comprehension upon the actions that must be deployed to attaining trustworthiness in the AI systems.

The Principled AI International Framework could be considered part of the attempt to specify the idea exposed in the previous paragraph in the direction of achieving trustworthy AI as a business model, specifically centered in safeguarding the individual's rights that might be affected by decisions produced by flawed AI solutions. However, previous findings described in chapters two and three highlight that there are significant differences along the language used in the regulatory documents forming the referred framework, which can compromise its proper implementation. For instance, the multiple definitions of the objective variables trustworthy AI are intended to be founded upon are listed among the described gaps.

The present study expands on those previous analyses and over the Principled AI International Framework itself where divergences in language, among other difficulties regarding the framework assimilation as a methodological reference for AIS development, were highlighted. We thought it would be pertinent to explore what extent the variables among the principles and their agency to propitiate trustworthiness in AIS were compromised due to the ambiguities and lack of precision in the use of language. We then conducted an exploratory survey on the notions of discrimination, bias, and fairness to comprehend how they can be articulated in the pursue of trustworthiness, in the context of ADM systems.

## 4.2 Related Research

Abhishek and others proposed a framework for trustworthy AI systems (Abhishek et al., 2020) providing a data-centric level of abstractions for ethical consideration within the AIS and Data Science contexts that would encompass three levels: data, algorithm, and practice. They defined a set of requirements for trustworthy AIS design. The intended variables coincide with other related studies, although the same cannot be said about some of the proposed definitions, an issue we have already encountered in the principled AI framework. The same happens with others (Brundage et al., 2020; Wickramasinghe et al., 2020) and the great majority of the referenced research.

The referenced studies concur on their method, which usually consists of a bibliographic survey determining and defining the variables each paper presents and should be used to build trustworthiness on AI systems. These studies (Abhishek, 2020; Brundage et al., 2020; Smith, 2020; Wickramasinghe et al., 2020) distinguish themselves by proposing mechanisms to support the trustworthy AI design by incorporating the variables they each define according to their respective conceptualizations.

These differences might seem trivial, however in a context totally dependent of the operationalization of the objective variables, this is an aspect that gains major relevance. In fact, it determines the success of the proposed mechanisms and the subsequent achievement of trustworthiness when there is no ambiguity on their implementation.

One of the more critical positions in this approach (Mittelstadt, 2019) stresses that the principles show deep political and normative disagreement, and highlights that the studied principled framework for AI development lacks common aims and fiduciary duties, proven methods to translate principles into practice, and robust legal and professional accountability mechanisms when comparing it with other frameworks used in fields like healthcare. This is indeed a very strong criticism of both of the principled framework itself and the ways in which can be operationalized in real contexts of AI software development and assessment.

### 4.3 Analysis of the Variable Discrimination

Automated learning aims to mimic some of the learning natural processes existing in nature, the difference being that in automated learning the learning is mainly based on a set of examples rather than following defined indications and rules that describe a given context. Similar to what happens with humans, ML often produces predictions and recommend decisions that end up being discriminatory to individuals or groups.

Among the available definitions of “Discrimination” in the context of ML and AI systems (Verma & Rubin, 2018), is Verma and Rubin’s approach describing discrimination as the direct or indirect relation between a protected attribute and the resulting prediction/classification/suggested decision. This is seconded by Mehrabi et al. (2019) where direct discrimination is distinguished by the direct relation between protected attributes and the produced prediction/classification/decision with a negative consequence for the object being targeted by the decision. It expands by declaring that indirect discrimination not only relates to an indirect relation between the mentioned taxonomy but is also manifested when the implicit effects of protected attributes are considered. For instance, the use of an individual postal code in loan and insurance premium calculations are two examples showing how apparently less sensitive individual features may lead to a discriminatory decision.

According to Zhang, Wu & Wu (2017), residential areas often offer a representative distribution of its inhabitants in regard to attributes like race, household income, etc. However, the zip code is not usually a protected attribute in the decision-making process because the law does not register it as a feature triggering discriminatory decisions, like other features as race or gender. In the literature, it is stated that a set of attributes the law suggests being treated as protected are exhibited in an attempt to help avoiding discrimination in the aforementioned scenarios and others such as recruitment (Jiahao et al., 2019). These examples allowed us to understand that discrimination is a variable that need to be dealt with casuistically, in every new project, for every new and old scenario, across cultures.



It can be said that Discrimination, in the context of ML and AI systems, has a statistical root when the information learned, by means of pattern discoveries, frequency measure, correlations among attributes, etc., about a group is used to judge an individual with similar characteristics. Hence the importance of data and data collection procedures carried out according to the scope of the intended decision or prediction.

The continued use of statistical methods in decision-making and/or the arrival of predictions leads to a systematization of discrimination. Therefore, it can be understood that ML has scaled the impact of discrimination, and "unintentionally institutionalized" these discriminatory methods through AI, and it has created a perpetual cycle where the object of discrimination itself becomes part of the knowledge base used in subsequent estimates, that, hence, become equally discriminatory. That is, a recommending software used within an enterprise with a given gender distribution will tend to reproduce the same unbalanced current gender distribution in their selection process while hiring new candidates. The referred distribution might not only be fit in correspondence to the enterprise's training base but also in correspondence with available knowledge about the top performers distribution in the guild the particular enterprise is part of what will result in perpetuating the gender distribution in the workforce and conditioning future hiring if the same method is used over time. This is the reason why discriminatory decisions are nowadays generally attributed to prediction, selection/ estimation algorithms, etc. (Jago & Laurin; Loi & Christen; 2021) and not to other aspects equally important like data gathering, data cleaning and data processing, as an example.

Mehrabi et al. (2019) adds that discrimination can be classified as explainable and non-explainable according to the possibility of justifying or not justifying the produced decision/prediction from the triggering attributes. That is, explainable discrimination is close to what we understand as prejudice, where there is a clear parameter influencing the discriminatory decision or prediction. While non-explainable discrimination happens when there is a discriminatory outcome that cannot be justified, the specific trigger cannot be identified. Either classification lacks ethical support.

Another study (Schmidt et al., 2019) conceptualizes discrimination in the context of ML and AI systems similarly to Mehari's study, and it justifies the use of these unintended discriminatory AI models by providing two main reasons: first, the model is able to provide a decision/prediction according to the need of the business; and second, the lack of a less discriminatory alternative model. This simply represents an attitude of resignation and acceptance of discrimination and the subsequent bias.

Also, in the literature (Martínez-Plumed, Ferri, & Nieve, 2019), discrimination is defined using six classifications for bias: (1) sample or selection bias, when the sample representativity gets compromised with significant unbalance; (2) measurement bias referring to systematic errors regarding data correctness, compromising the values supporting the estimations; (3) self-reporting (survey) bias, related with the completeness of data; compromising the statistical significance and the accuracy backing the predictions; (4) confirmation (observer) bias, resulting from the researcher own prejudice while he or she information backing his or her working hypothesis; (5) prejudice (human) bias, when the model/algorithm result reflects a pre-existent bias on the knowledge base used for training; and (6) algorithm bias, when the model/algorithm creates or amplifies bias from the training dataset in an attempt for overcoming processing needs, what is usually true when working with multiple samples of different sizes.

As can be appreciated, discriminating upon the characteristics of an object is not intrinsic to humans. Technology reproduces and amplifies such behavior. The specialized literature exhibits a tendency to hold machine learning algorithms accountable for the problem created by their inability to adequately deal with bias, as analyzed in chapter one; however, the data used in training, and the data collection methods are equally responsible for discriminatory predictions and recommendations.

Lastly, it can be highlighted that discrimination has both an origin and cause of bias, once the outcomes of today's discriminatory decisions based in yesterday's biases, populates tomorrow's datasets. In the field of the software industry, both variables: discrimination and bias, are closely related because of the speed at which the whole cycle

occurs, and because of cycle's many iterations. The following section presents bias as variable of analysis.

#### 4.4 Analysis of the Variable Bias

Similar to what happens with human prejudice, the bias in ML leads to discriminatory predictions and recommendations. Consequently, many researchers are pursuing optimization of the methods in which ML identifies and eliminates bias. There are two marked methodological trends on that regard. The first trend pertains to algorithm calibration (Chouldechova, 2017; Feldman et al., 2015; Fish et al., 2016; Hardt et al., 2016; Lazar Reich & Vijaykumar, 2020; Pedreschi et al., 2007; Solon & Selbst, 2016; Zafar et al., 2015), while most recent trends (Holstein et al., 2019; Varona, 2018, 2020a, 2020b and 2020c; Veale et al., 2018) are trying to tackle the problem from early stages of AI algorithms/model's design.

Among the documents forming the Principled AI International Framework (Fjeld et al., 2020), the UNI Global Union 2017 report (UNI Global Union & The future world of work, 2017) describes bias as the action of using features like gender, race, sexual orientation, and others, as discriminatory elements in a decision with a negative impact somehow harmful to the human being. Then, the difference of bias with respect to "Discrimination" is that "bias" represents the action while discrimination manifests itself in the result, of using certain attributes in the decision-making process. The dependence among these two variables could be located in this relation. It is also important to note that such a definition emphasizes the negative impact of the decision so that it seems to not to consider "bias" when such effect might be positive.

Another report, authored by the G20 (Abreu et al., 2018), describes bias as the product of human activity with a given effect on individual rights and other contexts inherent to humans, while it declares that algorithms can unintentionally produce both bias and discrimination. The report also highlights the existence of two types of sources for bias: the method, either in the design of the algorithm or in the way the data is collected; and in the distortion/corruption of the data used as the training basis for the model/algorithm.

We suggest the existence of two referents for the definition of bias within AIS: the statistical referent, and the social referent. In that regard, Access Now Organization (2018) presents the statistical referent as the distance between the AIS produced estimation/prediction and the actual occurrence of the estimated/predicted event. It explains that, when there is statistical bias there is evidence that the data represents a social bias, what is described as social bias by the same report.

Then, it is accurate to say that we are in the presence of an unfair dataset every time that a discriminatory or biased conclusion is drawn, and that any instance of an algorithm using that dataset for training will produce equally unfair decisions and predictions. That does not mean the same happens in the opposite direction. The fact that an algorithm does not produce a discriminatory or biased decision/prediction does not indicate we are using a fair dataset. That, along with some related principles from the framework analyzed in chapters two and three, is the reason we suggest as a specific practice across our model (where applicable) the use of data pipeline dedicated frameworks, to stress and exhaust datasets being used for algorithm and model training.

In that respect, the obligation of fairness defined in by Access Now Organization, (2018) and The Public Voice Coalition (2018) first suggests the existence of two benchmarks for the definition of bias in AI. The statistical reference, expressed as the deviation of the prediction in contrast with the event's actual occurrence; and the social reference, from the evidence of statistical bias within the data representing a social bias. Second, it recognizes that decisions/predictions reflecting bias and discrimination should not be normatively unfair. This means that decisions which are unfair and reflect biases must not only be assessed quantitatively, but also evaluated with regard of their context - with a case-by-case approach. This is to understand how to avoid them and create a norm/standard rather than being the exception to the rule. And third, it clarifies that the single evaluation of the outcomes (previously mentioned algorithm calibration) is not enough to determine the fairness of the algorithm or model. This idea was first explored in chapter one. Consequently, Access Now Organization (2018) and The Public Voice Coalition (2018) proposes the evaluation of pre-existing conditions in the data that can be further amplified by the AI system before its design is even considered. This report

shows an inclination towards the emerging trend of recognizing in the data an origin for discriminatory and biased decisions, in contrast with the rooted trend of solely holding the algorithms accountable for the negative outcomes produced by AIS.

Also, the House of Lords Select Committee on Artificial Intelligence (2018) and Martinho-Truswell et al. (2018) criticize the methods of learning developed in machine learning, specifically how data is used during training. Per the House of Lords Select Committee on Artificial Intelligence (2018), while learning, systems are designed to spot patterns, and if the training data is unrepresentative, then the resulting identified patterns will reflect those same patterns of prejudice and, consequently, they will produce unrepresentative or discriminatory decisions/ predictions as well. Martinho-Truswell et al. (2018) highlights that good-quality data is essential for the widespread implementation of AI technologies, however the study argues that if the data is nonrepresentative, poorly structured, or incomplete, then there exists the potential for the AI to make the wrong decisions. Both reports define bias over the basis of misleading decisions produced from such compromised datasets.

Acknowledging the role of data in the introduction of bias is a relatively new approach.<sup>28</sup> Mehrabi's (2019) comprehensive survey provides several definitions of types of biases originated in the data. The author enriches upon the already mentioned historical and representation biases by providing further classifications. From the definitions provided by Mehrabi (2019), we thought pertinent to highlight the following due to the focus not on the data distribution per se but on the introduced bias resulting from a misuse of the dataset.

---

<sup>28</sup> This is different from the Garbage In Garbage Out (GIGO) approach to explain the relation of trashy data input with faulty outputs. The GIGO approach links specific data issues like duplicity of information, absence of information, and noise in information, just to provide a few examples; and bad programming with faulty output from systems. The relatively new approach of pointing out the datasets as an origin for discriminatory decisions refers to those datasets that even when not being trashy are biased and triggers discriminatory patterns in ADM systems. It is a new approach as the origin of discriminatory ADM systems' outcomes where mainly linked to biased algorithms, ignoring that datasets and the development team had a role introducing bias into the system.

First, we wanted to note Measurement bias, which takes place when using a particular feature of the object of the decision when building judgment, just because that feature has been historically over measured. This particular action has a fuzzy line with human introduced bias as it is explained later in the classifications provided by IBM.

The overall evidence shows that there exist some population groups that are more assessed and controlled (policed) than others, and therefore have higher rates of arrests if we use the example of recidivism and risk assessment within the judicial system, turning those populations into groups vulnerable to this kind of bias.

Second, we wanted to point the Evaluation bias, that compromises the model validation when using inappropriate and disproportionated benchmarks in the verification process. The IJB-A benchmark known as the “Face Challenge” in face recognition was used to exemplify the matter because of its failures when considering skin color and gender.

There were four particularly interesting biases described in the study. First, Aggregation bias, when false assumptions are made because of the use of conclusions produced by previously flawed models; The Simpson’s Paradox related bias, referring to the different bias appreciations when looking at different data groupings within the analyzed dataset; the Linking bias, which arises when variables like network sampling, method of interaction, and time are not considered when building a network around the object of the decision; and what they denominate Emergent bias, resulting of the user experiences with deployed products through the graphical user interface, where possible habits of prospective users were estimated from the design stages.

IBM (2019) adds a human edge to the binomial data-algorithmic bias origin while presenting a set of unconscious bias definitions expressed in terms of their manifestation among the general population that engineers need to be consciously aware of when designing and developing for AI. Despite the IBM’s classification in three main focus areas (Shortcut biases, Impartiality biases, and Self-interest biases) we group those definitions in three main points of interest of project management as presented below. This new organization fits the context of our research as it moves the focus of the IBM’s

classification from the individual to the project stage in which such biases can be introduced.

#### 4.4.1 The First Point of Interest is Project Conceptualization

We gathered the IBM's Sunk Cost bias and Status Quo bias definitions under the project conceptualization point of interest. They both refer to the tendency to justify past choices and to maintain the current situation, even though they no longer seem valid or when better alternatives exist. In that sense AI practitioners need to be aware every new project involves a unique business reality. Some highly specialized teams will try to accommodate their expertise rather than study emerging methods when designing their solution approach. Sommerville (2015) and the CHAOS report (The Standish Group, 2020) stressed that issue as one of the main causes of project failure. Deciding a wrong project approach could be the first step onto an unfair AI system.

#### 4.4.2 The Second Point of Interest is Project Design

We gathered the IBM's Not Invented Here bias, Self-Serving bias, and bias Blind Spot definitions under the design point of interest. We also divided this point of interest into two subcategories: Data affairs and Algorithm functioning affairs as described below.

The Not Invented Here bias and the bias Blind Spot are somehow connected. The former refers to the aversion to contact with or use products, research, standards, or knowledge developed outside the own group; and the latter refers to the tendency to see oneself as less biased than others, or to be able to identify more cognitive biases in others than in oneself, something that might exhibit a cause-effect relation. The Self-Serving bias states the tendency to focus on strengths/achievements rather than on faults/failures. This suggests that AI practitioners should avoid discriminating against pre-existent approaches which could save significant amount of time and effort, and provide valuable knowledge based not only on proven hypotheses but on errors or rejected hypotheses as well.

#### 4.4.2.1 Data Affairs Subcategory

Under the Data affairs subcategory we listed the Base Rate Fallacy, referring to the tendency to ignore general information and focus on specific information (a certain case) providing an individualistic opinion upon the decision's object. This is somehow related to the idea of stepping afar from generalizing based on previously available knowledge given a group of subjects sharing some of their traits with the object of the decision. And, the Availability bias, that focuses on overestimating events with greater "availability" in memory, influenced by how recent, unusual, or emotionally charged those memories may be.

#### 4.4.2.2 Algorithm Functioning Affairs Subcategory

On the other hand, we listed the Congruence bias, Empathy Gap bias, Anchoring bias, and Bandwagon bias under the Algorithm functioning affairs subcategory.

The Congruence bias represents the tendency to test hypotheses exclusively through direct testing, instead of testing alternative hypotheses. This approach ignores other variables that might affect the business being modeled, overlooking possible scenarios where the algorithm/model might behave different regardless of the tested hypothesis's outcome.

Similarly, the Empathy Gap bias represents the tendency to underestimate the influence or strength of feelings, in either oneself or others. This and the Congruence bias can be connected whereas the inclination towards a given hypothesis ends up being accommodated.

Different from the Congruence and the Empathy Gap biases, the Anchoring bias relies almost entirely on one trait or piece of information when making decisions, usually the first piece of information that we acquire on the subject being targeted by the intended decision. It conceives a false illusion of objectivity, when we separate ourselves from untested assumptions, such as our hypotheses and our feelings. However, the resulting decision ends up being biased because of the probable unrepresentativeness of the used data over the reality being modeled.



Finally, the Bandwagon bias, portrays the tendency to do or believe things because many other people do. That kind of group thinking is wrong, because following the general norm (when making decisions) contrary to making a decision as an individual, might be forcing us to perpetuate bias. This is dangerous, because doing so avoids the needed paradigm rupture in given situations, where the general historically agreed upon decisions are outdated.

#### 4.4.3 The Third Point of Interest is Project Verification and Validation

We then gathered Confirmation bias, Halo Effect, and Ingroup/Outgroup bias under the project verification and validation point of interest.

The Confirmation bias explains the tendency to search for, interpret, or focus on information in a way that confirms one's preconceptions. It might represent the previously referred connection between empathy gap and congruence biases. Either way the introduced bias, in this case, is supported by underrepresentation of data used to reinforce one's own preconception.

The Halo Effect bias can be expressed by the predisposition of an overall impression to influence the observer. Positive feelings in one area causes ambiguous or neutral traits to be viewed adequately. This is not only important during the business modeling but also during verification tasks where the evaluator is too familiarized with the work being verified, measured, or audited.

The Ingroup/Outgroup bias, which describes the tendency or pattern to favor members of one's ingroup over outgroup members, favoring the institutionalization of bias.

Wrapping up the variable analysis, we can now state with support (Independent High Level Expert Group on AI, 2019; Smart Dubai Office, 2019), that bias can be perceived as an intentional or unintentional predisposition toward prejudice in favor or against a person, object, or position. It has multiple origins within the context framed by the AI systems. Such origins include information represented within the data, logic of

algorithmic functioning, engineering methods and practices for data collection, data processing, and algorithmic design; it also can derive from human intrinsic biases for both designers and prospective users, and the contexts in which systems are used.

## 4.5 Analysis of Variable Fairness

By definition, heavy methodologies for software projects helps developers and stakeholders to understand that efforts are needed along the software project lifecycle for verification and validation tasks. We can find several quality variables (Pressman, 2010; Sommerville, 2015) that software projects have proactively managed in an attempt to avoid unintended outcomes from the systems they produce. Nowadays, with the use of AI systems, and particularly ML models and algorithms (National Science and Technology Council & Committee on Technology, 2016), consequential decisions are being automatically generated about people. The automation of bias, the incapacity of AI systems to bring neutrality to the decisions they produce, the perpetuation of bias, and the amplification of the historical discriminations are leading to concerns about how to ensure fairness. On one side, software practitioners strive to prevent intentional discrimination or failure, to avoid unintended consequences, and to generate the evidence needed to give stakeholders justified confidence that unintended failures are unlikely. On the other side, policymakers work to regulate the design and consumption of such systems, so they are not harmful to human beings and that the necessary amendments are made in case they were required.

From a technical point of view, (Demiaux et al., 2017) fairness is defined as the actions performed to optimize search engines or ranking services without altering or manipulating them for purposes unrelated to the users' interest. Expanding that idea, in the UNI Global Union & The future world of work literature (2017) it is acknowledged that fairness tasks should be planned during the design and maintenance phase of software development, and that those tasks should seek to control negative or harmful human bias so that they are not propagated by the system.

Some studies (Independent High Level Expert Group on AI, 2019; T20, 2019) relate fairness to inclusion. For instance, (Independent High Level Expert Group on AI,

2019) stresses that fairness is expressed by means of inclusion and diversity by ensuring equal access through inclusive design and equal treatment. In (T20, 2019) it is stated that AI systems should make the same recommendations for everyone with similar characteristics or qualifications. In consequence, software developers and software operators should be required to test the deployed solutions in the workplace on regular basis to ensure that the system is built for purpose, and it is not harmfully influenced by bias of any kind—gender, race, sexual orientation, age, religion, income, family status and so on—exposing the variable character of fairness over time. The report also states that AI solutions should adopt inclusive design efforts to anticipate any potential deployment issues that could unintentionally exclude people. Both studies believe necessary the involvement of all affected stakeholders along the project lifecycle. This is a work philosophy that is shared by companies like Telefónica (2018), based in Spain, and one of the main telecommunication operators in Europe. Several of the techniques and metrics available describing how ML pursues fairness are mathematically formalized in the literature (Mehrabi et al., 2019; Verma & Rubin, 2018). A critical analysis of metrics and techniques like those formalized in both studies were criticized in chapter one.

A cultural attachment is also presented (Mehrabi et al., 2019) while defining the fairness variable when the authors state that different preferences and outlooks within different cultures condition the current situation of having multiple concepts for the term. The situation is aggravated by the fact that available definitions of fairness in philosophy, psychology, and computer science supporting algorithmic constraints are mostly based on Western culture. This led the authors to define fairness as the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making.

An even broader definition is being proposed by the Vatican (2020) while using impartiality to explain fairness. The Vatican's working concept gathers the development and consumption of AI systems when it says, "do not create or act according to bias" and it connects the outcome of working to ensure fairness with its human focus when it says, "safeguarding fairness and human dignity."

To wrap up the analysis of the fairness variable, we wish to point out that these studies (Mehrabi et al.; T20, 2019; UNI Global Union & The future world of work, 2017; Vatican, 2020; Verma & Rubin, 2018;) define fairness as the AIS's ability to treat all similar individual or groups equally, and as the AIS's inability to produce harm in any possible way. This is indeed a noble but still very broad definition, and it shows the lack of agreement among the scientific community to achieve a definition of fairness that can be widely accepted. The Indian National Strategy for AI (NITI Aayog, 2018) locates the issue of fairness at the forefront of discussion in academic, research and policy fora, something that definitely merits a multidisciplinary dialogue and sustained research to come to an acceptable resolution, and it suggests identifying the in-built biases to assess their impact, and in turn to find ways to reduce the biases until techniques to bring neutrality to data feeding AI solutions, or to build AI solutions that ensure neutrality despite inherent biases, are developed. In that regard, we need to stress that (Mehrabi et al. (2019) indicates it is crucial to understand the different kinds of discrimination that may occur given the numerous distinct available definitions of fairness.

The analysis evidences a steering of the majority of the elements describing machine learning's traditional approach (Chouldechova, 2017; Hardt et al., 2016; Solon & Selbst, 2016) to cope with bias and discrimination, moving away from its reactive character towards a more proactive style. Hence, it is appropriate to state that, in order to produce less discriminatory outcomes, in the context of AIS, the engineering focus needs to commute from fairness (as a nonfunctional requirement) onto trustworthy AI as a business model.

#### 4.6 Analysis of the Variable Trustworthiness

Several studies (Abhishek et al., 2020; Abolfazlian, 2020; Wickramasinghe, 2020; Wing, 2020; Smith, 2020) agree that it requires human agency and oversight, and the use of a set of overlapping properties to define trustworthiness in the context of AI systems development and consumption. Among the most frequent highlighted properties across the studied bibliography, the following can be found:

- 1) Reliability when the system does the right thing it was designed to and available when need to be accessed.
- 2) Reproducibility when the systems produce the same results in similar contexts.
- 3) Safety when the system induces no harm on people as a result of their outcomes.
- 4) Security when the systems are invulnerable or resilient to attacks.
- 5) Privacy when the system protects a person's identity and the integrity of data, indicates access permission and methods, data retention periods and how data will be destroyed at the end of such period, which ensures a person's right to be forgotten.
- 6) Accuracy when the system performs as expected despite of new unseen data compared to data on which it was trained and tested.
- 7) Robustness when the system is sensitive to the outcome and to a change in the input.
- 8) Fairness when the system's outcomes are unbiased.
- 9) Accountability when there are well defined responsibilities for the system's outcome so as the methods for auditing such outcomes.
- 10) Transparency when it is clear to an external observer how the system's outcome was produced and the decisions/predictions/classifications are traceable to the properties involved.
- 11) Explainability when the decisions/predictions/classifications produced by the system can be justified with an explanation that is easy to be understood by humans, while being also meaningful to the end user.
- 12) Other variables such as Data Governance, Diversity, Societal and Environmental Well-being/ Friendliness, Sustainability, Social impact, and Democracy.

Altogether, as supported by Brundage et al., (2020), it can help build a trustworthy methodology to ensure users are able to verify the claims made about the level of privacy protection guaranteed by AI systems, regulators are able to trace the steps leading to a decision/prediction/classification and evaluate them against the context described by the modeled business, academics are able to research the impacts associated with large-scale AI systems, and developers are able to verify best practices are set for each of the AI development stage within the project lifecycle.

In order to achieve Trustworthy AI, (Independent High Level Expert Group on AI, 2019) recommends enabling inclusion and diversity throughout the entire AI system's development project's life cycle involving all affected stakeholders throughout the process. Along with Abolfazlian (2020), both studies describe three components trustworthy AI should comply with throughout the system's entire life cycle: it should be lawful, complying with all applicable laws and regulations; it should be ethical, ensuring adherence to ethical principles and values; and it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Similarly, (Abhishek et al., 2020) proposes three other main components trustworthy AI systems should consist of the following: Ethics of algorithms (Respect for Human Autonomy, Prevention of Harm, Fairness, Explicability), Ethics of data (Human-Centered, Individual Data Control, Transparency, Accountability, Equality), and Ethics of Practice (Responsibility, Liability, Codes and Regulations).

This actually represents an attempt to harness unintended discrimination produced by AIS, from the perspective of the policymaking and legal norms, specifically with basis on the International Law of Human Rights. Given that engineering methods alone couldn't be sufficient enough to protect, according to Fjeld et al. (2020), the fundamental rights from unintended harms of AI systems. As seen above, the Principled AI International Framework presented by Fjeld et al. (2020) gathers a global effort to establish a set of policies and guidelines informed by principles as a methodological reference when designing AI. Despite the progress that this mechanism might represent from the legal point of view, it is yet insufficient as a methodological mechanism manageable by AI designers given their background, and the language (Varona, 2020a &

2020b) discrepancies among legal jargon and the software profession, better detailed in chapters two and three.

## 4.7 Conclusions of the Chapter

This chapter shows the lack of agreement among the scientific community in reaching a standardization of the studied variables to support trustworthy AI as a business model to be assimilated by software developers, specially by AIS designers, when designing AIS. That could be other of the reasons, along with the ones flagged already for the Principled AI International Framework principle's ambiguities described in chapters two and three.

Discrimination and bias are two entangled variables with a strong interdependency that results in one of them being the cause and the effect of the other. For the purposes of the present study, bias refers to the action of deciding upon an individual or group with a given potentially harmful impact because of their features, while discrimination is expressed by the outcome of the decision itself. That reasoning constitutes the mechanics for the logic behind the implementation of the proposed model, described in chapter five.

Discrimination, and by extension Fairness are culture dependable variables. In that regard, there must be required a dedicated assessment for every new project during the conceptualizing stage regardless of the scenario and how the variables will behave across cultures in which the projected ADM system will be deployed in. That consideration is incorporated as a specific practice in the proposed model, described in chapter five.

The study shows that ADMS's biased and discriminatory outcomes are not only a consequence of faulty algorithms and models but are also linked to other processes like data gathering, data cleaning and data processing; and also conditioned by the development team's own bias. Therefore, we have decided to design the proposed model's features and derived variables (mentioned here below) with three dimensions: (1) Algorithm, (2) Data, and (3) Practice, delimiting the scope of the indicated specific practices, detailed in chapter five.



The chapter also identifies the main variables that principled AI is suggesting trustworthy AI should be built upon (through fairness and non-discrimination) in order to design a capability and maturity model for trustworthy AI that takes those variables into consideration. Consequently, the proposed model -described in chapter five-, is orchestrated through fairness and non-discrimination oriented specific goals, and based on the following four derived features: (1) transparency, that involves specific related variables like explainability and accountability; (2) security, that involves specific related variables like safety and privacy; (3) project governance, that involves specific related variables like environmental commitment, societal wellbeing, diversity and inclusion, sustainability, social impact, and compliance with law and regulatory norms; and (4) bias management, that involves specific related variables like knowledge transfer, training, and data collection. These features and their derived variables are taken as checkpoints when determining the level of capability and maturity the AIS designers achieve in their development process when building a trustworthy product or system.

## 4.8 Works Cited in the Chapter

- Kumar, A., Braud, T., Tarkoma S., & Hui P. (2020). Trustworthy AI in the Age of Pervasive Computing and Big Data. *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 1-6, doi: 10.1109/PerComWorkshops48775.2020.9156127.
- Abolfazlian, K. (2020). Trustworthy AI needs unbiased dictators! artificial intelligence applications and innovations. *IFIP Advances in Information and Communication Technology*, vol 584.
- Abrieu, R., Aneja, U., Chetty, K., Rapetti, M., & Uhlig, A. (2018). The future of work and education for the digital age: technological innovation and the future of work: a view from the South. *Argentina : G20*.
- Access Now Organization. (2018). *Human rights in the age of AI*. AccessNowOrg.
- Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G.K., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P.W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensbold, J., O'Keefe, C., Koren, M., . . . Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *ArXiv*. doi:abs/2004.07213
- Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. *Proceedings of the Conference on Fairness, Accountability and Transparency*, 339–348.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, Vol. 5, 153-163.
- Demiaux, V., & Si A.Y. (2017). *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence*. The French Data Protection Authority (CNIL).
- Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C.E., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259-268.
- Fish, B., Kun, J. & Lelkes, Á. D. (2016). A confidence-based approach for balancing fairness and accuracy. *Proceedings of the 2016 SIAM International Conference on Data Mining*, 144-152.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society.
- Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. Cornell University.
- Holstein, K., Vaughan, J.W., Daumé, H., Dudík, M., & Wallach, H.M. (2019). Improving fairness in machine learning systems: What do industry practitioners

- need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.
- House of Lords Select Committee on Artificial Intelligence. (2018). *AI in the UK: Ready, willing and able?*. Authority of the House of Lords.
- IBM. (2019). *Everyday ethics for AI*. IBM.
- Independent High Level Expert Group on AI. (2019). *AI ethics guidelines for trustworthy AI*. European Commission.
- Jago, A.S., & Laurin, K. (2021). Assumptions about algorithms' capacity for discrimination. *Personality and Social Psychology Bulletin*, 1-14.  
doi:10.1177/01461672211016187
- Lazar Reich, C. & Vijaykumar, S. (2020). A possibility in algorithmic fairness: Calibrated scores for fair classifications. *arXivLabs*.
- Loi, M., & Christen, M. (2019). Insurance discrimination and fairness in machine learning: An ethical analysis. *SSRN Electronic Journal*.
- Loi, M., & Christen, M. (2021). Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosophy & Technology*, <https://doi.org/10.1007/s13347-021-00444-9> , Available at SSRN: <https://ssrn.com/abstract=3438823> or <http://dx.doi.org/10.2139/ssrn.3438823>
- Martínez-Plumed, F., Ferri, C., & Nieve, D. (2019). Fairness and missing values. *arXiv*. arXiv:1905.12728
- Martinho-Truswell, E., Miller, H., Nti Asare, I., Petheram, A., Stirling, R., Gomez Mont, C., & Martinez, C. (2018). *Towards an AI strategy in Mexico: Harnessing the AI revolution*. The British Embassy in Mexico.
- Mehrabi, N., Morstatter, F., Saxena, N.A., Lerman, K., & Galstyan, A.G. (2019). A survey on bias and fairness in machine learning. *Machine Learning*.
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*. 2019, Vol. 1, 501–507.
- National Science and Technology Council, Committee on Technology. (2016). *Preparing for the future of artificial intelligence*. Executive Office of the President.
- NITI Aayog. (2018). *National strategy for artificial intelligence*. NITI Aayog.
- Pedreschi, D., Ruggieri, S., & Franco, T.. 2007. *Discrimination-aware data mining Technical Report: TR-07-19*. Dipartimento di Informatica, Università di Pisa.
- Pressman, R. (2010). *Software engineering. A practitioner's approach (7th edition)*. McGrawHill Higher Education, 930.
- Schmidt, N., Siskin, B., & Mansur, S. (2019). How data scientists help regulators and banks ensure fairness when implementing machine learning and artificial intelligence models. *Conference on Fairness, Accountability, and Transparency*.
- Smart Dubai Office. (2019). *AI ethics, principles and guidelines*. Smart Dubai Office.

- Smith, C.J. (2020). Designing trustworthy AI: A human-machine teaming framework to guide development. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.
- Solon, B., & Selbst, A.D., (2016). Big data's disparate impact. *CALIF. L. REV*, 104, 671-732.
- Sommerville, I. (2015). *Software engineering*. Pearson Education.
- Task Force 7. (2020). *The future of work and education for the digital age*. T20.
- Telefónica. (2018). *AI principles of Telefónica*. Telefónica.
- The Public Voice Coalition. (2018). *Universal guidelines for AI*. The Public Voice Coalition.
- The Standish Group. (2020). *CHAOS Report 2020*. The Standish Group.
- UNI Global Union. (2017). *The future world of work. Top 10 principles for ethical artificial intelligence*. UNI Global Union.
- Wickramasinghe, C.S., Marino, D.L., Grandio, J., & Manic, M. (2020). Trustworthy AI development guidelines for human system interaction. *Proceedings of the 13th International Conference on Human System Interaction*.
- Wing, J.M., (2020). Trustworthy AI. *arXiv*. arXiv:2002.06276
- Varona, D. (2018). La responsabilidad ética del diseñador de sistemas en inteligencia artificial. *Revista de occidente*, 446-447, 104-114.
- Varona, D. (2020a). AI systems are not racists just because. *T-13 hours: Building Community Online in CSDH/SCHN2020*.
- Varona, D. (2020b). Artificial intelligence design guiding principles: Review of “European ethical charter on the use of AI in judicial systems and their environment”. Retrieved from <https://www.danielvarona.ca/2020/06/17/artificial-intelligence-design-guiding-principles-review-of-european-ethical-charter-on-the-use-of-ai-in-judicial-systems-and-their-environment/>. March 2021.
- Varona, D. (2020c). Artificial intelligence design guiding principles: Review of “Recommendation of the council on artificial intelligence. Retrieved from <https://www.danielvarona.ca/2020/06/28/artificial-intelligence-design-guiding-principles-review-of-recommendation-of-the-council-on-artificial-intelligence/>. March 2021.
- Vatican. (2020). *Rome call for AI Ethics*. Vatican.
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. *ACM/IEEE Proceedings of the International Workshop on Software Fairness (FairWare 2018)*.

- Yukun, Z., & Longsheng, Z. (2019). Fairness assessment for artificial intelligence in financial industry. *arXiv*. arXiv:1912.07211
- Zafar, M.B., Valera, I., Rodriguez, M.G., & Gummadi, K.P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv*. arXiv:1507.05259
- Zhang, L., Wu, Y., & Wu, X. (2017). A causal framework for discovering and removing direct and indirect discrimination. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 3929-3935.

## Chapter 5

### Proposal of a Capability and Maturity Model for Trustworthy ADM Systems

This chapter presents the result of the systematization of the study of the Principled AI International Framework —based on the international law for human rights supported policies and regulations—, the study of trustworthiness related variables, and available ML mechanisms to address the issue of fairness and discrimination produced by artificial decision-making systems. This chapter integrates the insights of studies described in the previous chapters into the design of a capability and maturity model for trustworthy ADM systems. The proposed model pursues five specific goals, accomplished through 19 specific objectives, expected to be achieved by the adoption of 74 specific practices oriented to three dimensions: data, algorithm, and engineering practice. The model is aligned with other quality assurance and development process practices in software development as proposed by Pressman, Sommerville, and the Software Engineering Institute.

#### 5.1 Introduction

The evolution of ML techniques, along with the data logs produced by many software aided processes in our society, among other factors, have conditioned the increasing demand for AI systems to support important business decisions. The discriminatory character of some of those AIS produced decisions in key domains like public security (Balaji et al., 2021), hiring (Garg et al., 2021), health care (Qayyum et al., 2020), etc. have been identified. Consequently, there is an increasing interest among researchers and scholars for identifying the causes of these discriminatory decisions, and eventually correct them.

The AIS discriminatory outcomes have been primarily attributed to the functioning of the algorithm (Fu et al., 2020; Sun, Nasraoui & Shafto, 2020), especially in solutions involving ML techniques and AI procedures. Recently it was recognized that data collection procedures were a source of discrimination and bias (Engstrom et al., 2020). The recognition that other elements beyond algorithms could be responsible for

discrimination and the ensuing flawed AIS resulted in researchers exploring further possible sources of discrimination and bias, resulting in the identification of the human factor (Cowgill et al., 2020) as an important source.

At the same time, related theories have evolved from conceptualizing bias (Mehrabi, et al., 2019; Yukun & Longsheng, 2019) and discrimination (Martínez-Plumed, Ferri, & Nieve, 2019; Verma & Rubin, 2018) to define fairness (Hughes et al., 2019), and lastly trustworthiness (Wickramasinghe et al., 2020). However, the literature suggests that it is difficult, from the software engineering discipline, to address algorithmic fairness, or AIS neutrality as it is also referred chapter one. International institutions like IEEE<sup>29</sup> and ISO/IEC,<sup>30</sup> historically leading the software development sector and establishing guidelines for software engineering process standardization, are still drafting standards targeting discrimination produced or amplified by AIS.

Alternatively, some researchers have been targeting the issue from a regulatory and public policy perspective (Rodrigues, 2020; Majumdar & Chattopadhyay). The researchers advocate for international human rights legislation to mitigate the negative impact of AIS from its design and acquisition stages. These researchers believe that international human rights law incorporates the necessary provisions to adequately manage the impact that might result as a consequence of a discriminatory decision produced by an AIS, while at the same time sets the course of action for legal remedies when needed. Feldj (Fjeld et al., 2020) mapped what can be considered as a Trustworthy AI International Framework, which has been further explored and criticized in chapters two and three.

The Trustworthy AI International Framework mapped by Feldj distinguishes itself over other similar efforts (Ryan & Carsten Stahl, 2021) in the accessibility to the documents included in the framework, and the author's analysis of their regional and

---

<sup>29</sup> Institute of Electrical and Electronics Engineers professional association.

<sup>30</sup> International Standardization Organization ISO/ International Electronic Commission IEC standards organizations.

international impact. An analysis of the framework identified several elements hindering its adoption from a software engineering perspective and revealed several issues with fairness and trustworthiness, which justified a further exploration of these variables, as described in chapter four.

This chapter aims to systematize previous findings into the design of a capability and maturity model based on trustworthy AI features resulting of the study of the referred Trustworthy AI International Framework and from the survey of the available specialized literature.

## 5.2 Related Research

### 5.2.1 Heavyweight Software Development Models

Heavyweight software development models or methodologies embody a set of procedures, techniques, and archive support guidelines for software development, framed in a step-by-step detailed process, where each and all tasks planned towards the desired software product are described.

According to Pressman (2010) and Sommerville (2015) heavyweight software development models include (1) the waterfall model, (2) the V-shaped model; and a set of models that can be grouped in categories like (3) incremental process models, (4) evolutionary process models, and (5) concurrent models. These heavyweight model categories are more tuned with the way processes are outlined than with their functioning. They serve as a basis for other groups of specialized models that, according to the chosen engineering approach, can be denominated as: (1) component-based development, when the development is divided into smaller deliverables or limited to a specific component; (2) formal models, to refer to mathematical formalization of a problem and its solution; and (3) aspect-oriented development models, when focusing on software characteristics. The philosophy behind the aspect-oriented development models provides a practical view about the way a capability and maturity model based on quality features can be executed, not only from an engineering standpoint but also of how it can be implemented as part of



the software solution if the features, turned into variables, are defined and modeled accordingly.

Jacobson combines the referred models into a single unified development process (RUP) (Jacobson, Booch, & Rumbaugh, 1999) which is the primary software development methodology utilized for medium and high complexity projects. Additionally, RUP is the primary software development methodology that quality assurance organizations consider as a reference for designing their evaluation and auditing mechanisms. Consequently, RUP is also taken as reference in the design of the proposed model described in the present study.

### 5.2.2 Agile Software Development Models

Opposite to traditional methodologies, agile models focus on delivering small pieces of software, while involving the client, from the start, in the creation process. This has advantages considering the software development industry's fast-paced and ever-changing related technologies and paradigms.

Some of the most popular agile software development methodologies are (1) extreme programming (XP), focused on coding activities with the model having an individual and an industrial version; (2) SCRUM, that distinguishes from XP in orchestrating the team's efforts in a common goal at a time; (3) agile modelling (AM), oriented to incremental prototypes; and (4) a RUP agile version called agile unified process (AUP), possibly the most used agile methodology after SCRUM.

The agile software development methods are more varied than traditional ones, as the software industry has become less centralized over time. Another factor favoring the shift towards agile development was identified in CHAOS biannual report (The Standish Group, 2020). It noted that project duration was a main failure element, and stressed that technological changes in short periods of time demand equally short development periods.

As the software development field is composed of both agile and the traditional approaches the proposed model is designed having both types of development

philosophies into account, so it can be applied regardless the software development method used.

### 5.2.3 Feature-Oriented Software Development Models

Regardless of the engineering style followed, traditional or agile approach, there is a group of development models that are oriented to specific features when building software. These models are chosen according the client's requirements, the context in which the software will be deployed, and the technological environment in which the software will be introduced, etc.

Some examples of feature-oriented development model are as follows: security (Peldszus et al., 2018), services (Rodriguez-Martinez et al., 2021), architecture (Santos et al., 2021), energy efficiency for internet of things (IoT) (Kumar et al., 2020). Quality-oriented development models are as follows: test driven development model (TDD) (Al-Saqqa et al., 2020), and the Capability Maturity Model Integrated (CMMi) (Software Engineering Institute SEI, 2010). While these examples represent models that are product or process oriented, there are other models that are team oriented, such as Jamie et al., (2020), which is mainly focused on inclusiveness and its impact on team members' well-being and productivity, as well as the offshore software development outsourcing (OSDO) model (Muhammad et al., 2020) which focuses on the management of geographically distributed teams. Lastly, aligned with recent concerns for trustworthiness, Knowles and Richards (2021) proposed a theoretical framework for Trusted AI research.

Knowles and Richards's model (2021) seeks to influence public's trust on AIS. It shares theories of trust with another related model proposed by Toreini et al. (2020). Both models focus on the public's trust in AI rather than in supporting the creation of AIS worthy of trust. They argue that trust in a particular technology differentiates from trust in people in a context where individuals are impacted by decisions made by AIS without their knowledge or consent with no other option but to accept the decision made by the AIS. In this context the level of trust in the technology will increase with well documented processes (useless to the average person with little or no tech literacy) or

decreased with an underdeveloped dedicated regulatory ecosystem and underdefined features for trustworthiness like the ones in the current AIS development landscape. Either way the general public is dependent on the institutions using AI and their ethical considerations.

Knowles and Richards's model (2021) aims to increase trust in AI by making it explainable. The model focuses on documenting an AI's artifacts across project instances to provide potential interested parties with data on the functioning of AI and its outcomes. The model outlines a process in which the resulting documentation supports the creation of regulations that are introduced back into the development process so it could be verifiable and auditable. However, the model ignores several other variables within the available definitions of trustworthy AI, and it fails to influence the AIS development process, something critical in the pursue of AIS worthy of trust from its design.

Toreini's model aims at influencing trust in AIS through an adapted version of the ABI model, supported by theories in the literature (Mayer et al., 1995; Sanders et al., 2006), fostering variables like ability, benevolence, integrity, and predictability through four dimensions consisting of humane, environmental, technical qualities, and the contrast between the initial and current level of trust in the target AIS in a given period of time. Although this model focuses on the general public's trust<sup>31</sup> towards AI it acknowledges that AIS's features of fairness, explainability, auditability, and provisions for safety are key elements to increase AIS's trustworthiness. Lastly, the model points at the ethical and principled driven implementation of the previously mentioned features as the link between the AIS technical disposition and the public's perception of benevolence.

Consequently, it is appropriate to infer that there are some features, like the ones highlighted by Knowles and Richard, and Toreini's models that can be planned,

---

<sup>31</sup> As the model aims at identifying the elements that influence public's trust on AIS so it can be modeled and conditioned.

implemented, verified, and measured as part of the AIS development process with the purpose of conditioning trust on AIS from its design stages. Hence, the authors of the present study believe that similar to how the CMMi model, specifically in its constellation oriented to the development process, ensures the overall quality of the developed solution by dissecting the project lifecycle in several process areas demarking capabilities and maturity levels. Then the referred features can be treated as AIS's quality characteristics also used to establish capabilities and maturity levels to show a standardized measure of trustworthiness for AIS. A standardized measure of trustworthiness for AIS can assist developers when selecting third parties' components without affecting the current quality of their solutions, institutions when selecting ethics and principled AIS, and the general public to have a scale to gauge trustworthy AI.

In order to understand the particularities of the development cycle of algorithmic decision-making systems, the following section is presented. Also, a study of the available algorithmic decision-making systems' specialized development models is provided.

#### 5.2.4 Algorithmic Decision-Making (ADM) Systems' Development Models

Like any other software solution, AIS and ML systems need to be analyzed, designed, implemented, verified, and maintained. However, there is a lack of specialized engineering practices within the software development industry for such systems as they are fundamentally different from traditional software systems. Building AIS and ML solutions requires extensive trial and error exploration for model selection, data cleaning, feature selection, and parameter tuning. Moreover, there is a lack of theoretical understanding that could be used to abstract away these subtleties. Conventional software engineering paradigms have not been designed to address challenges faced by AI and ML practitioners. This section gathers some of the available studies focused in exploring the engineering particularities of AIS and ML systems, with the objective of identifying shared elements with the available policies from the Principled AI International Framework.

A Microsoft based study conducted by Amershi and others (Saleema et al., 2019) where they observed teams as they develop AI-based applications concluded that engineers tend to adjust to a given nine-stage workflow to conventional software development methodologies (mostly agile). They also figured a set of best practices from Microsoft teams and discussed three fundamental differences in how software engineering applies to ML-centric components in contrast with previous application domains.

The referred nine-stages workflow included the following: (1) Model Requirements; (2) Data Collection; (3) Data Cleaning; (4) Data Labeling; (5) Feature Engineering; (6) Model Training, which may loop back to stage five; (7) Model Evaluation, which may loop back to any previous stage; (8) Model Deployment; and (9) Model Monitoring, also able to loop back to any of the precedent stages. These stages are similar to Shearer's (2000) six phases for data mining projects, which includes (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling, (5) evaluation, and (6) deployment. Amershi's workflow adapts to conventional software methodologies integrating the stage one to the Requirements Modeling phase, stages two to five with the Analysis and Design phases, stage six to Implementation, stage seven to Quality Assurance, stage eight to Deployment phase, and stage nine to the phases of Maintenance and Support. Although these suggested pairings might change in each methodology, it clearly implies more workload on the side of analysis, design, and quality assurance, than on implementation. That could be one of the reasons why the available specialized literature describes, on one hand, software developers aiming to address the issue of fairness calibrating and balancing algorithms and models once they are produced, and on the other hand, policy makers and researchers seeking to influence the solution from analysis and design stages. Neither of them focuses on the implementation stage.

Among the best practices highlighted by Amershi's study are: End-to-end pipeline support, a sort of controlled data environment where engineers can tune their models; Data availability, collection, cleaning, and management, as a philosophy of "internal" data openness where evolution of datasets resulting from the introduction of fresh data

and/or as result of several iterations of model tuning is acknowledged, and the consequent data configuration management is needed; Model Debugging and Interpretability, not only focused in the model tuning but also documenting those conditions in which the model fails; and Compliance, alluding to Microsoft's approach to AI principles. These best practices are part of an organization culture we would like to integrate into our model.

Amershi's study differentiates the software engineering applied to AIS and ML solutions from other application domains with three elements. The first, the complexity of discovering, managing, and versioning the data needed for ML applications, which is higher than in other software engineering approaches. The second, the skills required for model customization and model reuse are quite different than skills typically found in software teams. And last, the system modularity, which is more difficult to handle in AIS and ML solutions as distinct modules may be "entangled" in complex ways and experience non-monotonic error behavior in contrast to traditional software components or models. As can be noticed, all three differences lay on methods. Therefore, we believe our model's capability and maturity levels need to be method centered rather than feature oriented.

Similar to Amershi's findings, Ozkaya, on an editorial letter for IEEE Software (Ipek, 2020) explains that an AI-Enable Systems' software engineering approach, although more complex than other domains, is not necessarily different from those. The letter pinpoints the addition of a data scientist to the software engineering team as a critical stakeholder while acknowledges that data science processes do not always align with software engineering procedures when they follow a rigid setup. This is the reason why the organization's culture plays an important role in avoiding project failure.

Both studies, along with others (Kaestner, 2020; Rahimi et al., 2019; Subbaswamy et al., 2019; Frunal et al., 2019; Vrutik, & Gopalan, 2019) agree when presenting ML as an engineering requirement, that is mainly verification oriented. ML engineering approach does this, first, by focusing on an automated mining directed architecture and variant detection rather than in providing implementation guidelines, and

shortly after, by shifting the focus to verifying the product builds against the settled specifications, as evidenced in multiple general-scope researches such as (Ricchio et al., 2020; Xiaowei et al., 2020; Zhang et al., 2020), along with data testing focused studies like (Guy et al., 2019; Re et al., 2019) and data quality management-oriented studies as (Neoklis et al., 2019; Schelter et al., 2018), debugging focused research like (Yeounoh et al., 2019), and validation of ML frameworks as exhibited in (Siwakorn et al., 2018), for a few examples.

In that regard, an engineering model taking available AI principles when guiding the engineering process through the entire lifecycle, will complement the existing method, described before, resulting in an improved development process. Accordingly, it is pertinent to assume that quality assurance activities, conducted on every stage and phase of the engineering process, are an adequate way of influencing the overall product and process's trustworthiness when implementing available AI principles.

In this respect, the Algorithmic Decision Making (ADM) methodology proposed by Aysolmaz (Aysolmaz, Dau, & Iren, 2020), which integrates four relevant process frameworks CRISP-DM (Shearer, 2000), ASUM from IBM (IBM, 2016), DMLC (Hofmann & Tierney, 2009), and TDSP (Ericson et al., 2017), serves as a reference in the design of the model proposed in this study.

### 5.2.5 Quality Assurance of Algorithmic Decision-Making (ADM) Systems

There exist three marked trends, across the studied literature, for software quality assurance in the past two decades. A significant, and more traditional group of enterprises still follow the Six Sigma standardized methodology design for the Motorola company in 1986. The International Standardization Organization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) share the vast majority of modern available quality assurance standards nowadays, representing a second trend. The remaining trend is justified by the organizations who are certified or seek certification with the CMMI quality model. All three instruments are characterized in the following subsections.

### 5.2.5.1 Six Sigma Quality Assurance Methodology

The Six Sigma is a methodology oriented towards the continuous quality improvement of processes by identifying and removing failure causes in both products, and manufacturing and business processes, using statistical methods. It has two five-levels constellations, one used in projects focused on improving currently existing business and manufacturing processes, and the other used in projects focused in designing new processes and products. Basically, the Six Sigma methodology contrasts different variations of a same process and discards the ones with higher error rates. While the original Six Sigma methodology aims at reducing process variations, a lean version of it aims at reducing waste. However, both approaches are commonly merged, given their statistical core, and common focus on process improvement.

The use of the Six Sigma methodology is discouraged by organizations of 500 employees or less, as their processes variations are more obvious. Other usual criticism against the Six Sigma methodology highlights that, while focusing on continuously improving processes over the base of their own, it lacerates the opportunity for innovation and evolution what ends up limiting the organization adaptability. In contrast, Six Sigma advocates defend the business efficiency resulting of highly tuned processes. A balance is found, among the reviewed studies, in the existing belief that Six Sigma's statistical based methods can be executed during the measurement stage of CMM and CMMI quality assurance models.

### 5.2.5.2 ISO/IEC and IEEE International Standards

The International Standardization Organization (ISO) and the International Electrotechnical Commission (IEC) work together in providing a set of norms oriented to standardize manufacturing processes and product specifications. There are a number of standards dedicated to the software industry, specifically to quality assurance. In addition to the ISO and IEC's standardization efforts others can also be found done by the Institute of Electrical and Electronics Engineers (IEEE).



### 5.2.5.2.1 ISO/IEC Software Quality Assurance International Standards

The standard from which current ISO/IEC software quality assurance standards are derived from is the withdrawn standard ISO/IEC 9126 - Software Engineering - Product Quality published in 1991 (ISO & IEC, 1991). Interestingly, the ISO/IEC 9126 fundamental objective was to address human biases that can adversely affect the delivery and perception of a software development project. ISO/IEC 9126 definition of bias were expressed in terms of changing priorities and the unclear definition of notion of “success” after the start of a project. The standard aims at ensuring building software products with six measurable characteristics: Functionality, when the product satisfies stated or implied needs; Reliability, the product performs as expected, timely, and under specified conditions; Usability, the product is easy to use by prospective users; Efficiency, the relation performance-cost of resources, under stated conditions; Maintainability, the product is effortlessly modified upon organization, technological or business needs; and Portability, the ability of the software product to be transferred from one environment to another. All six are further divided into a total of 27 equally product-oriented sub-characteristics. The standard’s product-oriented scope justifies that is mainly applied by the use of internal, external, and in quality-in-use measurements.

The original standard ISO/IEC 9126 - Engineering - Product Quality was replaced in 2001 by four improved versions of itself (ISO & IEC, 2001, 2003a, 2003b & 2004). Each part of the new ISO/IEC 9126 expanded on one specific measurement type and their integration into the project lifecycle. The set of characteristics in ISO/IEC 9126 were reviewed and ISO/IEC 25010 added two new product-oriented quality characteristics to the original set: Compatibility, the ability of a software product to co-exist and interoperate with other software products; and Security, when software product properly verifies user’s identity, manages data access and data modification, is able to register action logs, and trace action to actioner. Consequently, the number of total sub-characteristics increased to 31, yet all product oriented, as can be verified in (ISO & IEC, 2010).

The ISO/IEC 25010 (ISO & IEC, 2010) is part of a set of standards that conform the SQuaRE Model, which is an abstract representation of a quality model, expressed in six divisions, each consisting of several standards: Quality Management [ISO/IEC 2500n<sup>32</sup>], defines all common models, terms and definitions further referred to by all other International Standards from the SQuaRE series, also provides requirements and guidance for a supporting function that is responsible for the management of the requirements, specification and evaluation of software product quality; Quality Model [ISO/IEC 2501n], provides detailed quality models and guidance on their implementation when evaluating computer systems and software products, quality in use, and data; Quality Measurement [ISO/IEC 2502n], offers a quality measurement reference model, quality measures, mathematical definitions, and practical guidance for their application in software projects; Quality Requirements [ISO/IEC 2503n], suggest guidelines for quality requirements specifications and their respective metrics, based on quality models and quality measures, to be used in the process of quality requirements elicitation; Quality Evaluation [ISO/IEC 2504n], delivers requirements, recommendations and guidelines for software product evaluation, whether performed by evaluators, acquirers or developers; and SQuaRE Extension [ISO/IEC 25050 – ISO/IEC 25099], includes requirements for quality of Commercial Off-The-Shelf software and Common Industry Formats for usability reports.

ISO/IEC standards are revised every five years, and as a result of this evaluation they may be improved or replaced. At the moment of the present investigation the standard ISO/IEC 25010:2011 is being reviewed and it is known it will be replaced by the standards ISO/IEC 25002<sup>33</sup> Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality models overview and usage,

---

<sup>32</sup> Where n is a number of a standard of the family of standards expressed by the precedent four digits. Different family of standards has a defined scope e.a. standards where n is “0” are glossary of terms for the rest of the family.

<sup>33</sup> ISO/IEC 25002 provides an overview of the SQuaRE quality model related norms and further guidance for its use

ISO/IEC 25010<sup>34</sup> Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Product quality model, and ISO/IEC 25019.2<sup>35</sup> Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality in use model, which are being drafted. Until it is made public, it is fair to believe they will follow the precedent ISO/IEC standards scope on measurable product-oriented quality characteristics and sub-characteristics.

### 5.2.5.2.2 IEEE Software Quality Assurance International Standards

Among IEEE's catalogue of standards there are a total of 43 dedicated to software and system engineering. Many of them incorporate ISO/IEC nomenclature so it is not unusual to find IEEE/ISO/IEC type standards, or IEEE standards whose title reference is the ISO/IEC counterpart. Both IEEE and ISO/IEC complement each other, while ISO/IEC standards are mainly oriented to the product and the product specifications, IEEE catalogue focuses on the process documentation, as a mechanism of influencing and standardizing the manufacturing process. As a result, they manage a smooth entanglement when IEEE adopts ISO/IEC product specifications on their own process definitions.

Similar to ISO/IEC standards, IEEE standards also experience regular revisions. Related to their area of application, software developers find the following active IEEE standards particularly useful.

First, the standard IEEE/ISO/IEC 29148-2018 - ISO/IEC/IEEE International Standard - Systems and Software Engineering - Life Cycle Processes - Requirements Engineering (WG\_LCP - Working Group for Life Cycle Processes, 2018) supporting software engineering requirements throughout the project's life cycle by defining

---

<sup>34</sup> ISO/IEC 25010 provides a product centered quality model overview of the SQuaRE model for system and software development.

<sup>35</sup> ISO/IEC 25019.2 provides a use centered quality model overview of the SQuaRE model for system and software development.

constructs of good requirements, suggesting attributes and characteristics for requirements, and discussing the iterative and recursive application of requirements processes throughout the life cycle. This standard expands on and uses as reference ISO/IEC/IEEE 12207 and ISO/IEC/IEEE 15288 but include no revision of the working quality variables from their ethical and social non-functional dimensions. Therefore, the standard's scope still ignores emerging business engineering challenges related to bias and fairness.

Second, the standard IEEE/ISO/IEC 21839-2019 - ISO/IEC/IEEE International Standard - Systems and Software Engineering - System of Systems (SoS) considerations in life cycle stages of a system (WG\_LCP - Working Group for Life Cycle Processes, 2019) provides a set of considerations to be addressed at key points in the life cycle of systems created by humans that will interact in a system of systems as the system of interest (SoI). This is a particular niche of software products that is constantly evolving and integrating more AI and ML components every day in domains such as transportation: air traffic management, the European rail network, and cargo transport; health care: emergency response service, and personal health management: and defense: missile and shield, and networked sensors; for example. Similar to IEEE/ISO/IEC 29148-2018 this standard aligns with ISO/IEC/IEEE 15288 and ISO/IEC/IEEE 24748 framework for system life cycle stages and associated terminology. Hence, it exhibits similar scope limitations.

Third, the standard IEEE/ISO/IEC 15026-1\_Revision-2019 - ISO/IEC/IEEE International Standard - Systems and Software Engineering - Systems and Software Assurance - Part 1: Concepts and Vocabulary (WG\_LCP - Working Group for Life Cycle Processes, 2019), defines a relational map of assurance-related concepts, thereby establishing a basis for a shared understanding of the terminology and principles central to all parts of ISO/IEC/IEEE 15026 across its user communities. It also expands on providing information regarding successive parts of ISO/IEC/IEEE 15026, their expected use as a single standard, and their expected used when being combined. Again, guiding the documentation process of assurance related tasks while using pre-established conceptions, although it still ignores the issue of bias, fairness, and trustworthiness.

The referred IEEE/ISO/IEC P15288 standard has been recently revised (WG\_LCP - Working Group for Life Cycle Processes, 2021). So, it is accurate to believe the two previous standards might soon experience some sort of revisions themselves. The available information states that the revised version establishes a common framework of process descriptions for describing the life cycle of systems created by humans, defining a set of processes and associated terminology from an engineering viewpoint. This would be applied at any level in the hierarchy of a system's structure, while involving stakeholders, with the ultimate goal of achieving customer satisfaction. It also states that the revised version provides processes which support the definition, control and improvement of the system life cycle processes used within an organization or a project. This is particularly important for organizations and project managers who can use these processes when acquiring and supplying systems to other organizations and project teams. Still, there is no mention of whether the new standard contemplates trustworthiness related variables.

Last, the standard IEEE 2755.1-2019 - IEEE Guide for Taxonomy for Intelligent Process Automation Product Features and Functionality (IPA - Intelligent Process Automation , 2019) provides a common understanding among individuals involved with Software-Based Intelligent Process Automation products so that industry participants may rely on the manufacturer's functionality claims about a product, understand the underlying technological methods used to produce its functionalities, and how one might approach evaluating the relative sophistication and importance of each function or feature. It uses terminology as established in IEEE Std 2755-2017, defining, and classifying approximately 150 features and functions across five core areas of technology capability in the family of new technology products collectively referred to as Intelligent Process Automation. The standard might be perceived as one initial step toward explainability and understanding at general levels of the functioning of processes executed when designing the solution leaving trust-related topics out of scope.

Fortunately, it is now known that IEEE working groups are drafting several standards pursuing to influence the design of ethical software solutions, especially,

systems with algorithmic decision-making components. The standardization projects include the following:

- (1) The IEEE P7000 - IEEE Draft Model Process for Addressing Ethical Concerns During System Design (EMELC - WG - Engineering Methodologies for Ethical Life-Cycle Concerns Working Group, 2021) establishes a group of processes by which organizations can include consideration of human ethical values during the stages of concept exploration and development, support management and engineering transparent communication with stakeholders for values elicitation and prioritization, involves traceability of ethical values through operational concepts, value propositions, and value dispositions in the system design, and integrates traceability of ethical values in the concept of operations, ethical requirements, and ethical risk-based design. It is claimed the standard will be applicable for all sizes and types of organizations using their own life cycle models.
- (2) The IEEE P7001 - IEEE Draft Standard for Transparency of Autonomous Systems (ASV WG\_P7001 - Autonomous Systems Validation Working Group\_P7001, 2021) aims at describing measurable, testable levels of transparency, so that autonomous systems can be objectively assessed, and levels of compliance determined.
- (3) The IEEE P7002 - IEEE Draft Standard for Data Privacy Process (PDP - Personal Data Privacy Working Group, 2021) guides the requirements of engineering process by providing definitions for privacy-oriented considerations regarding products, services, and systems utilizing employee, customer, or other external user's personal data, with impact across the life cycle from policy through development, quality assurance, and value realization. It applies to organizations and projects that are developing and deploying products, systems, processes, and applications that involve personal information.

- (4) The P7003 - Algorithmic Bias Considerations (ALGB - WG Algorithmic Bias Working Group, 2021) describes specific methodologies to help users certify how they worked to address and eliminate issues of negative bias in the creation of their algorithms and models, where "negative bias" refers to the usage of overly subjective or uniformed data sets or information known to be inconsistent with legislation concerning certain protected characteristics (such as race, gender, sexuality, etc.), or with instances of bias against groups not necessarily protected explicitly by legislation, but otherwise diminishing stakeholder or user wellbeing and for which there are good reasons to be considered inappropriate. Possible elements include (but are not limited to): benchmarking procedures and criteria for the selection of validation data sets for bias quality control; guidelines on establishing and communicating the application boundaries for which the algorithm has been designed and validated to guard against unintended consequences arising from out-of-bound application of algorithms; suggestions for user expectation management to mitigate bias due to incorrect interpretation of systems outputs by users (e.g. correlation vs. causation).
- (5) The IEEE P7004: Standard for Child and Student Data Governance (EDP - WG Employer Data Governance Working Group, 2021) provides, on the one hand, specific methodologies to help users certify how they approach accessing, collecting, storing, utilizing, sharing, and destroying child and student data; while on the other hand, it provides a system of metrics and conformance criteria regarding these types of uses from trusted global partners and how vendors and educational institutions can meet them.
- (6) The IEEE P7005 - IEEE Draft Standard for Transparent Employer Data Governance (EDG-WG Employer Data Governance Working Group, 2021) is similar to the P7004 but focuses on employees rather than children and students. It defines specific methodologies to help employers in accessing, collecting, storing, utilizing, sharing, and destroying employee data, and provides specific metrics and conformance criteria regarding these types of

uses from trusted global partners and how third parties and employers can meet them. Both, P7004 and P7005 leave out of their scope certification processes, success criteria, and execution procedures.

And last, the P7006 - Standard for Personal Data Artificial Intelligence (AI) Agent (WG-PDAI - Personal Data AI Agent Working Group, 2021) describes the technical elements required to create and grant access to a personalized AI that will comprise inputs, learning, ethics, rules, and values controlled by individuals.

### 5.2.5.3 Software Development and Software Quality Assurance Models

Surveys on software quality assurance models conducted by Suman and Wadhwa, and Miguel and others (Miguel, Mauricio, & Rodriguez, 2014; Wadhwa & Wadhwa, 2014), in 2014, show comparative studies of models from as early as 1977 to 2013, for a total of 28 models. If we exclude from that list the ISO/IEC and IEEE standards groups they treated as models, the SATC's model, the Aspect-Oriented Software Quality model, and the Component Based Software Development Quality Model, which are strictly based on the aforementioned standards, the number decreases to 22. Both studies highlight a set of quality characteristics and to what degree software products are expected to satisfy them, across the remaining 22 models. The definitions of these characteristics are aligned with the terminology standardized by ISO/IEC norms and whose processes of assurance were influenced by the IEEE documenting guidelines.

The period covered by Suman-Wadhwa's, and Miguel's studies is described by a strong trend towards verification and validation of software products by means of measuring their adherence with engineered requirements and the subsequent client's satisfaction, check listing the architecture and design, and debugging and testing the code builds. During this timeframe, the volume of software solutions incorporating AI and ML components has not yet rocketed, nor is it a timeframe characterized by an aggressive pool of big companies competing for the data supremacy. Therefore, it is understandable those models lack attention for possible ethical issues produced or aggravated by the solutions they were assuring.



Recently, Galli and others (Galli, Chiclana & Siewe, 2020) expanded on Miguel's revision of quality assurance models by adding 12 other models to the analysis. This time models were ranked on relevance. And relevance was curiously defined as the degree in which the models complied with a greater number of quality characteristics specified in IEEE/ISO/IEC terminology standards. Not surprisingly the top ranked models were the same ISO/IEC 25010 and 9126. As stated in the previous subsection the currently active ISO/IEC and IEEE standards designed their quality characteristics with focus on the product and the documental process of the development effort.

It is intriguing to note how these referenced studies included a previous version of the Capability Maturity Model CMM, without specifying which constellation (product, development, etc.) they refer to, and excluded its integrated version CMMi, or the latest version CMMi 2.0, both well known among software engineers and software engineering researchers. It could be thought that CMMi is studied with moderation given that it is a branded model, with expensive training, evaluation, and certification systems centralized by the CMMI Institute and the Software Engineering Institute (SEI) Center, based in USA. Those are usually the elements pointed by its critics. Also, there is some skepticism regarding the model's survival to the post-agile development context. However, the CMMi certification is broadly pursued by major software developer enterprises around the world.

The reason why CMMi might be seen incompatible with an agile and post-agile development is that it focuses on processes and their interoperability rather than product quality characteristics. The model follows a proactive approach supporting engineers in an adequate definition, monitoring, and evaluation of processes through several key process areas transversal to most software projects. On the other hand, the training, evaluation, and certification process require more time, effort, and logistics than the adoption of other models. The fact that the model is process-oriented moves it to certify the project team instead of the product, what plays against its popularity, forcing managers to evaluate their team's stability before seeking a CMMi seal.

Section 1.4 of the present study points to differences among software methodologies for traditional software solutions, establishing how the requirement engineering and quality assurance related tasks gain relevance over other development efforts in AI and ML projects. This combines with the fact that available software quality assurance models, defined before 2013, evidently overlook particularities of AI and ML development projects in which the quality assurance tasks are usually concentrated in evaluating quality of data and data completeness, and accuracy of training models. Consequently, they fail to meet current software engineering challenges regarding the increasing concerns among researchers, academics, and policymakers regarding the impact of AIS and ML solutions in terms of negative discrimination, and creation, amplification, and perpetuation of pre-existing biases.

In an intent to address such concerns a group of Japanese researchers (Hamada et al., 2020) proposed a set of guidelines for quality assurance of AIS and ML. The suggested guidelines are driven by a balance of five axes: Customer expectation, Process agility, Data integrity, Model robustness, and System quality, which are tuned through an extensive effort of testing, which the authors agreed is the main practical activity on their proposal. Unfortunately, similar to precedent quality assurance models, the proposed guidelines follow a technical approach and do not consider any of the variables used to conceptualize discrimination, bias, and fairness in the context of AIS and ML. Hence, there is still a need for a mechanism to complement currently available software quality assurance mechanisms, as the aforementioned, in order to include the social aspect in software engineering to favor a more ethical software design.

### 5.3 Method

The present study was performed firstly as exploratory research which later acquired a relational character since the primary goal was to increase the degree of familiarity with variables like bias, discrimination, fairness, and trustworthiness in the context of ML and AI to further determine their interdependence and a way to integrate them into a capability and maturity model for software development based on those variables as quality features. A causative non-experimental research design was conducted aiming to determine the relationship between the elements within the

mentioned variables. Given the particularities and demands that the implementation of the proposed model would carry, we were unable to conduct empirical experiments. In contrast, we used three study cases to illustrate the application of proposed model, which is presented at the end of this chapter. The proposed model presented in the present chapter was designed using theoretical research methods among which the following are included:

First, the inductive-deductive method when studying the available software development models, and standards to identify relevant elements for different AIS development phases in which integrate the results from the analysis of the regulatory instruments gathered in the Trustworthy AI International Framework. This method also allowed to incorporate revised versions and reuse the structure of available mechanisms with proven effectiveness in the area of software quality assurance, particularly the review of the quality features theories and the adoption of the CMMI for Development's philosophy to the proposed model.

Second, the analytical-synthetic method when reflecting on elements within the domain of the target variables, their relation to each other, and their adaptability with studied models, allowing to draw empiric and theoretical conclusions regarding the feasibility of turning them into quality features incorporated into the model as capability levels further arranged in maturity levels.

Third, the historical-logical method when determining causal conditions for current tendencies of software quality assurance aiming at fairer AIS and their integration into the software development process. Then, we used the modeling method when designing the proposed solution.

Lastly, the systematic method when ensuring that the different pieces resulting from the application of the other methods described above, which are components of the proposed model, are part of a whole that works harmoniously.

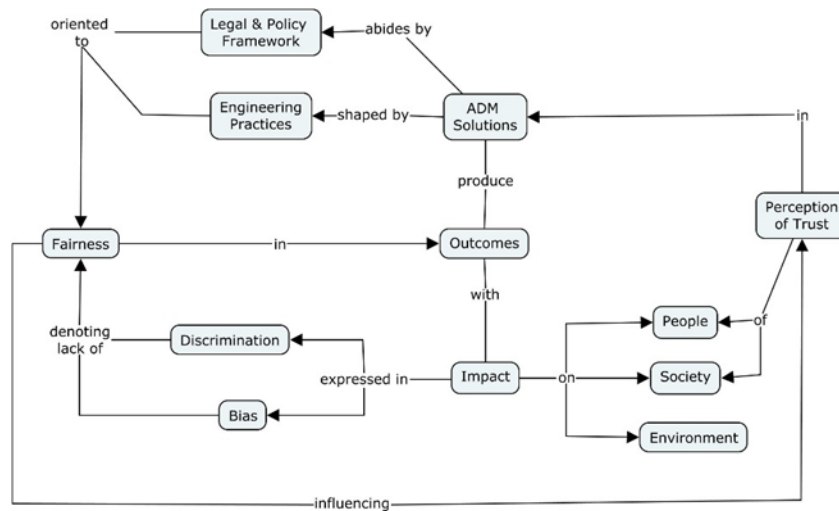
## 5.4 Capability and Maturity Model for Trustworthy Algorithmic Decision-Making (ADM) Systems

The present model proposal is completely aligned with the previous chapters. Chapter two allowed to understand the model environment and to conceptualize the model's architecture. Chapter three helped to define the capability levels, variables through which the goals of each capability level are orchestrated, and justify the multi-dimensional scope of the model, expressed in data, algorithm, and practice-oriented general objectives. This means that the specific objectives are arranged in correspondence with both the dimensions and the maturity level's purpose. Chapter four reinforced the model's three dimensions as ethics of data, ethics of algorithm, and ethics of practice; and supported the definition of the specific objectives.

### 5.4.1 Overview of the Model

The proposed model aims at complementing currently available software quality assurance efforts and should not be mistaken as an exclusive quality assurance endeavor. It seeks to cover the existing gap in software engineering when leaving aside ethical and social concerns of the produced solution while mainly focusing on client satisfaction through technical verification of processes and product specifications. This proposal strongly suggests we apply a technical oriented quality assurance model according to the project characteristics, together with the Capability and Maturity Model for Trustworthy Algorithmic Decision-Making systems.

The main objective of the model is to assure trustworthiness on algorithmic decision-making systems in terms of fairness and non-discrimination. It has as a reference the Principled AI International Framework, as defined in Chapter 2, and the available knowledge on the specialized literature described in section 5.2 of the present chapter while guiding actions towards an ethical design of the aforementioned software solutions. Figure 5-1 exhibits the characteristics of the model's domain.



**Figure 5-1: Simplified Model's Domain Conceptual Map [Own Elaboration].**

### 5.4.2 Model's Principles

The following are the guiding principles supporting the model:

1. Principles driven. The model is driven by the Principled AI International framework as the main reference corpus of international regulations concerning the ethics and design of ADMS as well as for software engineering methodologies, quality assurance models, and international technical norms and standards to decrease the number of biased and discriminatory ADMS' outcomes.
2. Integrative and low interference. The model is propelled by an adaptable and low interference philosophy through its integration within the quality assurance activities flow, regardless the project's engineering methodology style and adjusted to established processes generating the least possible number of extra processes, subprocesses, information/communication flows, or artifacts. When needed, additions must emphasize economy of action by keeping the technical approach and artifacts as

simple and concise as possible which will reduce the project team's resistance to the model.<sup>36</sup>

3. Manager-centered. The model is constrained by the manager's commitment, as quality assurance tasks, trustworthiness assurance needs the manager's commitment and support for its proper implementation. Also, the model is limited by the manager's understanding of the project's scope, and their awareness (during the project conceptualization stage) of the possible negative impact of the projected solution.

4. Collaborative and Cooperative. It is based on the collaborative and cooperative character of professional relations among project members, especially in the measuring process to facilitate data gathering and quality measurements.

5. Minimalist. In demand of minimum effort, the model must be planned and executed as part of the quality assurance processes with care not to exceed the project's effort estimation for such activities.

### 5.4.3 Model's Approach

The scientific approaches conducted in the model's design include the following:

1. Systemic, expressed through the orchestration of the different components forming the model's architecture, the integration of the referenced Principled AI International Framework, and the interaction of processes to produce the model's outcomes.
2. Strategic, expressed in the model application as a means for organizations to obtain information supporting subsequent managerial decisions in the pursuit of a trustworthy performance through the production of fairer and less discriminatory ADMS.

---

<sup>36</sup> Based on the theories in the Software Engineering Body of Knowledge SWBoK and other software engineering authors like Roger S. Pressman and Ian Sommerville.

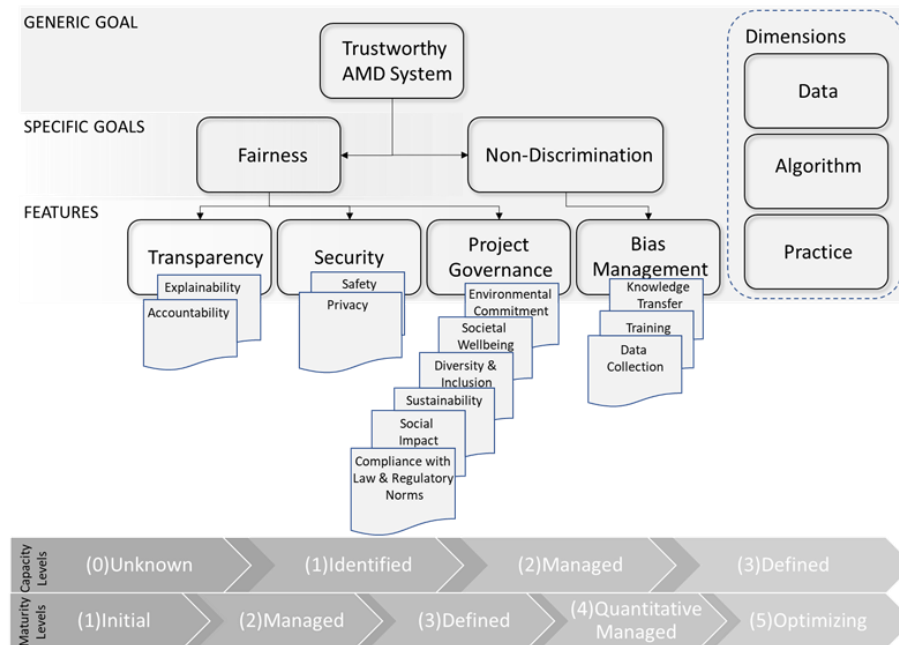
### 5.4.4 Model’s Quality and Premise

The quality distinguishing the model is expressed in the integration of the Principled AI International Framework into the software quality assurance tasks in order to add the trustworthiness scope into that workflow.

The premise referring to the model implementation is expressed through the organization’s willingness to elevate their process’s efficiency, specifically, the process of quality assurance management, by incorporating a trustworthy dimension to it.

### 5.4.5 Model’s Structure

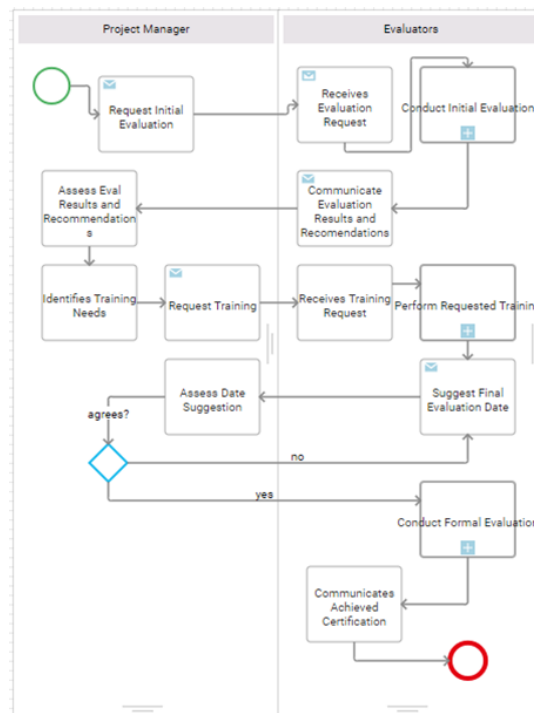
The proposed model consists of two stages. One stage on the software development project’s side, and the other on the external evaluation, training, and certification authority’s side. Figure 5-1 shows the model’s visual representation. The application of the model to the software development project’s side follows the workflow established by the quality manager according to the project’s methodology style and project type. The process map for the external evaluation, training, and certification authority’s side is specified in Figure 5-2.



**Figure 5-2: Model’s Structure [Own Elaboration].**

As shown in figure 5-2, the model adopts the capability and maturity level's philosophy from CMMI. Like the CMMI model, the present model concedes by default the capability level of (0) Unknown, and the maturity level of (1) Initial to all projects. The model's generic goal of trustworthy AMD systems is achieved through its specific goals regarding fairness and non-discrimination, according to the declared main objective of the model. The features of transparency, security, project governance, and bias management are handled through the specified capability and maturity levels attending to an array of specific goals, oriented towards defined variables and the listed three dimensions.

The model certification process can be seen in Figure 5-3. It is supported by three sub-processes, that is, initial evaluation, training, and formal evaluation, that end with an evaluation report and the issuance of a certificate indicating the project team's trustworthy ADM certificated level. The inputs and outputs for both stages of the model are detailed in the following section.



**Figure 5-3: Simplified Representation of the Model's Certification Process [Own Elaboration].**



### 5.4.6 Model's Inputs and Outputs

The two proposed model stages have different scopes and layouts; however, their inputs are similar. While the software development project's side stage focuses on assuring trustworthiness as an expanded quality characteristic the external evaluation, training, the certification authority's side stage aims at the model verification and auditing its execution.

As for the software development project's side stage, the inputs include the Project Plan, Quality Assurance Plan, Risk Management Plan, Analysis and Design Specifications, planned architecture, and any other managerial document according to the followed engineering style and project characteristics the Project Quality Manager esteems useful for planning the trustworthy assurance. The resulting list of documents experiences several updates along the project lifecycle; therefore, it needs to be consulted periodically. In contrast, among the model's main outputs the following artifacts can be found:

- 1) Project's Accountability Statement.
- 2) Project's Privacy and Safety Statement.
- 3) Project's Impact Study.
- 4) Project's Data Management Plan.
- 5) Project's Guidelines for Algorithm and Model Training, Tuning, and Knowledge Transfer.
- 6) Working controlled pipeline to test datasets, algorithms, and models in an ideal and controlled environment.
- 7) Update of project's expedient core plans like Risk Management Plan, Quality Assurance Plan, Requirements Management Plan, Configuration Management Plan, Version Control Plan, Test and Verification Plan, Security Management Plan, Project Team Development Plan, etc. and other documents within the

project's expedient like the Product Baseline, Product Specifications, and the User Manual.

Concerning the external evaluation, training, and certification authority's side stages, the inputs are framed by the project's file and interviews with part of the project team from different software engineering areas, when conducting the initial evaluation to identify which level the engineering team is at; during the training phase the inputs include training needs, based on the initial project evaluation, and a tailored training plan; and lastly, during the formal certification stage, the inputs gather the updated version of the project's expedient, and interviews team members from different engineering areas that haven't participated in the initial round of interviews.

The model's output for the external evaluation, training, and certification authority's side stage application is a Trustworthy Formal Evaluation Report, a Trustworthy Certificate, and a Seal endorsing the project team in one of the model's levels.

### 5.4.7 Models' Features, and Capability and Maturity levels

The proposed model aims at managing four features along a project's lifecycle. The capability approach follows a continuous representation expressing the project team's agency to add a given feature or group of features, to their managerial scope, and it is measured against the achievement of the specific goals for each feature. Similarly, the maturity approach follows a staged representation expressing the project's team agency to manage, and up to what state, the features they already have beneath their managerial scope. It is measured against the features themselves.

#### 5.4.7.1 Transparency Feature

##### 5.4.7.1.1 Transparency Features Specific Goal

Commit to design transparent solutions through explainable design and function of ADMS, and accountable authorship; for humans to easily perceive, detect, and understand the designed ADMS decision process.

#### 5.4.7.1.2 Transparency Feature's Specific Objectives

- SO1. Determine the liability of the project team upon possible negative impact on the population targeted by the ADM system that is being built.
- SO2. Partner with stakeholders to implement a reasonable accountability framework applicable to ADM Systems.
- SO3. Establish a public record of authorship upon the algorithms and decision-making models on the project's architecture's baseline reaffirming individual accountability.
- SO4. Foster explainable alternatives among ADM systems.

#### 5.4.7.1.3 Transparency Feature's Specific Practices

- SP1. Perform a risk analysis once the project is conceptualized considering the possible negative impact on the targeted population and the cultural background of such populations, taking into account the different geographical locations in which the ADM systems will be deployed to identify potential harm.
- SP2. Identify the project team's liability upon the identified potential harm given the context of the projected ADMS and communicate the high managers and clients the Risk Management Plan.
- SP3. Involve a multidisciplinary team of stakeholders to complement the group of functional experts with other disciplines related to the domain of potential harms the ADM system might cause on the targeted population once deployed.
- SP4. Determine roles, responsibilities, workflow, inputs and outputs, and monitor tasks in the design of an accountability framework particularized to the characteristics of the project and its target populations.
- SP5. Document the procedures for data collection, data preparation, and processing including the author of derived artifacts, data's metadata, results of a preliminary

test on data pipeline framework, negative outcomes, and the triggering conditions and dataset's characteristics.

SP6. Implement, as an ad-hoc service of the project, a trace log registry to record, when possible, data pivoting from parameters status to the decision/prediction final form, and any intermediate status, supporting subsequent reports on algorithmic functioning to enable those affected by an AI system to understand the outcome, and challenge it based on provided information on the factors, and the logic that served as the basis for the prediction, recommendation, or decision.

SP7. Document ADM system authorship for every implemented and acquired component, and the responsibilities derived from the ADM system's use for each known and projected context to support better informed decisions regarding its future use disclosing unintended bias present in the data and the algorithms, and to clarify the responsibilities of researchers, developers, users, and relevant parties.

SP8. Establish periodical verification procedures and internal audits to the architecture's baseline to ensure how the use of explainable variants are favored through the project lifecycle.

SP9. Encourage the re-use of previously developed components with unknown produced harm over time, which comply with the explainability, and accountability related practices herein described.

SP10. Conduct, as part of hired support and maintenance projects, monitoring of the original risk assessment on previously deployed ADM solutions to update the organization's risk bank for similar projects in similar contexts.

SP11. Monitor user complaints on previously deployed ADM that can be addressed in current open projects of similar scope, target population, and context.

## 5.4.7.2 Security Feature

### 5.4.7.2.1 Security Feature Specific Goal

Commit to design safe solutions considering personal data privacy and safety concerns in the design of ADM systems to avoid unreasonably curtail people's real or perceived liberty.

### 5.4.7.2.2 Security Feature Specific Objectives

SO1. Respect and guarantee individual privacy protection.

SO2. Acknowledge individual ownership and rights upon the data they generate.

SO3. Promote the use of state-of-the-art cryptography and security standards enabling trust and interoperability between ADMS and third party's solutions and services.

SO4. Partner with stakeholders to implement a privacy and safety framework applicable to ADM Systems.

SO5. Foster prudent alternatives among ADM systems, their components, and third parties they interoperate with.

### 5.4.7.2.3 Security Feature Specific Practices

SP1. Include, by default, rights to access of the target population of the ADM systems, manage and control the data they generate, and the ability to opt out from the data collection process into the requirements engineering workflow, regardless of their role (system users or target population) or without conditioning capacity to benefit from the offered digital service, in acknowledgment of their ownership and rights upon their data.

SP2. Balance the collection and processing of biometric data and other personally identifiable information in proportion to its stated purpose, based on justifiable need, scientifically recognized methods, and held and transmitted securely.

- SP3. Balance the ADM system's intrusion, and the subsequent personal and identifiable data acquisition and storing, in spaces where its target population is not subjected to surveillance or digital evaluation, based on justifiable need, scientifically recognized methods, and held and transmitted securely.
- SP4. Balance the ADM system's agency for profiling its target population and influencing their behavior without their free and informed consent, based on justifiable need, scientifically recognized methods, and held and transmitted securely.
- SP5. Perform a risk analysis once the project is conceptualized considering potential attacks on the data and data storage structure might be subjected to once the ADM built solution is deployed in a given network topography.
- SP6. Include a study of the impact of ADM over elements like the intimacy of thoughts, preferences, and individual emotions of their target population and the potential probability of the ADM system imposing moral judgments or segregation because of their lifestyle choices as part of the risk analysis.
- SP7. Include, ex-officio, ADM system's functionalities to guarantee the protection of the ADM system's target population's intimacy of thoughts, preferences, and individual emotions into the requirements engineering workflow to comply with the results of the dedicated conducted risk analysis.
- SP8. Include, ex-officio, ADM system's functionalities enabling data anonymization, and de-identification of digital identity into the requirements engineering workflow to protect personally identifiable information.
- SP9. Include, ex-officio, ADM system's functionalities enabling differentiated treatment to identity types like minors, students, workers, patients, convicts, and other people of interest, into the requirements engineering workflow to provide them with specific guarantees given their information's distinct sensitivity and vulnerability degrees.

- SP10. Include, ex-officio, ADM system's functionalities to guarantee human control over the system considering the specific context in which a particular system operates, into the requirements engineering workflow to provide an extra safety layer on specific domains such as defense and automated weaponry.
- SP11. Include, ex-officio, ADM system's functionalities guaranteeing the target population's universal right to be forgotten into the requirements engineering workflow ensuring periodical deletion of data segments linked to personal identification.
- SP12. Document and release to the client, as part of the project's expedient and product's specifications, known information on potential cyberattacks or hacks the AMD system being designed might be vulnerable to supporting better future consumer protection mechanisms.
- SP13. Incorporate state-of-the-art security standards involving data privacy and safety applicable to the context determined by the project and the target population characteristics given their different cultural backgrounds.
- SP14. Ensure that third party's components abide by shared safety and privacy organization's notions before its integration as part of the ADM system being built.
- SP15. Partner with specialized stakeholders given the project and ADM system's target population's characteristics in designing a framework for privacy and safety related issues, considering identified elements from the risk assessment and possible unexpected or undesirable distribution and use of personal data identified from regular testing, risk bank updates and monitoring procedures while deployment and maintenance stages.
- SP16. Determine roles, responsibilities, workflow, inputs and outputs, and monitoring tasks in the design of a privacy and safety framework particularized to the characteristics of the project and data sensitivity.

- SP17. Foster the use of an organization's internal open data policy, following data access protocols, to stress and study the data minimization, representativeness, storage limitation, integrity and confidentiality, and data structure resilience on a controlled data pipeline framework.
- SP18. Document negative outcomes and the triggering conditions negatively impacting expected data stability and data structure resilience.
- SP19. Implement, as an ad-hoc service of the project, a trace log registry to record data access and purpose of the access petition supporting subsequent reports to enable subsequent audits conditioning improvements to the defined data structure and access protocols.
- SP20. Include, as part of the project's test plan, the verification of potential harmful uses of the designed ADM system.
- SP21. Communicate, as part of the project's expedient, identified faults (along with their context, scope and triggering conditions) endangering the ADM system's target population safety and the probability of occurrence.
- SP22. Suggest public access and algorithm dissemination restrictions in correspondence with the identified ADM system faults compromising the target population safety.
- SP23. Conduct, as part of hired maintenance projects, monitoring of the original risk assessment on previously deployed ADM solutions to update the organization's risk bank for similar projects in similar contexts regarding personal privacy, safety, and how the projected data structure and data access protocols are influenced in a real-life scenario.
- SP24. Monitor user complaints, traceable to privacy and safety, on previously deployed ADM that can be addressed in current open projects of similar scope, target population, and context.



SP25. Encourage the re-use of previously developed components that have not produced harm over time, which comply with the privacy and safety related practices herein described.

### 5.4.7.3 Project Governance Feature

#### 5.4.7.3.1 Project Governance Feature Specific Goals

Commit to an inclusive and sustainable project management style, which is respectful the law and international regulations, oriented to a positive impact on society, the wellbeing of society, and environmental sustainability.

#### 5.4.7.3.2 Project Governance Feature Specific Objectives

SO1. Manage diversity and inclusivity concerns on the project team staffing process by integrating elements acknowledging candidates' personal characteristics and particular preferences to the traditional competence profile-oriented hiring approach, to enrich the team composition and provide equal opportunities to underrepresented candidates.

SO2. Comply with available specialized laws and international norms as applicable in the context of the project application domain, including those oriented on harnessing ADM systems' negative impact on society, to ensure an ethical and lawful design of the solution being built.

SO3. Avoid the design and deployment of ADM solutions that significantly compromise individual and society's wellbeing, or environmental sustainability.

SO4. Partner with stakeholders to conduct exploratory studies on the impact of the projected ADM solution on its target population at individual and societal levels, their environment and sustainability.

SO5. Establish mechanisms to support the professional growth of project team members. Acknowledge them as valuable members of the organization.

### 5.4.7.3.3 Project Governance Feature Specific Practices

- SP1. Incorporate, during project conceptualization, team composition related needs aligned with the project characteristics, developing multi-disciplinary team distributions, with diverse gender, ethnic, cultural, and socio-economic backgrounds to strengthen, from the project staffing stage, the project teams' awareness of potential biased outcomes produced by the projected ADM system.
- SP2. Perform, during project conceptualization, a comprehensive study of the available legal framework applicable to the project domain to ensure compliance through engineering methods, considering dimensions like data, algorithmic functionalities, and engineering practices within the scope of the study.
- SP3. Document, as part of the project, compliance with available and applicable international and local laws related to the projected ADM solution application domain.
- SP4. Document, as part of the project, the adoption of available and applicable international, local, and organizational engineering standards and norms.
- SP5. Include, ex-officio, a set of rules and constraints for the projected ADM system, as part of the requirement engineering phase to establish from the learning and training stage and implemented during the knowledge transference stage the ADM system functional adherence with an applicable legal framework according to its application domain.
- SP6. Align the projected ADM Systems set of rules and constraints related to the applicable legal framework with the different geographical and cultural contexts in which it is planned to be deployed.
- SP7. Design an organizational ethical code of conduct that project team members need to adhere to. Include elements that connect engineering practices and deployed biased products of current and past projects, to create, maintain and improve an organization's ethical culture.

- SP8. Design a response mechanism enabling the organization to deal with possible breaches of information once the projected ADM solution is deployed.
- SP9. Communicate with clients and end-users given the occurrence of an information breach once the projected ADM solution is deployed.
- SP10. Perform impact studies as part of the project conceptualization stage to identify potential harms to the target population, their environment and sustainability.
- SP11. Perform risk analysis impact studies to establish adequate risk management strategies considering the applicable legal framework.
- SP12. Balance the benefits and risks the projected ADM solution brings to its target population, general society, and environment to allow clients and end-users' to make better informed decisions.
- SP13. Design testing plans including features oriented at identifying (a) where the projected ADM solution may cause harm to the target population's living and working conditions, health, and universal rights; (b) limiting the pursue of their preferences when not causing harm to other sentient beings; (c) limiting the exercise of their mental and physical capacities; (d) increasing their stress, anxiety, or sense of harassment; (e) becoming a source of any other ill-being, unless it allows the achievement of a superior well-being than what could attain otherwise; (f) perpetuating the status-quo of underrepresented populations; amplifying economic, social, gender and other inequalities.
- SP14. Include, ex-officio, functional and non-functional requirements regarding the projected ADM solutions energy consumption, waste production, and pollution generation into the requirements engineering workflow, which extends to needed infrastructure within the scope of the project, to produce and deploy sustainable and eco-friendly solutions.
- SP15. Design mechanisms, as part of the organization's post-sales services, to handle the extraction of resources and the ultimate disposal of equipment, that were

deployed as part of the hired project when they have reached the end of its useful life, to minimize negative environmental impact.

SP16. Design recycling mechanisms to re-use and re-locate material in other projects or new projects to minimize the organization's negative impact on the environment, and costs.

SP17. Promote, whenever needed and possible, partnerships with other industry actors, academic institutions, and government organizations to increase innovation, increase product and service scalability, and attract financing ventures as the main expression of a sustainable project management style.

SP18. Promote project team members training opportunities on topics related to the organization's product portfolio and ethical issues intrinsic to project application domains, along with other technological and business-related training needs.

SP19. Incorporate the legal repercussions and available remedy mechanisms in response to the impact of possible negative outcomes from the product's function as a caution in the User Manual accompanying the projected ADM solution.

#### 5.4.7.4 Bias Management Feature

##### 5.4.7.4.1 Bias Management Feature's Specific Goal

Commit to design neutrality on the projected ADM system minimizing the number of subsequent produced biased outcomes.

##### 5.4.7.4.2 Bias Management Feature's Specific Objectives

SO1. Design a minimum-bias tolerance methodology for data collection and data preparation to avoid transferring and amplification of biases into designed algorithms and models.

SO2.Design mechanisms to ensure the neutrality of algorithms and models regardless of biases in the training datasets.

SO3.Design test and verification strategies to maximize the number of bias findings in the ADM system.

SO4.Manage engineering team members' biases to avoid their inclusion in the ADM system design.

SO5.Partner with stakeholders to implement a non-discrimination framework applicable to the ADM system according to project characteristics.

#### 5.4.7.4.3 Bias Management Feature's Specific Practices

SP1. Perform a risk analysis assessment once the project is conceptualized, considering potential biases the ADM system's target population might experience once deployed, based on demographical characteristics and cultural beliefs.

SP2. Conduct exploratory studies on the characteristics of the projected ADM system's target population to determine the characteristics of a representative dataset, and control datasets; exhaustively including all variables within the project domain.

SP3. Include a study of the impact of potential biases given the characteristics of the projected ADM system's target population and avoid the implementation of functionalities that socially impair individuals or groups.

SP4. Involve project domain specialized stakeholders in the definition of a tailored data integrity reference to periodically contrast datasets along the project lifecycle.

SP5. Extend configuration management to dataset related artifacts ensuring a baseline for training datasets, one or more for control datasets, and an adequate registry for dataset changes.

- SP6. Encourage, whenever possible, collecting data from scratch when building training datasets, for example by capturing data logs straight from the source, and avoid overusing pre-existing datasets for similar domains.
- SP7. Design a mechanism to evaluate pre-existing datasets within related project domains before its incorporation into the project training baseline to avoid the re-use of pre-existing biases through the projected ADM system.
- SP8. Minimize dataset sizes without affecting their representativeness of the target population within the context of the project.
- SP9. Document a dataset description as part of the project to provide a testable explanation of the type and purpose of data being gathered to avoid unnecessary data harvesting.
- SP10. Develop a pipeline in which datasets can be stressed and tested, in a controlled environment, to reveal hidden biases, and to avoid those biases during model training and embedded into the projected ADM system.
- SP11. Document the revealed biases within the dataset, the conditions which revealed them, and a detailed description of the impact on the demography of the population.
- SP12. Design an algorithm and model tuning mechanism, based on the “many-eyes” philosophy, involving several testers and several methods to mimic an expert consultation with as many factors involved to reach an informed decision.
- SP13. Conduct parallel verification of different versions of models and algorithms before incorporating them into the projected ADM system’s configuration baseline.
- SP14. Perform, ex-officio, periodical reviews for data integrity, algorithm and model permutations from its original deployed versions, to identify unintended biased outcomes at different levels of completeness of the data base projected growth, especially in sensitive application domains.

- SP15. Design a mechanism to evaluate pre-existing third-party ADM system components within related project domains before its incorporation into the project's baseline to avoid the incorporation of pre-existing biases into the projected ADM system.
- SP16. Ensure that every ADM system's outcome can be overridden or reversed by designated actors.
- SP17. Conduct, as part of hired maintenance projects, monitoring of the original bias considerations on previously deployed ADM solutions to update the organization's risk bank for similar projects in similar application domains and the conditions that revealed the uncovered biases.
- SP18. Monitor user's complaints, traceable to biased and discriminatory outcomes, on previously deployed ADM that can be addressed in current and future projects of similar scope, target population, and context.
- SP19. Encourage the re-use of previously developed components with that are unknown to produced discriminatory outcomes over time, which comply with the least-bias philosophy related practices herein described.

#### 5.4.7.5 Capability and Maturity Levels

As mentioned earlier, the model adapts CMMI quality assurance model's capability and maturity levels as follows:

➤ Capability levels

A project is considered to exhibit capability levels according to its ability to achieve the features generic goal up to that level. Every project is considered to be level (0) Unknown by default.

- Level (0) Unknown: describes projects where none of the model's features have been identified among the assurance needs.

- Level (1) Identified: describes projects where there is evidence of assurance tasks oriented towards the model's features, or towards the achievement of the model's features specific goal.
- Level (2) Managed: describes a project where (1) there is evidence of a planned process or processes, which are executed with the specific goal of the model's features; (2) employing skilled people; (3) having adequate resources to produce controlled outputs; (4) involving relevant stakeholders; (5) is monitored, controlled, and reviewed; (6) and is evaluated for adherence to its process description. That process or processes help to ensure that existing practices are retained during times of stress.
- Level (3) Defined: describes a project where (1) the processes are defined, institutionalized, and tailored from the organization's set of standard processes according to guidelines; (2) has a maintained process descriptions; (3) and contributes process related experiences back to the organizational process assets. It distinguishes from level two, where process instances can be quite different, by exhibiting more consistency and level of detail.

➤ Maturity levels

On the other hand, a project is considered to exhibit maturity levels according to its agency to implement the specific practices of the model's features providing a way of evaluating the project's performance. Every project is considered to have a level (0) Initial by default.

- Level (0) Initial: describes a project where the proposed feature's specific practices are implemented chaotically without a stable environment supporting their orchestration into a plan or process, successful implementation of proposed specific practices depends on the competence of team members rather than on the definition of proven



processes, procedures fluctuate in time of crisis or with frequent team composition variation.

- Level (1) **Managed:** describes a project where the proposed feature's specific practices are planned and executed according to guidelines; similar to capability level two, the project employs skilled people who have adequate resources to produce controlled outputs; involve relevant stakeholders; are monitored, controlled, and reviewed; and are evaluated for adherence to their process descriptions.
- Level (2) **Defined:** describes a project where the proposed feature's specific practices are orchestrated into defined processes, which follow standardize procedures, tools, and methods; and are used to establish consistency across the organization.
- Level (3) **Quantitative Managed:** describes a project where quantitative objectives are set to assess the project's performance in the execution of processes defined to implement the proposed model's feature's specific practices and use them as criteria in managing the project.
- Level (4) **Optimizing:** describes a project that is continually improving its processes based on quantitative criteria expressed through the analysis of the process variations and the causes of process outcomes, and its performance needs; there exist quality objectives established, continually revised to reflect changing business objectives and organizational performance, and used as criteria in managing process improvement.

The next section presents three examples of unfair decisions made by ADM systems discriminating individuals because of their race, gender, or their condition as landed immigrants, emphasizing the Hispanic situation in the USA, in particular.

## 5.5 Examples of Domains in Which ADM Systems Have Proven Discriminatory Outcomes

### 5.5.1 Example 1: Health Care System

There are algorithms used to approach optimization class problems that consider cost as a variable to measure their accuracy during training (Hileman, 2016). That can be problematic in cases where the decision concerns a patient's permanence in a hospital bed or the accommodation of home-visiting nurse services, where the expected focus variable must be illness (degree, risk, etc.) rather than associated costs (Vogeli, 2007; Bates et al., 2014).

The commercial approach of the USA health care has conditioned ADM systems to discriminate against non-white patients (Obermeyer et al., 2019), even when race was not part of the parameters considered on the decisions when hospitals and insurance companies identify which patients will benefit from “high-risk care management” programs.<sup>37</sup> The algorithms learned to signal, as required, sicker patients for more organized and specific attention, minimizing costs and maximizing client satisfaction; however, while doing so, it also learned to discriminate against some individuals with certain medical billing patterns.

In their study (2019), Obermeyer and others observed that black patients would pay bills for the same amount than less sick white patients given the former's trend to visit hospitals for more serious procedures while the latter are more able to receive prevention and prophylactic treatments. Although the study focused on the difference between white and black patients, it does stress that such racial gap accentuates when looking at other minorities such as the Hispanic community. The algorithm learned to equate similar billing patterns, without considering the actual need for medical attention,

---

<sup>37</sup> The “high-risk care management” program helps to provide chronically ill people with extended care at hospitals, access to specially trained nursing staff and allocate extra primary-care visits for closer monitoring.

heightening disparities with basis in race because of its correlation with cost affordability and actual request for medical attention variables.

### 5.5.2 Example 2: Hiring Processes

According to (Dastin, 2018)<sup>38</sup> big tech companies like Amazon, Facebook, Apple, Google and Microsoft exhibit a gender gap with male to female ratios among their employees of 60-40%, 64-36%, 68-32%, 69-31%, and 74-26% respectively. When looking at their employees in technical roles the ratios considerably favours male employees as follows: 77-23% in Apple, 78-22% in Facebook, 79-21% in Google, and 81-19% in Microsoft.

Several studies criticize Amazon’s algorithm for automated hiring. Particularly, Bornstein (Bornstein, 2019) explains that not only do the algorithm’s discriminatory patterns respond to the past hiring model of the company, by means of the Curriculum Vitae received in the prior 10 years and their resultant candidate selection, but that the algorithm learn from the language used in those CVs. That unexpected trait was denominated “Male language” and manifested by the use of certain terms like “Execute” and “Captured” verbs, and the discrimination of CVs containing chains like “Women’s chess club captain,” for example.

This is another example of how unexpected the sources for bias triggering discrimination can be. As language is intrinsically and fundamentally related to culture, it is appropriate to infer that there exists a greater gap within the Hispanic applicants wherever this hiring mechanism is applied to select candidates from a pool of applicants. In consequence, and agreeing with Ajunwa (2021) we suggest that algorithms used in the hiring process should be auditable and distinguished by a seal within the labor market expressing that: (1) the data used in training must represent the “what should be,” and not the “what it is,” (2) data and algorithms need to be auditable, including the data capturing

---

<sup>38</sup> Dastin’s study points that Amazon does not publish gender breakdown since 2017, which is the same year the company stopped a project that used AI to suggest the top five candidates to be hired out of a pool of resumes.

process; and (3) the model evaluation should include an evaluation conducted by social groups.

### 5.5.3 Example 3: Recidivism Risk Assessment During Pretrial, Sentencing, and Paroling Assessment

The USA's judicial system uses a software called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (rebranded to "Equivant" in 2017) to assess recidivism in defendants before trial, during trial in the determination of the sentence, and while assessing parole opportunities for already convicted individuals. The software is widely used across the USA. And it is estimated that it helps incarcerating a gross number of offenders every year.

Several studies have been interested in dissecting the "black box" algorithm in COMPAS which resulted in an ethical discussion about race disproportion in the inmate population in USA and subsequent loss of trust on such "supportive" software. On the one hand, some studies (Dressel & Farid, 2018; Kehl, Guo, & Kessler, 2017) criticize a racial bias in COMPAS, analyzing the false positive and false negative rates exhibited by it produced outcomes; on the other hand (Flores, Lowenkamp, & Bechtel, Sn.) defended the software against alleged racial biases.

Kehl and others (Kehl, Guo, and Kessler, 2017) conducted a historical review of USA efforts to predict what was then denominated dangerousness, until fourth generations predictive tools such as ML were incorporated to build models to assess what its currently known as recidivism. These recidivism tools respond to a trifold model: first, in response to the risk principle, which asserts predictability of risk, helping to determine differentiated and more intensive treatment to high risk flagged than low risk flagged offenders; second, following the needs principle, while supporting comprehension upon several features of criminal behavior, suggesting rehabilitative treatment, and sentencing decisions in response to criminogenic needs; and third, helping support the responsivity principle, suggesting tailored treatment to the specific offender. During sentencing, the recidivism prediction tools may be used to determine the proper punishment, including

diversion from prison to jail, diversion from jail to community service or home-arrest, and fines.

The study calls the attention to COMPAS as a proprietary software manufactured by Northpointe, that assesses 137 features under five main areas: criminal involvement, relationships/lifestyles, personality/attitudes, family, and social exclusion to arrive to a prediction, supporting judges in the process of sentencing. It also points the fact that COMPAS is not subject to federal oversight, what, along with the lack of transparency about its inner workings, including how it weighs the variables the model works with, outline the main reasons for which there is controversy around the software. Additionally, the study suggests that COMPAS disrupts the defendant's right to a fair trial when it violates the right to be sentenced based on accurate information as the proprietary nature of COMPAS prevented the defendant from assessing the accuracy of the assigned score, and the right to an individualized sentence as COMPAS relies on knowledge learn from a larger group with similar characteristics before making inferences about the defendant's personal likelihood to commit future crimes, among other reasons.

Lastly, Kehl's study references Kathy O'Neil's book "Weapons of Math Destruction" to stress how other elements like the defendant's residence postal code may trigger a racial discriminatory scoring. Kehl and Kathy agrees that individuals from lower income neighborhoods will exhibit a larger criminal record that individuals from higher income neighborhoods, merely because of more aggressive policing strategies. Therefore, individuals from low-income neighborhoods receive higher scores of risk assessment, and consequently they receive stricter punishments and larger sentences. That logic extends to all marginalized minorities such as the Hispanic population (Kehl, Guo, and Kessler, 2017).

Parallely, Dressel and Farid (2018) conducted a more qualitative critique to COMPAS. They performed an experiment in which they asked 20 participants, untrained in criminal law, and with no law administration expertise, to predict whether a person would reoffend within two years of their most recent crime. The participants judged a total of 1000 felons, in 20 sets of 50 subjects each, based on a brief description containing

the defendant's sex, age, and previous criminal history, but not their race. The human predictions were contrasted to COMPAS's using a simple lineal regression test.

The results show similar levels of accuracy, and the same false positive and false negative production rates in favor of white defendants over Black subjects. This suggests that COMPAS is no more suitable to predict recidivism than a group of random individuals. Additionally, it is worth to mention that only two features (age, and number of previous felonies) were found to be relevant when considering the human judges' predictions, in contrast to COMPAS's 137. That difference in the number of features that are relevant to the decision for that specific scenario, raises concern regarding the need for all other personal information COMPAS requires to arrive to an evaluation. It is an element that is considered within the scope of the right to privacy explored in previous chapters, regarding the governance of own information, and stating that institutions must access to what information based on a justified end.

COMPAS's racial bias was argued by Flores and other colleagues in (Flores, Lowenkamp, Bechtel, Sn.). However, their arguments vaguely justify that the software exhibits more accuracy than previous models and non-automated instruments. Interestingly, they present data regarding the racial distribution among the general USA population, and the inmate's racial breakdown as per the US census of 2014 as part of their argument.

The study points the 2014 census estimated that the racial breakdown of 318 million USA residents consisted of 62.1% white, 13.2% Black or African American, and 17.4% of Hispanic residents. Contrastingly, 37% of the prison population was categorized as black inmates, 32% was categorized as white convicts, and 22% categorized as Hispanic. Referencing Carson (Carson, 2015), to better visualize these distributions, 2.7% of black males (or 2,724 per 100,000 black male residents) and 1.1% of Hispanic males (1,090 per 100,000 Hispanic males) were serving sentences of at least 1 year in prison, compared to less than 0.5% of white males (465 per 100,000 white male residents). It is therefore correct to infer that the same distribution pattern from which the COMPAS software learned by means of a summarized dataset, is the same pattern it will

reproduce. Hence, disfavoring racial minorities like Hispanic and Black individuals. Eijo de Tezanos (2016) demonstrated that there is no statistical difference among the likelihood to be incarcerated without conviction when looking at parameters like age or gender, nor among races when comparing Hispanic and Black individuals, in contrast to when comparing Hispanic individuals with white subjects.

These three study cases exemplify how ADM technologies exacerbate social gaps within a society like the USA, and at the time it illustrates different ways in which the social contract resulted disrupted by the same matters that distinguished the focus of contemporary notions of justice in political philosophy such as the inequalities linked to gender, and race, in such determinant scenarios like access to health care, jobs, and freedom. The next section presents a brief description of how the discriminatory triggers from the three presented study cases are voided with the implementation of the proposed capability and maturity model for trustworthy ADM.

## 5.6 A Brief Description of How the Proposed Capability and Maturity Model for Trustworthy AI Helps Mitigate ADM Systems' Discriminatory Decisions

When inspecting the main causes triggering discriminatory outcomes on each one of the presented study cases in the previous section, it can be noticed that: for case one, race related features, specifically the habits for seeking medical attention, acted as a trigger for racial discrimination; similarly in case three, race related features, this time the area of residence, acted as an indirect (throughout the aggressive policing strategies given the neighborhood characteristics) trigger for racial discrimination; in contrast, for case two, gender related features like a certain use of language acted as a trigger for gender discrimination, along with the reproduction of the existent patterns of gender distribution, which also influenced the discriminatory outcome. This of course presents a simplified summary of a complex situation, although appropriate to the purpose of illustrating the functioning of the proposed model.

In regard to the discriminatory sources, it can be said that they are unexpectedly related to protected attributes like race and gender, as can be noticed a common aspect

across the three examples. This supports the review of fairness by unawareness criticized in chapter one, and it also illustrates the weakness of Rawls's "veil of ignorance" in the context of ADM solutions, where ML can make any sorts of inferences with available information.

In order to present a view of the implementation of the proposed model, a different feature of the model is used for each study case, showing different levels of capability and maturity, in a way that we can describe as much of the model as possible without being extensive.

### 5.6.1 Transparency Feature

The feature Transparency directs the project management's commitment with a transparent design in terms of explainability, and accountable authorship. All stated Specific Objectives SO1 to SO4 are applicable to case one. For the purpose of the example, let's assume there are evidence for the implementation of each of the specific objectives in the feature so that it can be said that, to the eye of an auditor/evaluator the model exhibits a capability level one "Identified."

Specifically, SO1 and SO2, are particularly useful to illustrate a maturity level two "Managed" through the descriptions of related specific practices SP1 to SP5. It must be noticed that this section shows only a partial simplified description of the implementation of the model, for illustration purposes.

According to the specific practice SP1, after the project conceptualization, as part of the risk analysis the Project Risk Manager or any other designated role (the model follows the principle of least interference) conducts an analysis of possible negative impacts of the projected solution on patients needing extended care. The resulting analysis is incorporated to the Risk Management Plan, registering the identified negative events, their probability of occurrence, response strategy, and responsible individual within the members of the project.

The risk description and their subsequent response strategy must reflect the characteristics and cultural background of the affected demographics in regard of the



locations in which the projected solution will be deployed to identify and address any potential harms traceable to those traits. That include, to provide some examples: 1) the risk of medical billing describing a different reality for different patients, given the relation medical bill and illness of the patient; and 2) the risk of racial related biases being triggered by other characteristics than race, to accommodate to the specifications of example one.

One possible strategy to mitigate risk 1), for example, could be creating different datasets with ranges of medical billings (in respect of the focus on costs) and the specifications of the related medical conditions, separating and combining other pieces of information like length of stay, needed care, etc. while monitoring accuracy changes with every pivoted variable. By separating each variable, it is easy to identify which one triggers an anomaly in the distribution of the observed in contrast to the expected results that may need further exploration within the scope of the proposed model's Bias Management feature.

The model conditionates the execution of this type of analysis, currently absent in software development methodologies more tuned with the functional aspect of the projected solutions and less with its ethical implications. However, it does not box nor restrain the development team in a specific way of doing things. The evaluators interview team members and review the project folder to determine what things correspond to the implementation of which specific practice, and assign a capability and maturity level, accordingly.

In correspondence with SP2's description, once the Risk Management Plan has incorporated all identified possible ethical and social negative triggers, the assigned role determines the project team's liability according to the identified risks and established response strategy. That could include, for example: the responsibilities of the data scientists and test specialists regarding the definition of a Data Capture Plan, to avoid the undesired racial bias permeated from other related features like the ones described in the previous paragraph. The Risk Management Plan is then communicated to the project managers and clients.

With the Risk Management Plan at hand, and the liabilities of the project's roles according to the specified Response Strategies, it is necessary to involve, as described in the specific practice SP3, a multidisciplinary team of stakeholders who will help the development team to gain more understanding of the environment of the projected solution, and to refine the response strategy accordingly. According to the particularities of the study case, the list of stake holders might include experts in medical billing, nurses with experience treating high risk patients at hospital, and nurses with experience treating high risk patients at home.

As a result of a refined Risk Management Plan, Risk Response Strategies, and Data Capture Plan, specific practices SP4 and SP5 assure the monitoring of roles' responsibilities, appropriate workflow, quality of inputs and outputs, as per the project's accountability framework; by documenting all procedure instances and their outcomes which anomalies can be used to feed back the project plan and the derived artifacts, like the ones referred at the beginning of the paragraph.

### 5.6.2 Security Feature

The security feature focuses on the project's commitment to safe design of the proposed solutions, taking special considerations towards personal data privacy and people's real or perceived liberty. Thus, this feature is more suitable to be illustrated with the help of the example three, specifically the unnecessary use of 135 personal features when the referenced experiment concluded the age and previous felony history were sufficient to arrive to the same conclusions.

All specific objectives SO1 to SO5 applies to the study case. The focus of the simplified example of an instance of the model implementation will be focused on specific objective SO1. The SO1 relates to the respect and guaranties of individual privacy protection. The description centers on SP2, which is linked to the referred specific objective SO1; in combination with SP17, which refers to the institution open data policy and the use of a controlled data pipeline framework to test the datasets. Although all other specific practices are applicable to the example, it must be clarified that SP1 and SP10 are excepted. SP1 relates to the particular cases in which the user

needs to grant permission for the collection of his/her data in order to receive a service. And SP10 aims at ensuring human control over ADM systems in the particular context of automated machinery and weaponry.

Specific practice SP2 of the Security Feature mainly deals with balance in the collection and processing of personal information in proportion to its stated purpose. The model is designed so that it suggests engineers to only incorporate the most essential attributes using the data pipeline framework described in SP17 of the same feature. In that regard, SP17 is an institutionally defined process that orchestrates the organization's internal open data policy to stress and study issues linked to data minimization, representativeness, among others; complementing the second half of SP2 that states "based on justifiable need, scientifically recognized methods, and held and transmitted securely." Although the model does not specifically provide guidelines to ensure the minimum necessary numbers of attributes to include in the decision, an example helping the auditor/evaluator to identify evidence of the execution of these specific practices may include stress tests to datasets, varying their dimensionality and monitoring accuracy to select the minimum necessary dimensionality and the corresponding attributes to be included in the decision, as part of the Test Plan.

For the purpose of diversity, the example description presented in the previous paragraph uses planned processes, defined, and standardized at the organization level and used in the project, to exemplify the proposed model's capability level two and maturity level three denominated as "Defined."

### 5.6.3 Bias Management Feature

The Bias Management feature, focuses on achieving neutrality in the projected ADM solution by minimizing the number of subsequent produced biased outcomes. The description provided below presents a summarized simplified instance of the implementation of the model in the context of example two, where an unexpected trait such as a certain use of language permeated into the ADM process influencing a gender biased outcome disfavoring female candidates when applying for a job opportunity. In this example, all specific objectives are applied through the 19 proposed specific

practices; however, specific objectives SO2 and SO3 provide enough information to visualize the implementation of the proposed model in this case. Specially, with the support of specific practices SP2, SP5, SP7, SP10, and SP11.

The example looks at Curriculum Vitae, the ADM needs to select the top 5 subjects, out of 1000 to be interviewed. Therefore, in following practice SP2, the team project conducts an exploratory study on the characteristics of expected CVs, and desired candidates to conform the particularities of a representative dataset. The identified variables resulting from the exploratory study helps the analysts (or other designated role) to build the necessary control datasets considering all variables of interest for the project established environment. They may include racial, gender, and other observable distributions in the group, and deviations from the desirable (ethical and socially just) distribution; along with sociolinguistics traits, as presented in the example, among other variables within the project domain. Parallely, test managers must design metrics to help quantitatively measure quality variables like representativeness of datasets, and accuracy-neutrality balance.

The results of the measures will allow the configuration manager (or another designated role) to ensure a stable baseline for training datasets, control datasets, and the registry for changes control of assessed datasets, as stipulated in SP5. Along with the suggestions of SP7, SP5 incorporates into the project baseline of already stressed datasets which biases have been extensively scrutinized, and hopefully identified using the data pipeline suggested in SP10. In the pipeline the datasets undergo stress tests to verify that the datasets are robust under tests conditions reflecting the identified risks and determined discrimination triggers. The test results are expected to be measured in order to qualitatively distinguish which datasets will be incorporated in the project training baseline, as well as to register the conditions under which the discrimination triggers are uncovered.

SP11 deals with the registry of the uncovered biases within the tested datasets, gathering the conditions that trigger discriminatory outcome so the analyst and designers can include them to their design as known restrictions. The SP11 also outlines the

necessary detailed description of the impact of identified unhandled biases so it can be added to the user manual and product specification. This way the projected solution follows the transparency principle and can attract more trust from end users.

SP14 and SP19 help bring neutrality to the evaluation process and datasets. These specific practices suggest periodical reviews of the data integrity, algorithm and model permutations from its original deployed versions, and unintended biased outcomes at different levels of completeness of the data base projected growth; and re-use of previously developed components with unknown produced discriminatory outcomes over time, contributing to the standardization of proven processes.

By applying the proposed model, as described in the example, it guarantees that data used in training AIS represents “What should be” instead of recycling pre-existing (most likely biased) scenarios and auditability of data and algorithms, and that the associated workflow from the data capturing stage is compatible with third parties allowing for independent evaluations.

The described instance of the implementation of the proposed model, as can be seen, represents a capability level three “Defined”, and maturity level four “Quantitative Managed”, as it exhibits an institutional define and standardized evaluation process, which is also quantitatively managed.

There is no need to describe an instance of the implementation of the Governance feature of the proposed model, as it gathers specific goals and practices related to the project management style rather than focusing on a particular study case.

## 5.7 Thesis Project’s Knowledge Mobilization Plan

This thesis project’s knowledge mobilization plan is designed to promote the proposed model among an audience of a) possible stakeholders to present them with a mechanism to evaluate the processes executed by ADMS development teams for the development of trustworthy solutions, b) ADMS project managers to help them understand the needed commitment with the ethical aspect of the impact of the solutions they build, and to introduce them with a model of methodological reference for the

development of trustworthy ADMS, and c) ADMS developers, especially to those involved in the quality assurance workflow to help them understand why it is important to consider the ethical impact of the solutions they build and how the proposed model could help them reduce the outcomes resulting in discriminatory decisions.

We will follow a two-stage approach to knowledge mobilization. The first stage will focus on communicating and disseminating the findings detailed as part of the present thesis document, and the second stage will focus on using the thesis document to develop a new research agenda focused on reflective and autoregulated ADMS. With respect to the first stage, we will communicate and disseminate the thesis document as follows:

First, we will present our findings in international conferences and to interest research groups such as the International Conference of Software Engineering ICSE and the Iberoamerican Network of Project Management, to promote our research findings with crosssectoral stakeholders and knowledge users. We will also share the thesis document with Canada's main AI research hubs like the Trustworthy AI Lab of the Ontario Tech University, the Ethics of AI lab of the University of Toronto, and the VECTOR(Toronto), MILA(Montreal), and CIFAR(Toronto) institutes, to mention just five examples.

Second, we will engage with interested parties from the mentioned in the previous paragraph, to prepare and deliver workshops of familiarization with the proposed model.

And third, we will publish a minimum of five peer-reviewed articles on: 1) the social context of ADMS discriminatory outcomes; 2) the analysis of the principles proposed by the International Framework for Principled AI; 3) the analysis of the variables of the proposed model environment; 4) the design of the capability and maturity model for trustworthy ADMS; and 5) an exploratory study of the feasibility of excluding of the algorithmic decision-making process the attributes that makes a person identifiable.

Lastly, the second stage of the thesis project's knowledge mobilization plan involves using the findings described in the thesis document and other currently open

questions derived from their discussion, to develop new research projects, based on the opportunities and gaps identified. CulturePlex Research Associates and Collaborators will use this thesis document as a discussion instrument for engagement with interested researchers and potential partners. The CulturePlex Director and the Lab's Research Network will join efforts to develop specific research proposals based on the report's results.

## 5.8 Conclusions of the Chapter

This chapter presents the design of a capability and maturity model for trustworthy ADM solutions that incorporates elements from the Principled AI International Framework, and engineering practices with proven success for the assurance of quality features in the software industry to help reducing the number of discriminatory outcomes from ADM solutions.

The model is supported by five guiding principles oriented to assure its applicability by following a philosophy of low interference with the project's chosen methodology and being respectful of the artifact's creation/verification efforts ratio.

The model adds a trustworthy facet to the project's quality assurance tasks through the inclusion of the ethical features like Transparency, Security, Project Governance, and Bias Management, via the implementation of 74 specific practices oriented at the data, algorithm, and engineering practice dimensions, as result of the analysis of the Principled AI International Framework's most relevant principles and the study of trustworthiness related variables.

The identified examples support our previous critiques regarding fairness by unawareness, which shows the inability of Rawls's "veil of ignorance" -criticized in chapter one-, to achieve fairness in a context where unfairness is exacerbated by ADM solutions, where unawareness will transform current discriminatory biases rather than eliminate them.

The study explored how features that are unexpectedly and non-obviously associated with sensitive attributes like race and gender can trigger biased and discriminatory outcome in the context of ADM solutions, demonstrating that our current social contract echoes the limitations of Rawls's, and helping visualize how those biases are amplified with ADM technologies.

A summarized simplified description of a theoretical implementation of the proposed model with regards to three project domains from real life was completed. The implementation highlights the changes that would have been made to the main causes for



discriminatory produced decisions that would have occurred without the use of the proposed model. Different capability and maturity levels are exhibited among the argued examples.

The implementation of the proposed model can complement current efforts in the pursue of trustworthy ADMs. This will hopefully lead to a more ordered and just society by using ADMs to reduce the gender and racial social gaps, including the discrimination experienced by the Hispanic community, which is the main goal of the thesis.

## 5.9 Works Cited in the Chapter

- Ajunwa, I. (2021). The auditing imperative for automated hiring. *Harv. J.L. & Tech.*. Retrieved from <https://ssrn.com/abstract=3437631> or <http://dx.doi.org/10.2139/ssrn.3437631>
- Algorithmic Bias Working Group. (2021). *IEEE P7003 - Algorithmic bias considerations*. IEEE Standard Association, IEEE Computer Society.
- Al-Saqqa, S., Sawalha, S., & AbdelNabi, H. (2020). Agile software development: Methodologies and trends. *International Journal of Interactive Mobile Technologies (IJIM)*, 14(11), 246-270. doi:<https://doi.org/10.3991/ijim.v14i11.13269>
- Autonomous Systems Validation Working Group P7001. (2021). *IEEE P7001 - IEEE draft standard for transparency of autonomous systems*. IEEE Standard Association, IEEE Computer Society.
- Aysolmaz, B., Dau, N., & Iren, D. (2020). Preventing algorithmic bias in the development of algorithmic decision-making systems: A delphy study. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.
- Balaji, S., Balamurugan, B., Kumar, A. T., Rajmohan, R., & Kumar, P. P. (2021). A brief survey on AI based face mask detection system for public places. *Irish Interdisciplinary Journal of Science & Research*, 5(1), 108-117.
- Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff*, 33, 1123–1131. doi:10.1377/hlthaff.2014.0041pmid:25006137
- Bornstein, S. (2019). Antidiscriminatory algorithms. *Alabama Law Review*, 70, 519-572.
- Carson, E. A. (2015). Prisoners in 2014. *Bureau of Justice Statistics*. Retrieved from: <http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5387>
- Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020). Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. *EC '20: Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 679-681. doi:<https://doi.org/10.1145/3391403.3399545>
- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* October 10th, 2018, withdrawn July 30th, 2021.
- Dressel, J. & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4(1) doi: 10.1126/sciadv.aao5580
- Eijo de Tezanos, P. M. A. (2016). Are Hispanics discriminated against in the US criminal justice system? *Graduate Research Posters. Poster 12*. <https://scholarscompass.vcu.edu/gradposters/12>
- Employer Data Governance Working Group. (2021a). *IEEE P7005 - IEEE draft standard for transparent employer data governance*. IEEE Standard Association, IEEE Computer Society.

- Employer Data Governance Working Group. (2021b). *IEEE P7004 - standard for child and student data governance*. IEEE Standard Association, IEEE Computer Society.
- Engineering Methodologies for Ethical Life-Cycle Concerns Working Group. (2021). *IEEE P7000 - IEEE draft model process for addressing ethical concerns during system design*. IEEE Standards Association, IEEE Computer Society.
- Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Steinhardt, J., & Madry, A. (2020). Identifying statistical bias in dataset replication. *Proceedings of the 37th International Conference on Machine Learning, PMLR*, 119, 2922-2932.
- Ericson, G., Rohm, W. A., Martens, J., Casey, C., Harvey, B., Gilley, S., & Poulton, K. J. (2017). What is the team data science process?
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center for Internet & Society*.
- Flores, A. W., Lowenkamp, C. T., & Bechtel, K. (n.d.). False positives, false negatives, and false analyses: A rejoinder to “machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” *Federal Probation*, 80(2).
- Frunal, B., Vrutik, S., & Gopalan, S. (2019). Machine learning: A software process reengineering in software development organization. *International Journal of Engineering and Advanced Technology*, 9(2), 4492-4500.
- Fu, S., Cutchin, S. M., Howell, K., & Ramachandran, S. (2020). Algorithm bias: Computer science student perceptions survey. *Proceedings of the 2020 ASEE PSW Section Conference*. Davis, California, USA.
- Fujii, G., Hamada, K., Ishikawa, F., Masuda, S., Matsuya, M., Myojin, T., Nishi, Y., Ogawa, H., Toku, T., Tokumoto, S., Tsuchiya, K., & Ujit, Y. (2020). Guidelines for quality assurance of machine learning-based artificial intelligence. *International Journal of Software Engineering and Knowledge Engineering*, 30(11-12), 1589-1606. doi:<https://doi.org/10.1142/S0218194020400227>
- Galli, T., Chiclana, F., & Siewe, F. (2020). Software product quality models, developments, trends, and evaluation. *SN Computer Science*, 1(154), 1-24. doi:<http://doi.org/10.1007/s42979-020-00140-z>
- Garg, S., Sinha, S., Kumar Kar, A., & Mani, M. (2021). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*. doi:<https://doi.org/10.1108/IJPPM-08-2020-0427>
- Guy, B., Farchi, E., Jayaraman, I., Raz, O., Tzoref-Brill, R., & Zalmanovici, M. (2019). Bridging the gap between ML solutions and their business requirements using features interactions. *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 1048-1058.

- Hileman, G. & Steele, S. (2016). Accuracy of claims-based risk scoring models. *Society of Actuaries*.
- Hofmann, M., & Tierney, B. (2009). Development of an enhanced generic data mining life cycle. *The ITB Journal*, 10(1), 50-71.
- Hughes, C., Robert, L., Frady, K., & Arroyos, A. (2019). Artificial intelligence, employee engagement, fairness, and job outcomes. In *Managing technology and middle- and low-skilled employees (The changing context of managing people)*. Emerald Publishing Limited.
- IBM. (2016). *Analytics solutions unified method - Implementations with agile principles*. IBM.
- Intelligent Process Automation. (2019). *IEEE 2755.1-2019 - IEEE guide for taxonomy for intelligent process automation product features and functionality*. IEEE Standards Association, IEEE Computer Society.
- Ipek, O. (2020). What is really different in engineering AI-enabled systems?. *IEEE Software*, 37(4), 3-6.
- ISO, IEC. (1991). *ISO/IEC 9126:1991 - Software engineering - Product quality*. International Standardization Organization (ISO), International Electronics Commission (IEC).
- ISO, IEC. (2001). *ISO/IEC 9126-1:2001 Software engineering - Product quality - Part 1: Quality model*. International Standardization Organization (ISO), International Electronics Commission (IEC).
- ISO, IEC. (2003). *ISO/IEC 9126-2:2003 Software engineering - Product quality - Part 2: External metrics*. International Standardization Organization (ISO), International Electronics Commission (IEC).
- ISO, IEC. (2003). *ISO/IEC 9126-3:2003 Software engineering - Product quality - Part 3: Internal metrics*. International Standardization Organization (ISO), International Electronics Commission (IEC).
- ISO, IEC. (2004). *ISO/IEC 9126-4:2004 Software engineering - Product quality - Part 4: Quality in use metrics*. International Standardization Organization (ISO), International Electronics Commission (IEC).
- ISO, IEC. (2010). *ISO/IEC 25010 Systems and software engineering - Systems and software quality requirements and evaluation (SQuaRE) - System and software quality models*. International Standardization Organization (ISO), International Electronics Commission (IEC).
- Jacobson, I., Booch, G., & Rumbaugh, J. E. (1999). *The unified software development process*. Addison-Wesley Longman Publishing Co. Inc.
- Jamie, D., Art, S., & Rajesh, A. (2020). Making agile more inclusive. *Software Quality Professional*, 23(1), 23-32.
- Kaestner, C. (2020). Machine learning is requirements engineering - On the role of bugs, verification, and validation in machine learning. *Medium Blog Post*.

- Kehl, D., Guo, P., & Kessler, S. (2017). *Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. responsive communities initiative*. Berkman Klein Center for Internet & Society, Harvard Law School.
- Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust In *AI. Conference on fairness, accountability, and transparency (FAccT '21)*, 262-271. doi:<https://doi.org/10.1145/3442188.3445890>
- Kumar, R., Maheshwary, P., & Malche, T. (2020). Novel software modeling technique for surveillance system. In R. Shukla, J. Agrawal, S. Sharma, N. Chaudhari, & K. Shukla (Eds.), *Social networking and computational intelligence. Lecture Notes in Networks and Systems*, Vol 100, 543-553. Springer. doi:<https://doi.org/10.1007/978-981-15-2071-643>
- Majumdar, D., & Chattopadhyay, H. K. (n.d.). AI and human rights: From business and policy perspectives. *International Journal of Business Marketing and Management*, 5(11), 51-60.
- Martínez-Plumed, F., Ferri, C., & Nieve, D. (2019). Fairness and missing values. *ArXiv*. doi:[arXiv:1905.12728v1](https://arxiv.org/abs/1905.12728v1)
- Mayer, R. C., Davis, J. H., & Schoorman, D. F. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ArXiv*. doi:[arXiv:1908.09635v2](https://arxiv.org/abs/1908.09635v2)
- Miguel, J. P., Mauricio, D., & Rodriguez, G. D. (2014). A review of software quality models for the evaluation of software products. *International Journal of Software Engineering & Applications*, 5(6), 31-54. doi:[10.5121/ijsea.2014.5603](https://doi.org/10.5121/ijsea.2014.5603)
- Muhammad, S., Qinghua, Z., Muhammad Azeem, A., Tahir, K., Faisal, M., & Muhammad Tanveer, R. (2020). Towards successful global software development. *Proceedings of the Evaluation and Assessment in Software Engineering EASE '20*, 445–450. Trondheim, Norway: Association for Computing Machinery. doi:[10.1145/3383219.3383283](https://doi.org/10.1145/3383219.3383283)
- Neoklis, P., Zinkevich, M., Roy, S., Breck, E., & Whang, S. (2019). Data validation for machine learning. *Proceedings of Machine Learning and Systems*, 334-347.
- Obermeyer, Ziad., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366 (6464), 447-453 doi: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342)
- Personal Data Privacy Working Group. (2021). *IEEE P7002 - IEEE draft standard for data privacy process*. IEEE Standard Association, IEEE Computer Society.
- Peldszus, S., Struber, D., & Jurjens, J. (2018). Model-based security analysis of feature-oriented software product lines. *ACM SIGPLAN Notices*, 53(9). doi:<https://doi.org/10.1145/3393934.3278126>
- Pressman, R. s. (2010). *Software engineering. A practitioner's approach (7th edition)*. McGrawHill Higher Education.

- Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14, 156 - 180. doi:10.1109/RBME.2020.3013489
- Rahimi, M., Guo, J. L., Sahar, K., & Chechik, M. (2019). toward requirements specification for machine-learned components. *IEEE 27th International Requirements Engineering Conference Workshops (REW)* 241-244. IEEE Computer Society.
- Re, C., Niu, F., Gudipati, P., & Srisuwananukorn, C. (2019). Overton: A data system for monitoring and improving machine-learned products. *ArXiv*. doi:arXiv:1909.05372
- Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., & Tonella, P. (2020). Testing machine learning based systems: A systematic mapping. *Empirical Software Engineering*, 1-62.
- Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, 4. doi:https://doi.org/10.1016/j.jrt.2020.100005
- Rodriguez-Martinez, L. C., Duran-Limon, H. A., & Mora, M. (2021). Service-oriented computing applications (SOCA) Development methodologies: A review of agility-rigor balance. In *Balancing agile and disciplined engineering and management approaches for IT services and software products*, 22. IGI Global. doi:10.4018/978-1-7998-4165-4.ch004
- Ryan, M., & Carsten Stahl, B. (2021). Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, 19(1), 61-86. doi:10.1108/JICES-12-2019-0138
- Saleema, A., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice*, 291-300.
- Sanders, K., Schyns, B., Dietz, G., & Den Hartog, D. N. (2006). Measuring trust inside organizations. *Personnel Review*.
- Santos, N. A., Ferreira, N., & Machado, R. J. (2021). AMPLA: An agile process for modeling logical architectures. In *Balancing agile and disciplined engineering and management approaches for IT services and software products*, 52-78. IGI Global. doi:10.4018/978-1-7998-4165-4.ch003
- Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 11(12), 1781-1794.
- Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13-22.

- Siwakorn, S., Wu, Z., Astorga, A., Alebiosu, O., & Xie, T. (2018). Multiple-implementation testing of supervised learning software. *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.
- Software Engineering Institute SEI. (2010). *CMMI for development version 1.3*. Carnegie Mellon University.
- Sommerville, I. (2015). *Software engineering*. Pearson Education.
- Subbaswamy, A., Schulam, P., & Saria, S. (2019). Preventing failures due to dataset shift: Learning predictive models that transport. *The 22nd International Conference on Artificial Intelligence and Statistics*, 3118-3127.
- Sun, W., Nasraoui, O., & Shafto, P. (2020). Evolution and impact of bias in human and machine learning algorithm interaction. *PLoS ONE*, 15(8), 1-39.  
doi:<https://doi.org/10.1371/journal.pone.0235502>
- The Standish Group. (2020). *CHAOS report 2020*. The Standish Group.
- Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Gonzalez Zelaya, C., & Van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 272-283.
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. ACM.  
doi:<https://doi.org/10.1145/3194770.3194776>
- Vogeli, C., Shields, A. E., Lee, T. A., Gibson, T. B., Marder, W. D., Weiss, K. B., & Blumenthal, D. (2007). Multiple chronic conditions: Prevalence, health consequences, and implications for quality, care management, and costs. *J. Gen. Intern. Med*, 22 (suppl. 3), 391–395 doi:10.1007/s11606-007-0322-1  
pmid:18026807
- Vogelsang, A., & Borg, M. (2019). Requirements engineering for machine learning: Perspectives from data scientists. *Proceedings of the 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*.
- Wadhwa, S., & Wadhwa, M. (2014). A comparative study of software quality models. *International Journal of Computer Science and Information Technologies*, 5(4), 5634-5638.
- Working Group for Life Cycle Processes. (2018). *IEEE/ISO/IEC 29148-2018 - ISO/IEC/IEEE international standard - Systems and software engineering - Life cycle processes - Requirements engineering*. IEEE Standards Association, IEEE Computer Society.
- Working Group for Life Cycle Processes. (2019a). *IEEE/ISO/IEC 15026-1\_Revision-2019 - ISO/IEC/IEEE international standard - Systems and software engineering - Systems and software assurance - Part 1: Concepts and vocabulary*. IEEE Standards Association, IEEE Computer Society.
- Working Group for Life Cycle Processes. (2019b). *IEEE/ISO/IEC 21839-2019 - ISO/IEC/IEEE international standard - Systems and software engineering -*

- System of systems (SoS) considerations in life cycle stages of a system.* IEEE Standards Association, IEEE Computer Society.
- Working Group for Life Cycle Processes. (2021). *IEEE/ISO/IEC P15288 - IEEE/ISO/IEC draft standard - Systems and software engineering - System life cycle processes.* IEEE Standards Association, IEEE Computer Society.
- Working Group for Personal Data AI Agent. (2021). *IEEE P7006 - Standard for personal data artificial intelligence (AI) agent.* IEEE Standard Association, IEEE Computer Society.
- Wickramasinghe, C. S., Marino, D. L., Grandio, J., & Manic, M. (2020). Trustworthy AI development guidelines for human system interaction. *13th International Conference on Human System Interaction.* IEEE Computer Society, HSI.
- Xiaowei, H., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., & Yi, X. (2020). A Survey of the safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defense, and interpretability. *Computer Science Reviews*, 37, 100-270.
- Yeounoh, C., Polyzotis, N., Tae, K., & Euijong Whang, S. (2019). Automated data slicing for model validation: A big data-AI integration approach. *IEEE Transactions on Knowledge and Data Engineering.*
- Yukun, Z., & Longsheng, Z. (2019). Fairness assessment for artificial intelligence in financial industry. *ArXiv*. doi:arXiv:1912.07211 [stat.ML]
- Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2020). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering.*



## Appendices

### Appendix A: List of documents included in the analyzed corpus to be referenced from chapter 2 as Principled AI International Framework.

ID Doc.	Year	Author	Document Title	Country	Author Type	Doc. Type
D01	2016	Partnership on AI	Tenets	USA	MultiStakeHolders	Commitment
D02	2016	U.S. National Science and Technology Council	Preparing for the Future of AI	USA	Government	Recommendations
D03	2017	UNI Global Union	Top 10 Principles for Ethical AI	Switzerland	Civil Society	Policy-Principles
D04	2017	Future of life	Asilomar AI Principles	USA	MultiStakeHolders	Principles
	2017	Tencent Institute	Six Principles of AI	China	Private Sector	Principles
D05	2017	ITI	AI Policy Principles	USA	Private Sector	Principles
D06	2017	The French Data Protection Authority (CNIL)	How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence	France	Government	Principles and Recommendations
D07	2018	Council of Europe: European Commission for the Efficiency of Justice CEPEJ	European Ethical Charter on the Use of AI in Judicial Systems and their environment	France	Inter-Governmental Organization	Policy-Usage
D08	2018	Amnesty International, AI Now	Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems	Canada	Civil Society	Policy-Principles
D09	2018	T20: Think20	Future of work and education for the digital age	Argentina	Civil Society	Policy-Principles
D10	2018	The public voice coalition	Universal Guidelines for AI	Belgium	Civil Society	Policy-Principles
D11	2018	Access Now	Human Rights in the age of AI	USA	Civil Society	Policy-Principles
D12	2018	University of Montreal	Montreal Declaration for responsible AI	Canada	MultiStakeHolders	Policy-Principles
D13	2018	Microsoft	Microsoft AI Principles	USA	Private Sector	Principles
D14	2018	Google	AI at Google: Our Principles	USA	Private Sector	Principles
D15	2018	Telefónica	AI Principles of Telefónica	Spain	Private Sector	Policy-Principles
D16	2018	Microsoft	Responsible bots: 10 guidelines for developers of conversational AI	USA	Private Sector	Guidelines Developers

D17	2018	Standards Administrations of China	White paper on AI Standardization	China	Government	Standardization Recommendations
D18	2018	Mission Assigned by the French Minister	For a Meaningful AI	France	Government	Considerations
D19	2018	UK House of Lords	AI in the UK	UK	Government	General Recommendations
D20	2018	Niti Aayog	National Strategy for AI	India	Government	Recommendations
D21	2018	British Embassy in Mexico City	Towards an AI Strategy in Mexico: Harnessing the AI Revolution	Mexico	Government	Recommendations
	2018	German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs	AI Strategy	Germany	Government	Considerations
D22	2018	Treasury Board of Canada Secretariat	Responsible Artificial Intelligence in the Government of Canada. Digital Disruption White Paper Series	Canada	Government	Recommendations
D23	2019	Organisation for Economic Co-operation and Development OECD	OECD Principles on AI	France	Inter-Governmental Organization	Policy-Principles
D24	2019	G20	G20 Principles on AI	Japan	Inter-Governmental Organization	Policy-Principles
D25	2019	IEEE Standard Association	Ethically Aligned Design	USA	MultiStakeHolders	Guidelines Developers
D26	2019	New York Times	Seeking Ground Rules for AI	USA	MultiStakeHolders	Principles
D27	2019	Beijing Academy of AI	Beijing AI Principles	China	MultiStakeHolders	Policy-Principles
	2019	AI Industry Alliance	AI Industry Code of Conduct	China	MultiStakeHolders	Policy-Principles
D28	2019	Telia Company	Guiding Principles on trusted AI Ethics	Sweden	Private Sector	Principles
	2019	IA Latam	Declaration of the Ethical Principles for AI	Chile	Private Sector	
D29	2019	IBM	IBM Everyday Ethics for AI	USA	Private Sector	Policy-Principles
D30	2019	Smart Dubai	AI Principles and Ethics	United Arab Emirates	Government	Principles
D31	2019	Monetary Authority of Singapore	Principles to promote FEAT AI in the Financial Sector	Singapore	Government	Principles
D32	2019	Government of Japan, Cabinet Office, Council for Science, Technology, and Innovation	Social Principles of Human-Centered AI	Japan	Government	Principles

D33	2019	European High-Level Expert Group on AI	Ethics Guidelines for Trustworthy-AI	Belgium	Government	Principles
D34	2019	Chinese National Governance Committee for AI	Governance Principles for a New Generation of AI	China	Government	Principles
D35	2019	IEEE Standard Association	IEC White Paper Artificial intelligence across industries	USA	MultiStakeHolders	Principles
D36	2020	Vatican	Rome Call for AI Ethics	Italy	Church	Principles
D37	2020	European Commission	AI for Europe	Belgium	Government	Action Plan

**Appendix B: List of principles (Summarized based on the page rank's importance score).**

PrincID	Community	PageRank Score	Principle Declaration	Principle Description
P06D01	2	0.0113	Work to maximize the benefits and address the potential challenges of AI technologies	:Working to protect the privacy and security a of individuals. Striving to understand and respect the interests of all parties that may be impacted by AI advances. Working to ensure that AI research and engineering communities remain socially responsible, sensitive, and engaged directly with the potential influences of AI technologies on wider society. Ensuring that AI research and technology is robust, reliable, trustworthy, and operates within secure constraints. Opposing development and use of AI technologies that would violate international conventions or human rights, and promoting safeguards and technologies that do no harm
P02D27	1	0.0110	For Humanity	:The R&D of AI should serve humanity and conform to human values as well as the overall interests of humankind. Human privacy, dignity, freedom, autonomy, and rights should be sufficiently respected. AI should not be used to against, utilize or harm human beings
P05D27	4	0.0107	Be Ethical	:AI R&D should take ethical design approaches to make the system trustworthy. This may include, but not limited to making the system as fair as possible, reducing possible discrimination and biases, improving its transparency, explainability, and predictability, and making the system more traceable, auditable, and accountable
P03D30	1	0.0104	HUMANITY	:AI should be beneficial to humans and aligned with human values, in both the long and short term; AI systems should be built to serve and inform, and not to deceive and manipulate. Nations should collaborate to avoid an arms race in lethal autonomous weapons, and such weapons should be tightly controlled. Active cooperation should be pursued to avoid corner-cutting on safety standards. Systems designed to inform significant decisions should do so impartially
P02D23	3	0.0090	Human-centered values and fairness	:a) AI actors should respect the rule of law, human rights, and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor and equality, diversity, fairness, social justice, and internationally recognized labor rights. b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art
P02D24	3	0.0090	Human-centered values and fairness	:a) AI actors should respect the rule of law, human rights, and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights. b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art

P06D28	2	0.0089	Safe and secure	:Our solutions are built and tested to prevent possible misuse and reduce the risk of being compromised or causing harm
P04D33	3	0.0089	The principle of respect for human autonomy	:The fundamental rights upon which the EU is founded are directed towards ensuring respect for the freedom and autonomy of human beings. Humans interacting with AI systems must be able to keep full and effective self-determination over themselves and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement, and empower human cognitive, social, and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight <sup>28</sup> over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work
P01D32	3	0.0088	The Human-Centric Principle	:The utilization of AI must not infringe upon the fundamental human rights guaranteed by the Constitution and international standards. AI should be developed, utilized, and implemented in society to expand the abilities of people and allow diverse people to pursue their own well-being. In a society making use of AI, it is desirable that we introduce appropriate mechanisms for literacy education and for the promotion of proper use of AI so that people do not become over-dependent on AI or misuse AI to manipulate other people's decision-making
P06D32	4	0.0085	The Principle of Fairness, Accountability, and Transparency	:The Principle of Fairness, Accountability, and Transparency: In an "AI-Ready Society", it is necessary to ensure fairness and transparency in decision-making, appropriate accountability for the results, and trust in the technology, so that people who use AI are not subject to undue discrimination with regard to personal background, or to unfair treatment in terms of human dignity
P04D15	2	0.0084	Privacy and security by design	:AI systems are fueled by data, and Telefonica is committed to respecting people's right to privacy and their personal data. The data used in AI systems can be personal or anonymous/aggregated. When processing personal data, according to Telefonica's privacy policy, we will at all times comply with the principles of lawfulness, fairness and transparency, data minimization, accuracy, storage limitation, integrity, and confidentiality. When using anonymized and/or aggregated data, we will use the principles set out in this document. In order to ensure compliance with our Privacy Policy we use a Privacy by Design methodology. When building AI systems, as with other systems, we follow Telefonica's Security by Design approach. We apply, according to Telefonica's privacy policy, in all the processing cycle phases, the technical and organizational measures required to guarantee a level of security adequate to the risk to which the personal information may be exposed and, in any case, in accordance with the security measures established in the law in force in each of the countries and/or regions in which we operate
P03D15	3	0.0081	Human-Centered AI	:AI should be at the service of society and generate tangible benefits for people. AI systems should always stay under human control and be driven by value-based considerations. Telefonica is conscious of the fact that the implementation of AI in our products and services should in no way lead to a negative impact on human rights or the achievement of the UN's Sustainable Development Goals. We are concerned about the potential use of AI for the creation or spreading of fake news, technology addiction and the potential reinforcement of societal bias in algorithms in general. We commit to working towards avoiding these tendencies to the extent it is within our realm of control
P04D23	2	0.0076	Robustness, security, and safety	:a) AI systems should be robust, secure, and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. c) AI actors should, based on their roles,

				the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias
P04D24	2	0.0076	Robustness, security, and safety	:a) AI systems should be robust, secure, and safe throughout their entire lifecycle so that, in 2 This Annex draws from the OECD principles and recommendations. conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. b) To this end, AI actors should ensure traceability, including in relation to datasets,
P04D07	4	0.0073	Principle of transparency, impartiality, and fairness	:Make data processing methods accessible and understandable, authorize external audits. a) A balance must be struck between the intellectual property of certain processing methods and the need for transparency (access to the design process), impartiality (absence of bias), fairness and intellectual integrity (prioritizing the interests of justice) when tools are used that may have legal consequences or may significantly affect people's lives. It should be made clear that these measures apply to the whole design and operating chain as the selection process and the quality and organization of data directly influence the learning phase. b) The first option is complete technical transparency (for example, open-source code and documentation), which is sometimes restricted by the protection of trade secrets. The system could also be explained in clear and familiar language (to describe how results are produced) by communicating, for example, the nature of the services offered, the tools that have been developed, performance and the risks of error. Independent authorities or experts could be tasked with certifying and auditing processing methods or providing advice beforehand. Public authorities could grant certification, to be regularly reviewed
P06D27	1	0.0072	Be Diverse and Inclusive	:The development of AI should reflect diversity and inclusiveness, and be designed to benefit as many people as possible, especially those who would otherwise be easily neglected or underrepresented in AI applications
P04D22	2	0.0072	Understanding the need to protect privacy and national security	:AI systems should be deployed in the most transparent manner possible
P03D19	2	0.0071	Access to, and control of, data	N/A
P06D36	2	0.0067	Security and privacy	:AI systems must work securely and respect the privacy of users
P01D30	2	0.0066	ETHICS	:AI systems should be fair; Data ingested should, where possible, be representative of the affected population, Algorithms should avoid non-operational bias. Steps should be taken to mitigate and disclose the biases inherent in datasets. Significant decisions should be provably fair. transparent; Developers should build systems whose failures can be traced and diagnosed. People should be told when significant decisions about them are being made by AI. Within the limits of privacy and the preservation of intellectual property, those who deploy AI systems should be transparent about the data and algorithms they use, accountable; Accountability for the outcomes of an AI system lies not with the system itself but is apportioned between those who design, develop, and deploy it. Developers should make efforts to mitigate the risks inherent in the systems they design. AI systems should have built-in appeals procedures whereby users can challenge significant decisions. AI systems should be developed by diverse teams which include experts in the area in which the system will be deployed; explainable, Decisions and methodologies of AI systems which have a significant effect on individuals should be explainable to them, to the extent permitted by available technology. It should be possible to ascertain the key factors leading to any specific decision that could have a significant effect on an individual. In the above situation we will provide channels through which people can request such explanations; and understandable

P04D27	2	0.0065	Control Risks	:Continuous efforts should be made to improve the maturity, robustness, reliability, and controllability of AI systems, so as to ensure the security for the data, the safety and security for the AI system itself, and the safety for the external environment where the AI system deploys
P05D12	4	0.0063	DEMOCRATIC PARTICIPATION PRINCIPLE	:AIS processes that make decisions affecting a person's life, quality of life, or reputation must be intelligible to their creators. The decisions made by AIS affecting a person's life, quality of life, or reputation should always be justifiable in a language that is understood by the people who use them or who are subjected to the consequences of their use. Justification consists in making transparent the most important factors and parameters shaping the decision and should take the same form as the justification we would demand of a human making the same kind of decision. The code for algorithms, whether public or private, must always be accessible to the relevant public authorities and stakeholders for verification and control purposes. The discovery of AIS operating errors, unexpected or undesirable effects, security breaches, and data leaks must imperatively be reported to the relevant public authorities, stakeholders, and those affected by the situation. In accordance with the transparency requirement for public decisions, the code for decision-making algorithms used by public authorities must be accessible to all, with the exception of algorithms that present a high risk of serious danger if misused. For public AIS that has a significant impact on the life of citizens, citizens should have the opportunity and skills to deliberate on the social parameters of these AIS, their objectives, and the limits of their use. We must at all times be able to verify that AIS are doing what they were programmed for and what they are used for. Any person using a service should know if a decision concerning them or affecting them was made by an AIS. Any user of a service employing chatbots should be able to easily identify whether they are interacting with an AIS or a real person. Artificial intelligence research should remain open and accessible to all
P18D02	2	0.0062	Educational organizations should include ethics, and related topics in security, privacy, and safety, as an integral part of curricula on AI, machine learning, computer science, and data science	N/A

## Curriculum Vitae

**Name:** Daniel Varona Cordero

**Post-secondary Education and Degrees:** University of Informatic Sciences  
Havana, Cuba  
2003-2008 B.Sc.

The University of Western Ontario  
London, Ontario, Canada  
2016-2021 Ph.D.

**Honors and Awards:** Provost Award, University of Informatic Sciences  
2008

Emerging Leader of Americas Program ELAP Research  
Scholarship (ELAP)  
Research Fellowship, Western University  
2011

Accelerate Internship Program grant of MITACS (PI: Juan Luis  
Suárez)  
Research Project, CulturePlex Laboratory  
2019

Voucher for Innovation and Productivity grant of the Ontario  
Centre of Excellence (OCE) (PI: Juan Luis Suárez)  
Research Project, CulturePlex Laboratory  
2020

**Related Work Experience** Teaching Assistant, Research Assistant  
CulturePlex Laboratory, The University of Western Ontario  
2016-2021

### Publications:

Suárez, Juan L. and Varona, Daniel. (2021). “The Ethical Skills We Are Not Teaching. An Evaluation of University Level Courses on Artificial Intelligence, Ethics, and Society”. A report to the Social Sciences and Humanities Research Council Knowledge Synthesis Grants Program. CulturePlex Lab, Western university, Canada.

Varona, Daniel and Suarez, Juan L. (2021). “Analysis of the Principled-AI Framework’s constraints in becoming a methodological reference for Trustworthy-AI Design” Handbook of Computational Social Science Volume 1 Theory, Case Studies and Ethics.

(Eds) Uwe Engel, Anabel Quan-Haase, Sunny Xun-Liu and Lars Lyberg. Francis and Taylor, UK.

Varona, Daniel; Lizama-Mue, Yadira; and Suarez, Juan L. (2020). Machine learning's limitations in avoiding automation of bias. *AI & Society*. <https://doi.org/10.1007/s00146-020-00996-y>.

Varona, Daniel. (2018). La responsabilidad ética del diseñador de sistemas en inteligencia artificial (Ethic responsibility of the system designer in artificial intelligence). *Revista de Occidente* (Madrid, Spain: 1923) Julio-Agosto (446-447).