UNIVERSITY
INSTITUTE
OF LISBON

Impact of Artificial Intelligence on Public Policy Automation: Benefits, Risks, and Pathways for Responsible Governance

André Monteiro Repolho dos Santos

Master in Computer Engineering

Supervisor:
PhD Luís Miguel Martins Nunes, Associate Professor with Aggregation,
Iscte – Instituto Universitário de Lisboa

Supervisor:
PhD Ana Maria Carvalho de Almeida, Associate Professor with Aggregation,
Iscte – Instituto Universitário de Lisboa

July, 2026

[This page is intentionally left blank.]

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Impact of Artificial Intelligence on Public Policy Automation: Benefits, Risks, and Pathways for Responsible Governance

André Monteiro Repolho dos Santos

Master in Computer Engineering

Supervisor:
PhD Luís Miguel Martins Nunes, Associate Professor with Aggregation,
Iscte – Instituto Universitário de Lisboa

Supervisor:
PhD Ana Maria Carvalho de Almeida, Associate Professor with Aggregation,
Iscte – Instituto Universitário de Lisboa

July, 2026

[This page is intentionally left blank.]

To my lovely girlfriend, Leonor,
for the unconditional love, unwavering support,
and encouragement that made this journey possible.
Thank you for always believing in me!

[This page is intentionally left blank.]

Acknowledgments

I would like to express my gratitude to my supervisors, Professor Luís Nunes and Professor Ana de Almeida. I am grateful not only for their embracing and believing in my proposed topic, showing enthusiasm from the very beginning, but also for their constant guidance and sharing of knowledge. This research process was initially unfamiliar to me and their working methodology and guidance were crucial in enabling me to develop this dissertation in a structured, rigorous, and calm manner.

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT), ISTAR-Iscte Pluriannual Projects UID/04466/2025 (`https://doi.org/10.54499/UID/04466/2025`).

A sincere thank you to my family and friends for their unconditional support in my daily life. I am deeply grateful to those who have always wished for my success and fought to provide me with the necessary conditions to reach this point. Your genuine interest in the development of this work and your constant concern for my well-being throughout this period were immensely important to me.

Finally, with very special emphasis, I reserve my deepest thanks for my girlfriend, Leonor. Thank you for being by my side every day, through the good times and the most challenging ones, and for never letting me break down or give up when fatigue spoke louder and that seemed like the easiest option. I appreciate all your companionship and patience, for choosing to always be present, and for the endless support you give me. You played a crucial role behind the scenes of this work and this achievement. I will always be grateful to you for making me the happiest man in the world, for helping me grow, and for motivating me daily to be the best version of myself. Without you, none of this would have been possible.

[This page is intentionally left blank.]

Resumo

A integração da Inteligência Artificial na Administração Pública Portuguesa cria um impasse técnico-legal, porque os sistemas modernos de *machine learning* são estruturalmente incapazes de cumprir o dever de fundamentação (*iter cognoscitivo*) exigido pelos artigos 152.º e 153.º do Código do Procedimento Administrativo. Apesar da entrada em vigor do *EU AI Act*, persiste um vazio operacional crítico, deixando as entidades públicas sem ferramentas para traduzir a conformidade abstrata em fluxos de trabalho de engenharia.

Esta dissertação aborda essa lacuna através de uma metodologia de *Design Science Research*, produzindo um artefacto de governação, o *Portuguese Algorithmic Impact Assessment* (AIA-PT), fundamentado numa Revisão Sistemática da Literatura (PRISMA) de 107 estudos revistos por pares e 18 documentos de literatura cinzenta.

O AIA-PT estrutura-se em três pilares: classificação de risco com escalonamento processual (Pilar I), requisitos de engenharia e monitorização do ciclo de vida da IA (Pilar II), e transparência algorítmica e ética desde a conceção (Pilar III).

A *framework* é avaliada através de demonstrações *ex ante* em quatro domínios: triagem preditiva (Saúde), reincidência criminal (Justiça), evasão fiscal (Finanças) e *chatbots* municipais (Administração Local). A avaliação revela que o AIA-PT atua como uma rede de segurança regulatória vital, captando funções soberanas de alto risco, como a deteção de fraude financeira pela Autoridade Tributária, isenta do escrutínio de alto risco ao abrigo do Anexo III do *EU AI Act*. Em última análise, o AIA-PT equipa os gestores públicos portugueses com as ferramentas operacionais para implementar uma IA eficaz, constitucionalmente conforme e democraticamente escrutinável.

Palavras-Chave: Inteligência Artificial; Administração Pública; Governação Algorítmica; EU AI Act; Avaliação de Impacto Algorítmico; Design Science Research.

[This page is intentionally left blank.]

Abstract

The integration of Artificial Intelligence (AI) into the Portuguese Public Administration creates a legal-technical deadlock, because modern Machine Learning (ML) systems are structurally incapable of fulfilling the duty to state reasons (*iter cognoscitivo*) mandated by Articles 152 and 153 of the Administrative Procedure Code (*Código do Procedimento Administrativo*). Despite the entry into force of the EU AI Act, a critical operational void remains, leaving public entities without tools to translate abstract compliance into daily engineering workflows.

This dissertation addresses that gap through a Design Science Research methodology, producing a governance artifact, the Portuguese Algorithmic Impact Assessment (AIA-PT), grounded in a Systematic Literature Review (PRISMA) of 107 peer-reviewed studies and 18 grey literature sources.

The AIA-PT is structured around three pillars: risk classification with procedural escalation (Pillar I), engineering requirements and continuous AI lifecycle monitoring (Pillar II), and algorithmic transparency and ethics by design (Pillar III).

The framework is evaluated through *ex ante* scenario-based demonstration across four domains: predictive triage (Health), recidivism risk assessment (Justice), tax evasion detection (Finance), and a municipal FAQ chatbot (Local Administration). The evaluation reveals that the AIA-PT acts as a vital regulatory safety net, successfully capturing high-stakes sovereign functions, such as financial fraud detection by the Tax Authority, that are explicitly exempted from high-risk scrutiny under Annex III of the EU AI Act. Ultimately, the AIA-PT equips Portuguese public managers with the operational tools to deploy AI that is effective, constitutionally compliant, and democratically accountable.

KEYWORDS: Artificial Intelligence; Public Administration; Algorithmic Governance; EU AI Act; Algorithmic Impact Assessment; Design Science Research.

[This page is intentionally left blank.]

Contents

Acknowledgments i

Resumo iii

Abstract v

List of Figures ix

List of Tables xi

List of Acronyms xiii

Chapter 1. Introduction 1
1. Contextualization 1
2. Motivation and Problem Statement 2
3. Objectives and Research Questions 2
4. Expected Contributions 3
5. Methodology Overview 4
6. Dissertation Structure 4

Chapter 2. State of the Art 7
1. Systematic Review 7
2. AI in Public Sector: Foundations 9
3. AI in Public Health 12
4. AI in Justice 15
5. AI in Public Finance 18
6. Cross-Domain Risks 22
7. Portuguese Context & Conclusion 24

Chapter 3. Comparative Framework Analysis 27
1. Introduction 27
2. The "Hard Law": Impact of the EU AI Act on Public Administration 28
3. The "Soft Law": The Best Practice Benchmarks 31
4. National Diagnosis: The Operational Gap in Portugal 46

Chapter 4. Methodology for Guidelines Development 49
1. Research Approach: Design Science Research 49
2. Phases 1 & 2: Problem Identification and Objectives Definition 49
3. Phase 3: Design and Development 50

4. Phases 4 & 5: Demonstration and Evaluation 51
5. Phase 6: Communication 52

Chapter 5. Guidelines for Portuguese Public Administration 53
1. Introduction and Architecture of the Guidelines 53
2. Pillar I: The Portuguese Algorithmic Impact Assessment (AIA-PT) 54
3. Pillar II: Engineering Requirements and Continuous Test, Evaluation, Verification, and Validation (TEVV) 59
4. Pillar III: Team Ethics and Proactive Transparency 62

Chapter 6. Discussion 67
1. From Theory to Practice 67
2. Demonstration: Multi-Scenario Application 67
3. Evaluation 82

Chapter 7. Conclusion 87
1. Summary of Contributions 87
2. Theoretical and Practical Implications 89
3. Limitations 90
4. Directions for Future Work 90
5. Closing Remarks 91

References 93

List of Figures

1.1 The Design Science Research (DSR) Process Model (from Peffers et al. [5]). 4

2.1 PRISMA Flow Diagram of the literature selection process. 9

3.1 The 5 Dimensions of the OECD Classification Framework (Source: OECD [55]). 33

3.2 Application of an AI system "in the lab" or "in the field" (Source: OECD [55]). 34

3.3 Classification of algorithmic transparency instruments (Source: Global Partnership on Artificial Intelligence (GPAI) [56]). 35

3.4 Geographical distribution of active algorithmic registries (Source: Global Partnership on Artificial Intelligence (GPAI) [56]). 35

3.5 Example of Meaningful Transparency: The UK Algorithmic Transparency Recording Standard (from gov.uk/algorithmic-transparency-records). 36

3.6 First Action Scorecard (Source: UK Government Digital Service [58]). 39

3.7 Example of the Self-Assessment Scorecard (Source: UK Government Digital Service [58]). 39

3.8 Characteristics of trustworthy AI systems (Source: US National Institute of Standards and Technology [59]). 42

3.9 The Core Functions of the AI RMF (Source: US National Institute of Standards and Technology [59]). 43

5.1 High-level architecture of the proposed Guidelines 53

5.2 The AIA-PT workflow: Taxonomic Mapping, Risk Classification, and Mitigation Requirements. 54

5.3 Shifting from a static software procurement model to a continuous AI TEVV lifecycle. 60

[This page is intentionally left blank.]

List of Tables

2.1 Refined Search Strings applied in Scopus 7
2.2 Eligibility Criteria applied in the SLR 8
3.1 Gap Analysis: International Benchmarks versus Portuguese Strategy 48
5.1 AIA-PT Risk Scoring Rubric 56
5.2 AIA-PT Procedural Escalation Matrix by Impact Level 59
5.3 AI Public Procurement Mandatory Requirements Checklist 61
5.4 Structure of the National Algorithmic Transparency Registry (Two-Tier Model) 64
6.1 Summary of the AIA-PT Multi-Scenario Demonstration 82

[This page is intentionally left blank.]

List of Acronyms

ADM	Automated Decision-Making
AI	Artificial Intelligence
AIA	Algorithmic Impact Assessment
ALTAI	Assessment List for Trustworthy Artificial Intelligence
ANN	Artificial Neural Network
AT	Autoridade Tributária
ATRS	Algorithmic Transparency Recording Standard
AUROC	Area Under the Receiver Operating Characteristic
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CPA	Portuguese Code of Administrative Procedure
DESI	Digital Economy and Society Index
DGRSP	Direção Geral de Reinserção e Serviços Prisionais
DL	Deep Learning
DSR	Design Science Research
ED	Emergency Department
EDN	Estratégia Digital Nacional
ESI	Emergency Severity Index
EU	European Union
FRIA	Fundamental Rights Impact Assessment
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GOFAIL	Good Old Fashioned AI and Law
GPAI	Global Partnership on Artificial Intelligence
HART	Harm Assessment Risk Tool
HUNOD	Hybrid Unsupervised Outlier Detection
IT	Information Technology
LADA	Lei de Acesso aos Documentos Administrativos
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
ML	Machine Learning
MTS	Manchester Triage System
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OCR	Optical Character Recognition

OECD	Organisation for Economic Co-operation and Development
OR	Operating Room
PA	Public Administration
PACU	Post-Anesthesia Care Unit
PET	Privacy-enhancing technology
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRR	Recovery and Resilience Plan
RAG	Retrieval-Augmented Generation
RCN	Recidivism Clustering Network
RMF	Risk Management Framework
RSE	Registo de Saúde Eletrónico
SHAP	Shapley Additive Explanations
SLR	Systematic Literature Review
SMOTE	Synthetic Minority Over-sampling Technique
SNA	Social Network Analysis
SNS	Serviço Nacional de Saúde
SOFA	Sequential Organ Failure Assessment
STJ	Supreme Court of Justice
SVM	Support Vector Machine
SyRI	System Risk Indication
TEVV	Test, Evaluation, Verification, and Validation
XAI	Explainable AI

CHAPTER 1

Introduction

The digital transformation of Public Administration (PA) has evolved from mere digitization of paper processes to the implementation of intelligent systems capable of autonomous decision-making. Artificial Intelligence (AI) is currently reshaping how governments interact with citizens, promising significant improvements in efficiency, consistency, and speed in service delivery. However, the deployment of AI in high-stakes domains such as Public Health, Justice, and Public Finance introduces complex challenges regarding transparency, accountability, and fairness.

This chapter establishes the theoretical and practical foundations of this dissertation. It begins by contextualizing the rise of algorithmic governance using recent academic literature, leading to the identification of the critical problem: the gap between strategic ambition and operational reality. From this problem statement, the research objectives are derived, which in turn shape the specific research questions that guide this study.

1. Contextualization

Globally, the adoption of AI in public services is accelerating, driven by the need for more efficient resource management and responsive governance. Recent academic literature highlights that governments are increasingly moving beyond simple automation to the deployment of Machine Learning (ML) solutions for complex decision-making. In the healthcare sector, systematic reviews indicate that AI is already revolutionizing emergency management, while in public finance, algorithms are improving budgeting accuracy and fraud detection. However, this transition is not merely technical but deeply sociopolitical. In the justice sector, scholars like Završnik [1] warn that the shift towards "predictive justice" and policing tools entails a fundamental change in how the state exercises power, carrying profound risks for human rights and discrimination.

While international frameworks, such as the Organisation for Economic Co-operation and Development (OECD) AI Principles and the recently approved European Union (EU) AI Act (2024), highlight the urgency of regulating these technologies, the operational reality often lags behind. In Portugal, the commitment to digital acceleration is evident in strategic documents like the *Estratégia Digital Nacional 2030* and *AI Portugal 2030*.

According to the Digital Economy and Society Index (DESI) 2025 report [2], the country excels in digital public services, scoring 84.5 for citizens (above the EU average of 82.3) and achieving 98.7% in 5G coverage. This success is driven by the Recovery and Resilience Plan (PRR), which allocates €4.5 billion to the digital transition (21% of the total allocation) [2]. Yet, the operational adoption of AI lags significantly behind this

infrastructural maturity. European Commission [2] reveals that only 8.6% of Portuguese enterprises utilized AI technologies in 2024, a significant gap relative to the European target of 75% by 2030. This gap between strategic ambition and operational reality creates a critical challenge according to Magalhães, Costa, and Matos [3]. The Portuguese Code of Administrative Procedure (CPA) mandates a strict duty to provide reasons for administrative decisions (Articles 152 and 153). This legal requirement clashes with the opacity of "black box" algorithms, creating a tension between the efficiency promised by *AI Portugal 2030* and the transparency required by law.

As the country moves towards algorithmic governance, bridging this gap requires more than technology adoption, it demands specific operational guidelines to ensure that the "third actor" in administrative relations, the algorithm, remains accountable and compliant.

2. Motivation and Problem Statement

The motivation for this research stems from a dual observation regarding the current state of technology in the public sector. On one hand, the literature documents substantial benefits of AI automation, such as increased diagnostic precision and operational efficiency. On the other hand, a critical "Accountability Gap" emerges. As identified by Crawford and Schultz [4], when decision-making power is delegated to "black box" algorithms, often developed by private third parties, the state risks abdicating its sovereign responsibility, leaving public institutions unable to explain or justify automated decisions to citizens. This tension is exacerbated by the risk of social exclusion. Evidence from welfare automation systems, such as the Universal Credit in the UK, demonstrates that rigid digital conditionality can disproportionately harm vulnerable citizens, eroding the concept of social citizenship.

Despite the strategic enthusiasm in Portugal, there is a noticeable lack of operational guidance for the actual implementation of these systems. Public managers and software engineers currently lack a localized framework that translates high-level ethical principles into technical and procedural requirements. Consequently, the central problem addressed by this dissertation is the absence of actionable guidelines for the responsible deployment of AI in the Portuguese Public Administration, creating risks of algorithmic bias, lack of transparency, and legal non-compliance in high-risk sectors.

3. Objectives and Research Questions

Deriving directly from the identified problem, the gap between high-level strategy and operational reality, the main objective of this dissertation is to develop practical and operational guidelines for the responsible implementation of AI in the Portuguese Public Administration, ensuring alignment with European regulations (EU AI Act) and national strategic goals.

To achieve this, specific objectives were defined to structure the research process:

i. **Analyze the State of the Art:** Map current AI applications, benefits, and reported risks in Public Health, Justice, and Finance through a systematic literature review.

ii. **Evaluate Regulatory Frameworks:** Critically analyze international frameworks (EU, OECD, National Institute of Standards and Technology (NIST), the Government of Canada) to identify best practices adaptable to the Portuguese reality.

iii. **Identify Implementation Gaps:** Diagnose the specific regulatory and technical gaps within the current Portuguese institutional context.

iv. **Propose Operational Guidelines:** Design a set of actionable technical, ethical, and organizational recommendations, specifically tailored for Portuguese public entities.

To fulfill these objectives and address the core problem of the "Accountability Gap," the investigation is guided by the following main research question: *"How can the Portuguese Public Administration implement AI systems in high-stakes domains while ensuring accountability, fairness, and compliance with European regulations?"*

This main inquiry is decomposed into three specific research questions:

- **RQ1:** What are the primary uses of AI in public policies regarding Justice, Health, and Finance, and what are the reported benefits and risks associated with these applications? (Addressing Objective i.)
- **RQ2:** Which guidelines, mechanisms, and best practices for responsible governance are currently proposed in the literature and which are transferable to Portugal? (Addressing Objective ii.)
- **RQ3:** How can these international frameworks be adapted to create an operational roadmap for the Portuguese context? (Addressing Objectives iii. and iv.)

4. Expected Contributions

By answering these questions, this dissertation aims to contribute to the field of Computer Science and Information Systems applied to the Public Sector by providing a Systematic Literature Review (SLR) following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)[1] protocol that synthesizes research at the intersection of AI and Public Policy.

Furthermore, it not only offers a comparative analysis of major international AI governance frameworks, but also presents a set of practical guidelines specifically adapted to the Portuguese legal and technical reality, thus providing an operational tool for public decision-makers and Information Technology (IT) engineers.

[1]Available at: `https://www.prisma-statement.org/`

5. Methodology Overview

This research follows a Design Science Research (DSR) methodology, visually represented in Figure 1.1. The methodology is structured in two main phases to ensure both theoretical rigor and practical relevance.

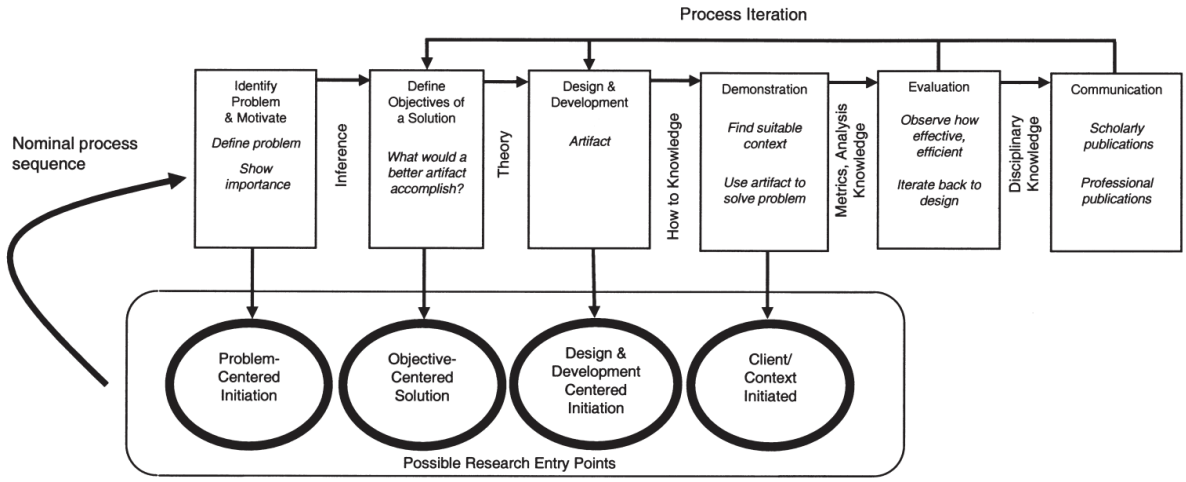


FIGURE 1.1. The DSR Process Model (from Peffers et al. [5]).

The first phase consists of an SLR following the PRISMA protocol, analyzing 107 academic articles and 18 grey literature documents to establish the theoretical foundation. The second phase involves a Comparative Framework Analysis and the synthesis of guidelines based on the identified gaps and best international practices.

A detailed description of the methodology will be presented in Chapter 4 (Methodology).

6. Dissertation Structure

The remainder of this document will be organized into six chapters that guide the transition from theoretical analysis to practical application. Chapter 2 State of the Art, establishes the academic foundation through a systematic literature review, mapping AI applications in Health, Justice, and Finance, while identifying cross-domain risks. This will be complemented by Chapter 3 Comparative Framework Analysis, which critically examines international regulatory frameworks, such as the EU AI Act and the Canadian Directive, to inform the normative basis of this study.

The process for synthesizing these findings into actionable artifacts will be detailed in Chapter 4 Methodology, which will outline the methodological approach adopted for the development of the proposed guidelines. The core contribution of this dissertation will be presented in Chapter 5 Guidelines for Portuguese PA, which will present the operational and technical guidelines specifically tailored for the Portuguese Public Administration. The feasibility, implications, and limitations of these recommendations will be critically examined in Chapter 6 Discussion.

Finally, Chapter 7 Conclusion will synthesize the main findings and suggest future research directions to further advance responsible algorithmic governance in the public sector.

[This page is intentionally left blank.]

CHAPTER 2

State of the Art

1. Systematic Review

To ensure a rigorous and replicable analysis of the state of the art, this dissertation followed an SLR methodology based on the PRISMA statement guidelines. Given the intersection of this research with public policy and regulation, the review adopted a hybrid approach: a systematic search of academic databases complemented by a targeted selection of "grey literature" (official government reports and regulations), a strategy essential for capturing both the technical and legal dimensions of AI in the Public Sector.

The systematic search was primarily conducted on Scopus, chosen for its extensive coverage of interdisciplinary fields including Computer Science, Social Sciences, and Decision Sciences, and PubMed, to ensure adequate coverage of the Health domain. An initial exploratory search was performed in September 2025 to identify main keywords. Based on preliminary results, the search strings were refined in October 2025 to increase precision, utilising proximity operators (e.g., `W/3`) to specifically link AI technologies to public policy contexts such as "Public Health Policy" or "Judicial Decision". This refinement was crucial to exclude purely technical algorithm descriptions lacking administrative application. The final search queries applied to each domain (Health, Justice, and Finance) are detailed in Table 2.1.

TABLE 2.1. Refined Search Strings applied in Scopus

Domain	**Query String**
Health	(("artificial intelligence" OR "machine learning" OR "algorithmic decision-making" OR "predictive analytics" OR "decision support system") W/3 ("public health policy" OR "health policy" OR "resource allocation" OR "triage" OR "hospital management" OR "healthcare management")) AND ("efficiency" OR "accuracy" OR "workflow optimization" OR "resource optimization") AND ("public sector" OR "public administration" OR "public policy" OR "public policies" OR "governance" OR "government*" OR "public health system*" OR "public healthcare" OR "health policy" OR "policy-making" OR "policy making" OR "national health" OR "health governance" OR "public hospital*" OR "state hospital*")

Domain	Query String
Justice	(("artificial intelligence" OR "machine learning" OR "algorithmic decision-making" OR "predictive analytics" OR "decision support system") W/3 ("judicial decision" OR "justice system" OR "predictive policing" OR "criminal justice" OR "court" OR "case management")) AND ("efficiency" OR "accuracy" OR "risk assessment" OR "workflow optimization" OR "decision support") AND ("public sector" OR "public administration" OR "public policy" OR "public policies" OR "governance" OR "government*" OR "criminal justice policy" OR "justice policy" OR "policy-making" OR "policy making" OR "law enforcement" OR "public safety" OR "justice reform" OR "judicial reform" OR "state court*" OR "public prosecution")
Finance	(("artificial intelligence" OR "machine learning" OR "algorithmic decision-making" OR "predictive analytics" OR "decision support system") W/3 ("fraud detection" OR "tax evasion" OR "public finance" OR "social scoring" OR "government auditing" OR "resource allocation")) AND ("efficiency" OR "accuracy" OR "cost reduction" OR "workflow optimization" OR "resource optimization") AND ("public sector" OR "public administration" OR "public policy" OR "public policies" OR "governance" OR "government*" OR "fiscal policy" OR "tax administration" OR "public finance" OR "public financial management" OR "tax authority" OR "tax authorities" OR "revenue service*" OR "public revenue" OR "government budget*" OR "public expenditure" OR "welfare policy" OR "social welfare policy")

To ensure the quality and relevance of the included studies, strict eligibility criteria were defined. The review focused on the timeframe between 2018 and 2025, capturing the impact of the latest AI advancements and recent regulatory shifts. Only peer-reviewed articles and reviews written in English or Portuguese were considered. Studies focusing exclusively on clinical diagnosis without administrative implications or private sector applications were excluded. A summary of these inclusion and exclusion criteria is presented in Table 2.2.

TABLE 2.2. Eligibility Criteria applied in the SLR

Criterion	Inclusion	Exclusion
Timeframe	Published between 2018 and 2025 (PUBYEAR > 2017)	Published before 2018
Language	English and Portuguese	Other languages
Document Type	Peer-reviewed Articles and Reviews	Conference papers (if not peer-reviewed), editorials, short notes
Subject	Application of AI in Health, Justice, or Public Finance with focus on policy, governance, or efficiency.	Purely technical papers without public sector application; Clinical studies without management/policy focus.

The selection process is visually summarized in the PRISMA flow diagram presented in Figure 2.1. The database searches yielded a total of 1,046 results. To build a theoretically sufficient and manageable corpus for each domain, results were ranked by relevance and citation count, with the highest-ranked records retained for full-text screening. Records that did not rank sufficiently in either dimension were excluded before screening ($n = 937$). The remaining 109 records were screened by title and abstract, and two clearly irrelevant articles were subsequently excluded, yielding a final corpus of 107 peer-reviewed articles.

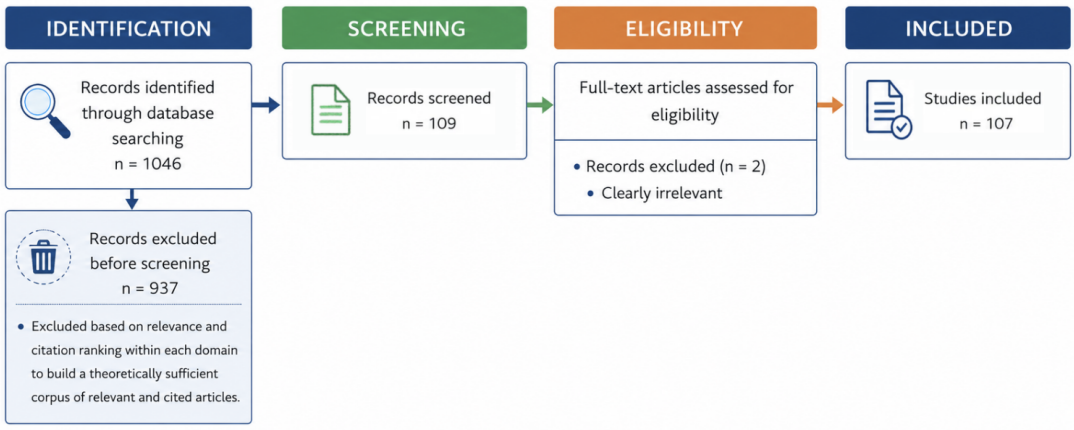


FIGURE 2.1. PRISMA Flow Diagram of the literature selection process.

Complementing the database search, grey literature was identified through targeted web searches of official institutional websites, focusing on national Portuguese legislation and policy documents, European Union regulatory instruments and documentation from countries with administrative traditions similar to the Portuguese. No formal database was used in this component. The selection was dictated by direct relevance to the research domains.

2. AI in Public Sector: Foundations

The transition described in the introductory chapter, from traditional bureaucratic administration to algorithmic governance represents more than a technological update, as it constitutes a paradigm shift in the exercise of state power. Unlike the previous wave of digitization, which focused on the dematerialization of processes, the integration of AI introduces the "third actor" into the administrative relationship [3]. This chapter deconstructs the theoretical and technical foundations of this shift, analyzing how predictive architectures are reshaping the ontology of administrative acts, transforming them from human-centered deliberations into probabilistic outcomes based on data correlations.

2.1. From Digitization to Algorithmic Intelligence

While the initial phase of e-government focused on the cataloguing and transaction of digital forms [6], the current wave of technological adoption is characterized by systems that not only process data but also emulate intelligent behavior. In the legal and administrative domain, this distinction is often framed as the evolution from Good Old

Fashioned AI and Law (GOFAIL) to modern ML. GOFAIL systems typically relied on a close partnership between legal experts and computer specialists to obtain knowledge and represent it in an executable form. In contrast, the modern AI systems currently being deployed in high-stakes domains operate inductively: they learn patterns and rules from vast datasets rather than being explicitly programmed with them [7].

This shift is technically supported by the adoption of ML and Deep Learning (DL). ML allows systems to "learn" from data by recognizing patterns and making predictions without rigid instructions [3]. Within this subset, DL utilizes Artificial Neural Networks (ANNs) with multiple layers to process complex data structures, relying on architectures inspired by biological neural networks yet fundamentally distinct in their operational logic [8]. The application of these technologies in the public sector creates what Magalhães, Costa, and Matos [3] describe as a "third actor" in legal-administrative relations, altering the traditional dyad between the Administration and the citizen.

However, the suitability of this "third actor" is strictly dependent on the nature of the administrative function in question. The literature indicates that AI demonstrates high efficacy and legal safety in "bound acts", where the law strictly defines the premises and consequences of a decision, allowing for automation based on clear, pre-defined logic [3]. Conversely, in "discretionary acts" involving indeterminate legal concepts (e.g., "grave negligence" or "public interest"), the substitution of human judgment by an algorithm is legally perilous. Current AI systems lack the sensitivity to perform the value judgments and weighting of interests required by the principle of administrative discretion [3]. Therefore, the foundational deployment of AI in the public sector must be understood not merely as a technical upgrade, but as a bounded integration subject to the specific legal nature of the administrative act.

2.2. Technological Architectures in Governance

The operational reality of algorithmic governance relies on specific algorithmic architectures that determine how decisions are computed. The literature identifies several dominant techniques applied across Justice, Health, and Finance:

2.2.1. *Supervised Learning*

This is the most prevalent approach in administrative applications, where algorithms are trained on labeled historical data to predict future outcomes. For example, in the justice sector, recidivism prediction tools like Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) or the Recidivism Clustering Network (RCN) utilize historical offender profiles to forecast re-offending risks [9], [10]. Similarly, in healthcare triage, supervised models like Support Vector Machines (SVMs), Random Forest, and Gradient Boosting are trained on past patient records to predict hospital admission needs and acuity levels [11], [12]. In finance, algorithms such as Logistic Regression and Random Forest are extensively used to classify credit card transactions as fraudulent or legitimate based on past labeled examples [8], [13].

2.2.2. *Unsupervised Learning*

In domains where labeled data is scarce or the objective is anomaly detection, unsupervised learning is employed [14]. This is particularly relevant in financial oversight, such as tax evasion detection, where clustering algorithms (e.g., K-means) group taxpayers based on behavioral patterns to identify outliers indicative of fraud without prior labeling of evasion cases [15].

2.2.3. *Natural Language Processing (NLP)*

With the public sector being heavily document-centric, NLP has become foundational. It is used to analyze unstructured text in electronic health records for triage [11], [16] and to power conversational agents (chatbots) in healthcare services [17].

However, the reliance on these architectures introduces the "Black Box" phenomenon. Especially in DL models, the internal logic connecting inputs to outputs is often unintelligible even to the developers, creating a barrier to human understanding and, consequently, to administrative transparency [1], [14]. This opacity challenges the democratic requirement that administrative decisions must be explainable and justifiable to the affected citizens [3], [12].

2.3. The Transformation of Administrative Discretion

The integration of AI fundamentally alters the concept of administrative discretion. Traditionally, discretion involves a human official weighing complex factors to reach a decision within the bounds of the law. In the era of algorithmic governance, discretion is increasingly mediated or replaced by statistical probability.

Scholars note that while AI offers consistency, eliminating the "noise" and variability of human judgment caused by fatigue or mood, it lacks the capacity for semantic interpretation and causal understanding required in law [18]. An algorithm can identify a correlation between a variable and an outcome, but it cannot understand the context or the social causality of that relationship [7], [18]. For instance, in judicial decision-making, while AI can process massive amounts of data to predict outcomes, it struggles with value judgments and the "warmth" of human empathy required for equitable justice [18].

This limitation leads to a tension between "automation" and "augmentation." In some contexts, AI is positioned as a tool to streamline efficiency by handling repetitive tasks, theoretically freeing human judges for complex reasoning [18], [19]. However, there is a documented risk of "automation bias," where human decision-makers, under time pressure or lacking technical literacy, defer excessively to algorithmic recommendations, effectively ceding their discretionary power to the machine [19], [20]. This phenomenon creates a de facto automated decision-making system, even when a "human-in-the-loop" is theoretically present to validate the outcome [21]. Evidence suggests that humans may struggle to override algorithmic advice, especially when it confirms pre-existing stereotypes [19].

2.4. Data as the Foundation of Policy

Finally, the foundation of any AI system in the public sector is data. The literature emphasizes that algorithmic outputs are only as reliable as the data upon which they are trained, a concept often summarized as "garbage in, garbage out" [9]. In public policy, this is critical because historical data is rarely neutral; it reflects the structural inequalities and biases of past administrative practices [14], [22].

For instance, predictive policing algorithms trained on historical arrest data do not predict crime per se, but rather predict police activity, often reinforcing feedback loops of over-policing in specific communities [9]. Similarly, in recidivism prediction, the use of historical data can embed racial biases, leading to higher false positive rates for minority groups [9], [23]. In the health sector, data bias can result in algorithms that perform poorly for underrepresented demographic groups, exacerbating health disparities [14], [24].

Consequently, the "foundations" of AI in the public sector are not merely technical but deeply socio-political. The deployment of these systems implies a shift from a logic of causality and individualized justice to a logic of correlation and actuarial risk management [25]. This shift establishes the baseline for the specific benefits and risks that will be analyzed in the subsequent sections of this chapter.

3. AI in Public Health

The application of AI in the health sector represents one of the most mature domains of algorithmic governance. Unlike traditional administrative digitization, AI in health involves high-stakes decision-making where algorithmic outputs directly impact patient mortality, resource allocation, and civil liberties.

The literature reveals a distinct dichotomy, while quantitative studies consistently demonstrate the superior predictive performance of ML models over human judgment and traditional scales, qualitative and systematic reviews highlight profound barriers regarding generalization, ethical risks, and the "implementation gap" between theoretical performance and clinical reality.

3.1. Clinical Triage and Emergency Medicine

Emergency Departments (EDs) are characterized by overcrowding and the need for rapid, accurate decision-making, making them a primary target for AI automation [26], [27]. Traditional triage systems, such as the Emergency Severity Index (ESI) or the Manchester Triage System (MTS), rely heavily on subjective human judgment, leading to variability and potential error [12].

Recent systematic reviews and meta-analyses provide robust evidence of AI's technical superiority in this domain. Adebayo, Bhuiyan, and Ahmed [28] analyzed 14 studies, finding that AI/ML models significantly outperformed conventional triage tools. Specifically, for mortality prediction, AI models showed a mean Area Under the Receiver Operating

Characteristic (AUROC) difference of 0.09 compared to standard tools ($p = 0.008$). Similarly, for hospitalization and critical care admission, AI models demonstrated statistically significant improvements in predictive accuracy [28].

The technical architecture of these systems is evolving, while Logistic Regression remains common due to its interpretability [27], ensemble methods such as XGBoost (Extreme Gradient Boosting) and Random Forest are increasingly dominant due to their ability to handle non-linear relationships in complex clinical data [12], [29]. For instance, retrospective analyses indicate that XGBoost algorithms can achieve sensitivity rates as high as 96.2% for critical care outcomes, surpassing traditional clinical risk scores like the Sequential Organ Failure Assessment (SOFA) [30]. Furthermore, comparative studies suggest that ML models can reduce mistriage rates significantly and one study reported a reduction in the mistriage of critically ill patients to 0.9% compared to 1.2% using traditional systems [30].

However, the integration of these models into clinical workflows remains complex. Nover et al. [16] conducted a prospective study comparing an ensemble ML model against triage nurses in predicting hospital admissions. The ML model achieved an accuracy of 86.7% compared to the nurses' 81.6% [16]. Crucially, this study challenged the prevailing "human-in-the-loop" assumption, combining nurse predictions with the ML model did not improve overall accuracy, suggesting that, specifically in terms of raw predictive performance for structured tasks, algorithmic prediction may have reached a threshold where human input does not necessarily add incremental statistical value [16].

3.2. Operational Efficiency and Resource Management

Beyond clinical triage, AI is reshaping the administrative and operational layer of public health. Resource management, specifically the optimization of Operating Rooms (ORs) and Post-Anesthesia Care Units (PACUs), is a critical area of focus. Bellini et al. [29] highlighted that accurate prediction of surgical case duration is vital for cost reduction and scheduling efficiency [29]. In their systematic review, surgeon-specific ML models using XGBoost consistently outperformed service-specific averages in predicting case durations, allowing for more precise scheduling and reduced idle time.

Furthermore, AI applications are being deployed to predict surgical cancellations. Studies utilizing Random Forest algorithms have achieved accuracy rates of over 85% in identifying surgeries at high risk of cancellation, allowing hospitals to overbook or rearrange schedules proactively to maximize resource utilization [29]. In the context of bed management, ML models have demonstrated the ability to predict prolonged PACU stays, enabling hospitals to re-sequence cases and reduce after-hours staffing needs [29]. These applications directly address the efficiency imperatives of New Public Management, yet they rely heavily on the availability of high-quality, structured retrospective data, which remains a barrier in many institutions [12].

3.3. Chronic Disease Management and Public Health Surveillance

The scope of AI extends from acute care to the longitudinal management of chronic diseases. In China, where the aging population presents a significant public health challenge, AI is being utilized to personalize treatment plans for type 2 Diabetes. He and Li [24] describe systems that integrate data from wearable sensors and electronic medical records to provide real-time monitoring and early warning systems for glucose fluctuation. These systems not only improve patient adherence through personalized telemedicine platforms but also support "precision public health" by identifying high-risk populations for early intervention [24].

However, the shift towards predictive surveillance introduces significant ethical risks regarding civil liberties. Molldrem, Smith, and McClelland [31] analyzed the use of predictive analytics in HIV surveillance in the United States. They identified that public health programs are increasingly using integrated data systems, combining clinical data with records of incarceration, substance use, and social services, to predict individuals likely to fall out of care [31]. While the stated goal is to prioritize resources for vulnerable patients, this practice creates a hierarchy of "compliant" versus "deviant" biomedical subjects. Such "predictive public health" challenges the traditional regulatory frameworks designed for population-level benefits, shifting the focus to individual behavioral policing and raising profound concerns about data rights, consent, and function creep [31].

3.4. Emerging Technologies: Generative AI and LLMs

The emergence of Large Language Models (LLMs) represents the newest frontier in health AI. Gebrael et al. [17] retrospectively analyzed the performance of ChatGPT 4.0 in triaging patients with metastatic prostate cancer. The model demonstrated a high sensitivity (95.7%) in determining the need for admission and provided more comprehensive diagnostic lists than those generated by emergency physicians [17]. However, the study also exposed critical limitations inherent to Generative AI: the model exhibited specificity issues (18.2% in discharging patients) and instances of "hallucination", where it missed subtle clinical cues such as immunotherapy-related pancreatitis [17]. These findings suggest that while LLMs hold promise for administrative summarization and diagnostic support, their "black box" nature and potential for factual fabrication currently necessitate rigorous human oversight.

3.5. Critical Analysis: The Implementation and Accountability Gaps

Despite the abundance of literature validating the technical performance of AI, a critical "implementation gap" persists. Yin, Ngiam, and Teo [32] noted in a systematic review that the vast majority of health AI studies remain in the proof-of-concept phase using retrospective data, with very few applications successfully integrated into real-life clinical practice [32]. Several structural and ethical barriers contribute to this gap:

3.5.1. *Lack of Generalization and External Validation*

Piliuk and Tomforde [27] highlight that most AI studies in emergency medicine rely on private, single-center datasets [27]. This lack of external validation means that an algorithm trained in one hospital system often fails when transferred to another due to variations in patient demographics and data recording practices.

3.5.2. *The Risk of Self-Fulfilling Prophecies*

Nord-Bronzyk et al. [20] argue that AI implementation carries the risk of untraceable false positives. If an algorithm incorrectly predicts a patient has a high mortality risk and recommends palliative care over aggressive treatment, the patient's subsequent death "validates" the algorithm's prediction. This creates a closed loop where the algorithm's errors become self-fulfilling and undetectable by standard accuracy metrics.

3.5.3. *Automation Bias and Interpretability*

There is a tension between model performance and interpretability. Complex "black box" models like Deep Neural Networks often outperform interpretable models (e.g., Decision Trees) but lack transparency [33]. This opacity exacerbates "automation bias", where clinicians may over-rely on algorithmic outputs without critical evaluation, potentially leading to errors in high-pressure environments [20].

3.5.4. *Ethical Principles and Governance*

The deployment of these systems requires adherence to ethical principles that go beyond technical accuracy. Tiribelli et al. [34] propose a framework for AI-based telemedicine that emphasizes not just beneficence and non-maleficence, but also explicability and justice.

4. AI in Justice

The integration of AI into the justice sector represents a paradigm shift from mere digitization to the automation of cognitive legal tasks. While the adoption of these technologies offers the potential to address chronic inefficiencies, such as case backlogs and inconsistent sentencing, it simultaneously disrupts traditional legal principles, introducing profound risks to due process, accountability, and the presumption of innocence. This section analyzes the state of the art regarding AI in justice, categorizing the literature into applications, predictive mechanisms, ethical-legal challenges, and emerging pathways for trustworthy governance.

4.1. From Digitization to Cognitive Automation

A sharp divergence exists between international implementation models. Xu [18] highlights the deployment of "Smart Courts" in China, where systems like the "206 system" in Shanghai assist in evidence verification and sentencing guidelines, demonstrating that AI can process massive amounts of judicial data to ensure "similar cases are decided similarly" [18]. In contrast to this push towards full automation, the European context presents a distinct reality check. Scholars like Fortes [35] emphasize that the path to digital justice in Western legal orders points towards "technological support for decision-making rather

than to robots in robes making automated decisions with human intervention" [35]. They warn that, while automation is feasible for "bound acts" (where the law defines clear consequences), it faces severe legal limitations when applied to "discretionary acts" involving indeterminate legal concepts (e.g., "grave negligence"), which require a human value judgment that current AI lacks [3]. This distinction is critical, while Asian models explore autonomous adjudication, the European and Portuguese reality, as discussed by Magalhães, Costa, and Matos [3], focuses on a "Hybrid Administration" where AI is strictly auxiliary, ensuring that human judgment remains central to avoid the erosion of the rule of law.

4.2. Predictive Justice and Risk Assessment

A significant portion of the literature focuses on "predictive justice" particularly in recidivism risk calculation and policing. This domain illustrates the transition from traditional statistical methods to ML techniques that identify complex, non-linear patterns in historical data.

4.2.1. *Recidivism Prediction Tools*

Tools such as COMPAS and Harm Assessment Risk Tool (HART) are widely discussed. These systems use static factors (e.g., criminal history) and dynamic factors (e.g., employment status, social network) to forecast an individual's likelihood of reoffending [23] and while traditional models like Logistic Regression have been the standard, recent studies by Cavus et al. [36] propose advanced DL frameworks, such as the RCN, to improve prediction accuracy.

Crucially, in the specific context of Portugal, recent empirical research has begun to test these architectures against national data. Carvalho [37] conducted a pioneering study using datasets from the Direção Geral de Reinserção e Serviços Prisionais (DGRSP). The study found that ensemble methods, specifically Gradient Boosting algorithms, yielded the best predictive performance for recidivism in Portugal, outperforming Decision Trees and Random Forests. However, the same study noted that unsupervised learning techniques failed to identify distinct offender subgroups, suggesting a high heterogeneity in the national data that standard algorithmic profiling struggles to categorize [37].

4.2.2. *Predictive Policing*

Beyond individual risk, AI is used for "spatio-temporal predictive policing", forecasting crime hotspots based on historical data. Simmler et al. [21] identify that while these tools are used in Switzerland to support planning, they are rarely allowed to make autonomous decisions due to skepticism among practitioners regarding their effectiveness and the risk of "pseudo-legitimation" of police actions [21]. O'donnell [9] categorizes these tools into person-based (risk assessment of individuals) and place-based (geographic profiling), noting that both rely heavily on historical crime data which is often a reflection of police activity rather than actual crime incidence [9].

4.3. Operational AI in Portuguese Courts: The IRIS Project

While predictive tools remain largely in the research phase within the national territory, the Portuguese judicial system has actively begun deploying AI for administrative efficiency and data protection. A prime example is the IRIS Project, presented in October 2023 by the Supreme Court of Justice (STJ) in partnership with INESC-ID and ISCTE-IUL.

Unlike risk assessment tools, IRIS focuses on the "back-office" of justice. It utilizes NLP techniques to automate the pseudonymization of judicial decisions, ensuring compliance with the General Data Protection Regulation (GDPR) before public release [38]. Additionally, the system employs advanced Optical Character Recognition (OCR) to digitize historical case files and semantic search algorithms to improve the accessibility of jurisprudence. This implementation serves as a practical validation of the "Hybrid Administration" model identified by Magalhães, Costa, and Matos [3], demonstrating that AI can be effectively deployed to improve transparency and judicial productivity without encroaching on the core discretionary function of the human judge.

4.4. Challenges to Fundamental Rights and Due Process

The literature reveals a critical consensus, the deployment of AI in justice introduces severe risks to fundamental rights, creating a tension between technological capability and legal guarantees.

4.4.1. *Algorithmic Bias and Discrimination*

A predominant concern is that AI systems perpetuate and amplify existing societal biases. O'donnell [9] argues that ML algorithms trained on historical crime data inevitably learn to associate race with criminality because the input data itself reflects historical over-policing of minority communities. This creates a "feedback loop" where the algorithm directs police to specific neighborhoods, leading to more arrests in those areas, which in turn reinforces the algorithm's bias [9].

The critique of the COMPAS algorithm by Pro Publica [23], highlighting that it generated false positives at a higher rate for black defendants than for white defendants, remains a central case study. Although developers argue that the system is calibrated for predictive parity (equal accuracy across groups), Chiao [39] demonstrates that it is mathematically impossible to satisfy all definitions of fairness simultaneously when base rates of recidivism differ between groups [39]. This underscores that "fairness" in AI is not merely a technical metric but a political choice [25].

4.4.2. *Opacity and the "Black Box" Problem*

The transparency of AI systems is hindered by both technical complexity (deep learning opacity) and legal barriers (trade secrets). Peluso Lopes [19] identifies a "black box" problem where the internal logic of an algorithm is unintelligible to human operators, making it impossible for judges to fulfill their duty to state reasons for their decisions [19]. In the Loomis v. Wisconsin case, the court upheld the use of COMPAS but issued

warnings about its proprietary nature, which prevented the defendant from scrutinizing the algorithm's internal weighting of factors [35]. This opacity compromises the principle of "equality of arms", as defense teams cannot effectively challenge evidence generated by proprietary software.

4.4.3. *Automation Bias and Judicial Independence*

Even when AI is used merely as "decision support", there is a risk of "automation bias", where human judges overly rely on algorithmic recommendations due to time pressure or a lack of technical expertise [19], [21]. Simmler et al. [21] describe a "gut-feeling paradox": users trust the algorithm when it confirms their intuition but override it when it contradicts them, potentially using the algorithm only for "pseudo-legitimation" of subjective decisions [21]. This threatens judicial independence, as the judge may effectively delegate the decision-making authority to the machine, eroding the human responsibility required by law [10].

4.4.4. *Erosion of the Presumption of Innocence*

Završnik [1] warns that AI facilitates a shift from a "post-crime" justice system (reacting to past acts) to a "pre-crime" logic (anticipating future risks). This transition creates "suspects" based on statistical probability rather than factual evidence, fundamentally eroding the presumption of innocence.

4.5. Critical Analysis and Identified Gaps

A critical limitation in the current state of the art is the predominance of Anglo-Saxon studies, creating a transferability gap for Civil Law jurisdictions like Portugal. As noted by Hartel, + Wegberg, and Staalduinen [40], findings from US-centric research are often based on legal systems where algorithms are trained on judicial precedents to predict case outcomes. However, this approach faces fundamental theoretical barriers in Civil Law systems, where general statutory norms take precedence over case-by-case precedent. This jurisdictional mismatch implies that the uncritical importation of tools developed for Common Law contexts into the Portuguese legal order would not only be technically unreliable but potentially unconstitutional.

5. AI in Public Finance

The integration of AI into public finance represents a paradigm shift from traditional, rule-based bureaucratic administration to predictive, algorithmic governance. This transition is most visible in the detection of fiscal fraud, the optimization of public budgeting, and the automation of welfare distribution. However, the literature reveals a profound tension: while advanced ML architectures offer unprecedented accuracy in detecting anomalies and optimizing resources, they simultaneously introduce significant risks regarding transparency, exclusion, and the erosion of state accountability. This section critically analyzes these dimensions, synthesizing technical methodologies with socio-legal implications.

5.1. Algorithmic Approaches to Tax Evasion and Revenue Mobilization

Unlike private sector financial fraud, which often deals with high-frequency transaction data, public finance fraud, specifically tax evasion, involves complex networks of opacity and regulatory arbitrage. The literature highlights a move away from manual auditing toward unsupervised learning and graph-based analysis to detect these sophisticated schemes.

A critical challenge in tax administration is the lack of labeled data, which renders standard supervised learning difficult. To address this, Savić et al. [15] propose the Hybrid Unsupervised Outlier Detection (HUNOD) method, applied to personal income tax declarations in Serbia. By combining clustering algorithms (K-means) with representational learning (autoencoders), this hybrid approach allows for the internal validation of outliers without requiring a pre-labeled dataset of fraud instances. Crucially, this method incorporates domain knowledge by allowing tax experts to assign weights to specific features, thereby bridging the gap between statistical anomaly detection and economic context [15].

The complexity of modern tax evasion often involves networks of affiliated companies engaging in "circular trading" or transfer pricing manipulation to obscure liabilities. Nuryani et al. [41] demonstrate that traditional SQL-based matching is insufficient for these scenarios. Instead, they argue for the use of Social Network Analysis (SNA) and graph data structures to map relationships between taxpayers, shareholders, and family members. Their findings indicate that taxpayers with high "degree centrality" in these ownership networks are significantly more likely to engage in tax avoidance, necessitating graph-based AI models to detect patterns invisible to linear analysis [41].

Furthermore, the potential for using Open Data to enhance fiscal oversight has been demonstrated in Brazil. Xavier et al. [42] developed models using Random Forest and Neural Networks trained solely on public data sources to identify tax-evading companies. Their results, achieving over 98% accuracy, suggest that public administrations need not rely exclusively on sensitive, internal data to build effective predictive models, offering a pathway for more transparent fiscal monitoring [42].

5.2. Automation in Public Budgeting and Expenditure

Beyond fraud detection, the utility of AI in public finance extends to the optimization of resource allocation through behavioral interventions. Santschi et al. [43] explore the phenomenon of "budgetary slack", where public managers intentionally underestimate revenue or overestimate expenses to create safety buffers, a practice often driven by agency theory and risk aversion [43].

Their experimental study offers a fascinating insight: the deployment of AI serves not only a computational function but also a behavioral one. The authors found that the mere presence of an 'AI slack evaluation', where an algorithm benchmarks human estimates against historical data, significantly reduced the tendency of participants to include slack in their budgets. This suggests that AI can act as a mechanism for 'behavioral governance,' increasing the perceived accountability of human decision-makers. By

narrowing the information gap between agents and principals, the AI system effectively nudged participants towards greater honesty and fiscal discipline, demonstrating that the value of AI in public budgeting lies as much in its psychological impact on human officials as in its predictive accuracy [43].

5.3. Technical Methodologies: Transferability from Private to Public Sectors

While specific literature on public sector algorithms is growing, the technical foundations are largely adapted from the private financial sector (e.g., credit card and insurance fraud). Understanding these methodologies is essential for defining the technical state of the art available to public administrators.

5.3.1. *Advanced Classifiers and Ensemble Learning*

The consensus in recent literature is that ensemble methods outperform single classifiers. Khalid et al. [8] and Ileberi, Sun, and Wang [44] demonstrate that combining algorithms such as Random Forest, SVMs, and K-Nearest Neighbors (KNN) significantly improves detection rates. Specifically, XGBoost is repeatedly cited as a superior algorithm for financial fraud due to its speed and accuracy [45]. In the context of vehicle insurance, Khan, Olivia, and Shetty [46] achieved 96% accuracy using a stacked ensemble model, highlighting the potential for these architectures in government claims processing.

5.3.2. *Addressing Data Imbalance*

A pervasive technical challenge in fraud detection is the "imbalanced dataset" problem, where fraudulent cases are rare compared to legitimate transactions. If unaddressed, this leads to models that are biased toward the majority class (non-fraud). The literature proposes some solutions relevant to public administration. Alshawi [47] validate the use of Generative Adversarial Networks (GANs), specifically CTGAN for tabular data, to generate synthetic fraudulent examples. This allows models to be trained on balanced datasets without compromising user privacy and is also called Synthetic Data Generation. Oversampling techniques like Synthetic Minority Over-sampling Technique (SMOTE) remain a standard and effective approach for balancing data in fraud detection systems [8], [44].

5.3.3. *Feature Selection and Optimization*

To handle the high dimensionality of financial data, evolutionary algorithms are being employed. Patel and Panday [13] and Ileberi, Sun, and Wang [44] demonstrate the efficacy of Genetic Algorithms (GA) for feature selection, which optimize the input variables for classifiers like Random Forest, ensuring that the models focus on the most relevant indicators of fraud.

5.3.4. *Explainable AI (XAI)*

As public administrations face strict legal requirements for transparency, "Black Box" models pose validation challenges. Mill et al. [48] argue for a research agenda focused on Intrinsic XAI (models that are interpretable by design) and faithful post-hoc explanations

to engender trust. Khan, Olivia, and Shetty [46] apply techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to insurance fraud, proving that it is possible to achieve high accuracy while retaining the ability to explain why a specific claim was flagged, a crucial requirement for due process in public administration.

5.4. The Accountability Gap: Risks and Ethical Implications

While the technical literature focuses on accuracy and efficiency, the socio-legal literature identifies severe risks when these systems are deployed in government. The delegation of sovereign decision-making to algorithms creates what Crawford and Schultz [4] identify as an "Accountability Gap".

5.4.1. *The State Actor Problem*

Government agencies frequently outsource the development of AI systems to private vendors. Crawford and Schultz [4] document cases such as the MiDAS unemployment fraud system in Michigan, which falsely accused thousands of beneficiaries with a 93% error rate. Because the state lacked the technical capacity to understand the "black box" algorithms provided by vendors, and because private vendors claimed trade secrecy, constitutional accountability was difficult to enforce. The authors argue that AI systems performing core government functions should be legally treated as "State Actors" to close this liability loophole [4].

5.4.2. *Digital Conditionality and Exclusion*

The automation of welfare and finance creates a new form of "Digital Conditionality." Morrison [49] examines the Universal Credit system in the UK, arguing that the "digital by default" design effectively makes digital proficiency a condition for accessing social rights. This excludes vulnerable groups (the elderly, those with learning difficulties, or limited internet access), forcing them to rely on informal "digital carers" and eroding their autonomy. Hofmann, Sæbø, and Rydén [50] support this finding in the Norwegian context (NAV), showing that vulnerable citizens often struggle with the complexity and anonymity of digital self-services, leading to feelings of powerlessness and humiliation.

5.4.3. *Regulatory Fragmentation*

Finally, there is a lack of standardization in how these systems are implemented. Marin [6] notes that even within a single country (Romania), electronic procedures for social benefits vary wildly between counties and institutions, creating an uneven landscape of access. Morse [51] introduces the concept of "Government-to-Robot Enforcement", suggesting a future where government regulation might bypass individuals and target the compliance software itself (e.g., TurboTax), potentially solving under-enforcement but raising new questions about fairness and "law creep."

5.5. Critical Analysis and Identified Gaps

The state of the art in AI for public finance is characterized by a high degree of technical sophistication, utilizing GANs, Ensemble Learning, and Graph Neural Networks, capable of detecting complex evasion and optimizing budgets. However, the operational reality reveals a disconnect: these tools are often deployed without adequate safeguards for explainability, leading to legal challenges (e.g., MiDAS) and social exclusion. The literature points to an urgent need for frameworks that integrate XAI [46], [48] and clear legal accountability for algorithmic outputs [4] to ensure that efficiency does not come at the cost of fundamental rights, underscoring the necessity for the operational guidelines proposed in this dissertation.

6. Cross-Domain Risks

The integration of AI across public administration sectors promises improved efficiency and accuracy. However, the literature consistently highlights that these technological advancements introduce profound transversal risks that challenge democratic values, the rule of law, and social equity. This section identifies the cross-cutting patterns that only become visible through comparison, and draws from them the analytical conclusions that motivate the governance framework proposed in this dissertation.

6.1. Algorithmic Bias and Feedback Loops

One of the most pervasive risks documented in the literature is the potential for AI systems to perpetuate and exacerbate existing societal inequalities. O'donnell [9] argues that predictive policing algorithms function as "self-fulfilling prophecies" because they are trained on data generated by human policing practices that historically target specific demographics. Consequently, rather than providing objective assessments, these systems create positive feedback loops where marginalized groups are disproportionately targeted, generating more data that reinforces the algorithm's bias in subsequent iterations [39]. Similarly, Huq [23] posits that algorithmic tools in criminal justice can deepen racial stratification, arguing that constitutional law currently lacks the tools to effectively address the "disparate impact" of these technologies.

This risk is not confined to the judiciary. In the financial domain, specifically regarding fraud detection, the use of historical transaction data can embed biases against certain consumer profiles. While ML models like XGBoost and Random Forest show high accuracy in detecting credit card fraud [8], [13], the literature warns that without careful feature selection and handling of imbalanced datasets, these models may generate false positives that disproportionately affect specific groups, leading to financial exclusion.

Attempts to mitigate these biases through technical means, such as integrating SMOTE to balance datasets, show promise in reducing discriminatory outcomes [8]. However, scholars argue that technical fixes alone are insufficient without a normative re-evaluation of how these tools interact with social stratification [23].

6.2. The "Black Box" Problem and Opacity

The second major cross-domain risk concerns the opacity of AI systems, often referred to as the "black box" problem. In the medical field, the lack of interpretability is cited as a significant barrier to clinical adoption.

While DL models may outperform traditional methods in triage or diagnosis, their inability to provide a clear causal explanation for their outputs poses ethical risks regarding patient safety and informed consent [30], [52]. Mill et al. [48] argue that while post-hoc explanations (e.g., LIME, SHAP) offer some insight, they may lack faithfulness to the underlying model, creating a tension between the pursuit of accuracy and the requirement for explainability in public law.

6.3. Accountability Gaps

The automation of public services introduces a complex "accountability gap". When decision-making authority is delegated to algorithms, determining responsibility for errors, whether it lies with the vendor, the public agency, or the human operator, becomes legally and ethically fraught [4].

Crawford and Schultz [4] identify a critical issue where AI systems developed by private vendors are used to perform core state functions. This outsourcing can shield the state from constitutional liability due to trade secrecy laws that prevent the scrutiny of proprietary source code. This "vendor lock-in" limits the public administration's ability to oversee and correct system errors.

Furthermore, the phenomenon of "automation bias" suggests that human operators may over-rely on algorithmic outputs, effectively ceding discretionary power to the machine and eroding the human responsibility required by due process [21].

6.4. Digital Exclusion and the Erosion of Social Citizenship

Finally, the "digital by default" approach in welfare services risks creating a two-tier citizenship. Evidence from the UK's Universal Credit [49] and Norwegian welfare services [50] demonstrates that vulnerable populations often lack the digital literacy to navigate complex automated systems, leading to a loss of autonomy and access to social rights.

Similarly, in the context of Romanian social services, Marin [6] highlights that the lack of standardization and the persistence of the digital divide mean that e-government initiatives often fail to reach those most in need, reinforcing existing social stratifications.

In the healthcare sector, Tiribelli et al. [34] warn that AI-based telemedicine must account for "digital inclusion" to avoid violating the principle of non-maleficence, ensuring that minority groups are not systematically excluded from the benefits of digital health innovation.

7. Portuguese Context & Conclusion

The systematic literature review presented in this chapter reveals a fundamental paradox. While the global state of the art demonstrates technical maturity, the transposition of these architectures to the Portuguese reality faces challenges that extend beyond the general hazards discussed previously.

Specifically, while the cross-domain risks identified in Section 6, such as algorithmic bias, opacity, accountability gap and the digital exclusion, are fully applicable to the national jurisdiction, they are compounded by specific dogmatic and operational barriers inherent to the Portuguese legal-administrative tradition. This section synthesizes these unique gaps, arguing that the direct importation of international "black box" models would not only be risky but potentially unlawful under current administrative norms, thus establishing the central argument for the tailored guidelines proposed in this dissertation.

7.1. The Legal Administrative Block: The Civil Law Dilemma

A significant portion of the analyzed literature is based on Common Law systems, where decisions can sometimes accommodate models based on precedents. However, the direct application of these solutions in Portugal, a Civil Law system, is perilous. The Portuguese administrative system imposes a strict duty to substantiate administrative acts, enshrined in Articles 152 and 153 of the CPA. As highlighted by Magalhães, Costa, and Matos [3], the use of opaque "black box" algorithms collides frontally with this duty. If the Administration cannot explain the *iter cognoscitivo* that led to a decision, the act is liable to annulment, violating the constitutional right to effective judicial protection.

7.2. Limits of Discretion and the PRR Challenge

The analysis further reveals that AI cannot be uniformly applied across all administrative functions. It is imperative to distinguish between "bound acts" (where the law strictly defines assumptions and consequences, making automation feasible) and "discretionary acts" (which require value judgments and the weighing of competing interests). National legal literature emphasizes that current algorithms lack the "ethical sensitivity" necessary to interpret indeterminate legal concepts, such as "grave negligence" or "public interest", risking the erosion of the margin of appreciation constitutionally conferred upon the Human Administration.

As emphasized by Fortes [35], unlike certain Asian models that explore autonomous adjudication, the path to digital justice in Western legal orders points strictly towards "technological support for decision-making rather than to robots in robes making automated decisions" [35]. Consequently, the Portuguese reality, as discussed by Magalhães, Costa, and Matos [3], necessitates a model of "Hybrid Administration". In this model, AI remains strictly auxiliary, ensuring that human judgment remains central to avoid the erosion of the rule of law and to mitigate the risks of automation bias.

However, a critical tension arises with the implementation of the PRR. The drive for rapid digital transition risks overlooking this nuance. Without tailored guidelines that

enforce this "hybrid" approach, there is a danger that the pressure for efficiency might lead to the deployment of "digital by default" systems that inadvertently bypass the human discretionary layer, creating barriers for citizens who cannot effectively engage with the "third actor" in administrative relations.

7.3. Conclusion: The Operational Void

This chapter concludes that there is a critical "operational void." The global technology exists that promises efficiency, yet its direct application clashes with the guarantees of the Portuguese Rule of Law. The risk is not merely technical, but constitutional.

This collision justifies the urgent need for the operational guidelines proposed in this dissertation. To lay the groundwork for these guidelines, the following chapter, Chapter 3 (Comparative Framework Analysis), moves beyond theoretical risks to examine the normative reality. By analyzing international regulations (such as the EU AI Act) and strategic government documents, it will benchmark Portugal's current maturity against global best practices. The objective is to understand how other countries are navigating these challenges and to identify specific policy frameworks that can be adapted to resolve the legal-technical deadlock within the Portuguese Public Administration.

[This page is intentionally left blank.]

CHAPTER 3

Comparative Framework Analysis

1. Introduction

The review of the State of the Art in the previous chapter demonstrated that the integration of AI into the public sphere is not without risks. Dangers such as algorithmic opacity (the "black box" phenomenon), data bias, and a lack of accountability are not merely theoretical hypotheses but tangible threats to the integrity of administrative functions. However, the identification of these risks leads to a pragmatic and urgent question: How prepared are the current legal framework and governance strategies to mitigate these dangers?

This chapter aims to transform the discussion of abstract risks into a concrete normative reality. Its central purpose is to conduct a diagnosis of institutional maturity, contrasting the emerging regulatory requirements of the European Union and international best practices with the operational reality of the Portuguese Public Administration.

This diagnosis is motivated by the "legal-technical deadlock" identified in the problem statement of this dissertation: the tension between the efficiency promised by predictive algorithms and the duty to justify administrative acts, as established in Articles 152 and 153 of the CPA. In a state governed by the rule of law, the Public Administration cannot simply decide effectively, it must be able to explain the reasoning behind the decision to the citizen. Without mechanisms for transparency and explainability, the use of AI in the public sector runs the risk of generating administrative acts that are null and void or voidable due to a lack of justification.

To address this challenge, this chapter is structured around three levels of analysis:

(1) **The Legal Framework** (Hard Law): This section analyzes the impact of the EU AI Act [53]. This regulation shifts the paradigm from ethical voluntarism to mandatory legal compliance, classifying the majority of systems used in the public sector as "High Risk". The requirements for transparency and human oversight will be examined, as they are no longer optional but have become prerequisites for legality.

(2) **Best Practice Benchmarks** (Soft Law): This section observes how other jurisdictions and international bodies are implementing these principles. It examines the OECD frameworks for system classification and the data ethics models of the United Kingdom and Canada, which offer practical self-assessment tools that are currently nonexistent in Portugal.

(3) **The National Diagnosis:** This section assesses Portugal's strategic maturity. Although documents such as the National Digital Strategy and AI Portugal 2030 demonstrate a clear ambition for innovation and technical qualification, a critical gap is identified regarding governance procedures. The analysis will highlight that, despite promising initiatives such as the future "National AI Agenda", an "operational void" persists: public workers possess the technology but lack formal guidelines on how to utilize it in compliance with the new European regulatory framework.

2. The "Hard Law": Impact of the EU AI Act on Public Administration

This section analyzes a definitive transition in the governance of technology. With the approval of Regulation (EU) 2024/1689, commonly known as the EU AI Act, the adoption of AI by the Public Administration is no longer exclusively a matter of innovation policy or modernization strategy. Instead, it becomes a rigorous exercise in compliance with fundamental rights protection. The Regulation establishes a uniform legal framework across the Union, aiming to mitigate the specific asymmetries of power that arise when the State utilizes automated systems to make decisions affecting its citizens.

2.1. The Ethical "Red Lines": Prohibited Practices (Article 5)

Before examining the compliance obligations for high-risk systems, it is imperative to establish the absolute boundaries of State intervention defined by the Regulation. Article 5 presents a catalog of prohibited AI practices considered to pose an "unacceptable risk" to fundamental rights [53]. For the Portuguese Public Administration, these prohibitions act as a safeguard against the dystopian drift of algorithmic governance, specifically in two critical areas.

Starting with the Prohibition of Social Scoring, the Regulation explicitly forbids public authorities from using AI to evaluate or classify individuals over a certain period of time based on their social behavior or known/predicted personality characteristics [53]. This provision is vital for the Portuguese context of digital interoperability, as it ensures that data collected in one administrative context (e.g., social security debts) cannot be used by an AI system to generate a "score" that leads to detrimental or unfavorable treatment in an unrelated context (e.g., access to healthcare or education).

Also with limits on Predictive Policing, in the domain of justice and security, the Act prohibits AI systems used to assess the risk of an individual committing criminal offenses based solely on profiling or assessing personality traits [53]. This reinforces the presumption of innocence. It mandates that the State cannot rely on algorithmic probability alone to determine criminal risk. Any such assessment must be supported by human evaluation based on objective, verifiable facts, effectively barring fully automated "pre-crime" mechanisms in Portugal.

By establishing these "red lines", the European Parliament and Council of the European Union [53] clarify that while administrative efficiency is a goal of modernization, it cannot supersede the dignity of the citizen.

2.2. The Risk-Based Approach and the Public Sector

The regulatory architecture of the AI Act is built upon a "risk-based approach", which imposes burdens proportionate to the potential harm a system may cause. For the Portuguese PA, the implications of this approach are profound, particularly when analyzing Annex III of the Regulation. Unlike the private sector, where many AI applications might fall into low-risk categories, the core functions of the welfare state are almost entirely classified as high-risk.

Specifically, the Regulation identifies that AI systems intended to be used by public authorities for evaluating the eligibility of natural persons for essential public assistance benefits and services such as social security, housing support, or unemployment benefits carry a significant potential for material and immaterial harm. Similarly, tools used in migration, asylum, and border control management, as well as those employed in law enforcement for individual risk assessment, are subject to the same stringent classification.

The legal consequence of this classification is that public entities cannot simply adopt solutions that are ready to use. Before any system enters into service, it must undergo a rigorous conformity assessment. This effectively ends the era of experimentalism in live environments for critical public services, the State must now prove the safety and robustness of its algorithms before they impact a single citizen.

2.3. Transparency as a Prerequisite for Administrative Legality

One of the most significant contributions of the AI Act to the administrative domain is the codification of transparency requirements, which directly addresses the "black box" problem identified in the literature. Article 13 mandates that high-risk AI systems must be designed and developed in a way that ensures their operation is sufficiently transparent to enable implementers, in this case public authorities, to interpret the outcomes of the system and use them appropriately.

This requirement is not merely technical, it acts as a bridge to administrative law. In the Portuguese context, the CPA imposes a duty to state reasons for any act that affects the rights of citizens. By enforcing technical transparency through concise, complete, and accessible instructions for use, Article 13 provides civil servants with the necessary information to understand the algorithmic rationale.

Furthermore, the Regulation introduces a new right in Article 86: the Right to explanations regarding individual decisions. This provision empowers any person affected by a decision taken by a high-risk AI system to request a clear and meaningful explanation of the role the AI played in that decision [53]. This effectively creates a new layer of administrative due process, mandating that the *iter cognoscitivo*, the reasoning behind the

decision, must be reconstructible and intelligible, preventing the use of incomprehensible algorithms in public policy.

2.4. Human Oversight and the Imperative of AI Literacy

To combat the phenomenon of "automation bias", the psychological tendency of humans to rely excessively on automated recommendations, Article 14 establishes Human Oversight as a non-negotiable operational requirement. The law dictates that high-risk systems must be designed with human-machine interfaces that allow authorized personnel to effectively supervise their operation [53]. Crucially, this oversight is not passive. The Regulation specifies that the human in the loop must possess the authority and competence to interpret outputs and, if necessary, to disregard, override, or reverse the system's decision. This validates the concept of "Hybrid Administration," where the human official retains the final sovereignty over the decision-making process.

However, effective oversight requires competence. Recognizing this, Article 4 introduces a mandatory requirement for AI Literacy. Public entities are now legally bound to ensure that their staff, including those who deploy and oversee AI systems, have a sufficient understanding of how these systems work [53]. This provision transforms training and capacity building from a human resources recommendation into a legal compliance obligation. Without a literate workforce capable of critical engagement with the technology, the deployment of high-risk AI becomes unlawful.

2.5. The Fundamental Rights Impact Assessment

Perhaps the most distinct and consequential obligation for public sector deployers, as opposed to private companies, is established in Article 27. Prior to deploying a high-risk AI system, public bodies (or private entities providing public services) must conduct a Fundamental Rights Impact Assessment (FRIA) [53]. This provision acknowledges the unique power asymmetry between the State and the citizen, establishing a procedural safeguard to ensure that efficiency gains do not come at the expense of constitutional rights.

The Regulation stipulates that this assessment must be a comprehensive exercise, shifting the burden of due diligence firmly onto the deploying institution. Operationally, this assessment requires the public entity to map the specific temporal and procedural context in which the system will operate, explicitly identifying the categories of individuals and vulnerable groups that are likely to be affected.

This moves beyond abstract risk analysis to a concrete evaluation of the system's impact on the specific population it will serve. It ensures that efficiency gains promised by AI adoption are weighed against their potential societal costs before implementation, thereby aligning the technological deployment with the constitutional values of the Portuguese Republic.

3. The "Soft Law": The Best Practice Benchmarks

While the EU AI Act establishes the mandatory legal boundaries and fundamental rights safeguards, it does not provide the operational manuals required for daily administrative management. The frameworks developed by global organizations and pioneering nations are analyzed not merely as theoretical references, but as practical toolkits designed to translate legislative mandates into actionable administrative procedures.

3.1. The Operational Void and the Function of Soft Law

The analysis of the European "Hard Law" in the previous section reveals a landscape of robust legal obligations, namely the duties of transparency, human oversight, and non-discrimination. However, the legislation creates what can be designated as an "operational void": the regulation dictates *what* the Public Administration must guarantee, but it does not specify *how* to technically build, procure, or audit a system that meets these stringent requirements.

It is in this context that international Soft Law assumes a critical function, not as "weak law" but as the repository of technical standards, engineering toolkits, and governance directives that allow for the operationalization of the law. As highlighted by the OECD and UNESCO [54], the role of governments transcends mere regulation, it requires the creation of an "enabling environment." This environment is constructed through practical guidance on public procurement, skills development and continuous monitoring, transforming abstract ethical principles into actionable system requirements.

3.2. Taxonomies and Risk Classification: The Technical Foundation

The ability to define and categorize the regulated object is a basic prerequisite for any successful regulation. Portugal currently lacks an official taxonomy that would enable public decision-makers to technically distinguish between high-risk and low-risk systems.

The OECD [55] is the international reference tool used to close this gap. Rather than a straightforward binary classification, this model suggests a multidimensional analysis that enables the mapping of the policy implications of particular systems. The structure, illustrated in Figure 3.1, is based on five interdependent dimensions, the detailed analysis of which is imperative for national algorithmic governance:

(1) **People and Planet:** This core dimension evaluates the system's potential to promote human-centric and trustworthy AI. It goes beyond identifying the direct operators (e.g., public employees) to include a broader mapping of all individuals and passive stakeholders who interact with or are affected by the applied AI system. Crucially for the public sector, it demands a rigorous assessment of the system's social footprint, including its impact on fundamental human rights, citizen well-being, environmental sustainability, and the labor market. It also assesses the "optionality" of the application, that is, the level to which a citizen can opt out of the automated process without suffering detrimental consequences.

(2) **Economic Context:** This dimension maps the specific sectoral and organizational environment in which the applied AI system operates. In the context of Public Administration, it requires defining the specific public service domain (such as healthcare triage, criminal justice, or tax fraud detection) and the administrative function the system serves. Importantly, it guides decision-makers to evaluate the "critical nature" of the deployment, determining whether system failure would disrupt essential public services as well as assessing the scale of its impact and the maturity of the underlying technology prior to public procurement.

(3) **Data and Input:** Focusing on the raw materials of algorithmic decision-making, this dimension analyzes the data and expert inputs used to build the system's computational representation of the environment. It mandates a detailed scrutiny of data provenance, collection methods (distinguishing between human-led and machine-automated collection), and structural formats. A critical contribution of this framework is the distinction it draws between data used to train the system ("in the lab") and the live data it processes during real-world operation ("in the field"). For public entities, this rigorous data profiling is essential to detect historical biases at the source and ensure compliance with privacy regulations.

(4) **AI Model:** This dimension deconstructs the computational core, the "black box", that represents the external environment and internal decision-making logic. It requires public managers to classify the technical architecture of the system, distinguishing between models built via explicitly programmed expert knowledge (symbolic AI), those developed through data-driven ML (statistical AI), or hybrid approaches. Understanding this underlying architecture is directly tied to the system's explainability, dictating how its performance objectives can be audited by regulatory bodies and how its logic can be made intelligible to the citizens it affects.

(5) **Task and Output:** The final dimension evaluates the operational execution of the system, what it actually does and the tangible consequences of its outputs. It categorizes the specific tasks performed by the algorithm, such as pattern recognition, forecasting risks (e.g., recidivism prediction in justice), or goal-driven optimization (e.g., resource allocation in hospitals). Above all, it assesses the system's "action autonomy," determining whether the AI merely recommends an action to a human official or independently executes an administrative decision. This evaluation is critical for defining the necessary level of human oversight and establishing appropriate performance metrics.

3.2.1. *The Crucial Distinction: "In the Lab" vs. "In the Field"*

Beyond the static classification of an algorithm, one of the most sophisticated contributions of the OECD framework and a critical concept for the Portuguese public sector

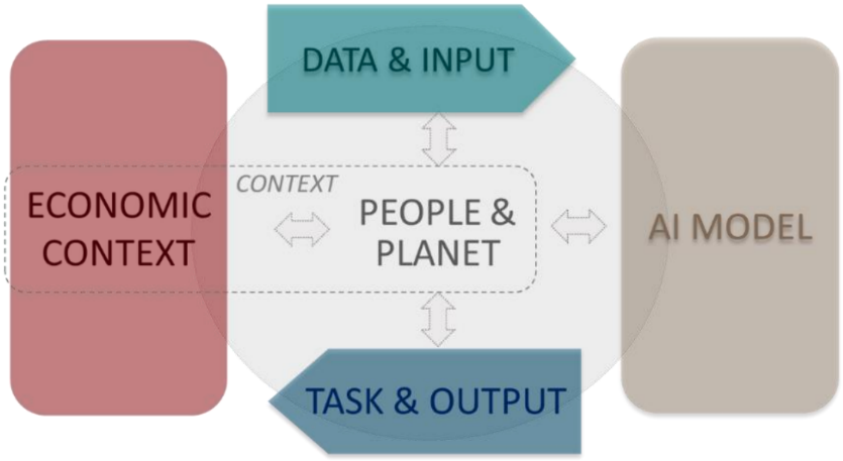


FIGURE 3.1. The 5 Dimensions of the OECD Classification Framework (Source: OECD [55]).

strategy is the operational distinction between evaluating an AI system "in the lab" versus "in the field" [55]. As delineated in Figure 3.2, the "in the lab" context refers to the conception and development phase prior to deployment. At this stage, risk management is strictly *ex ante*, focusing primarily on the technical dimensions of the framework: Data & Input (e.g., qualifying the historical data), the AI Model (e.g., training the initial architecture), and the intended Task & Output. For public procurers, this represents the technical validation threshold that must be crossed before a system is acquired.

Deployment "in the field" refers to its active use and continuous evolution. The evaluation must shift towards *ex post* risk management. Crucially, it is only "in the field" that all five dimensions of the framework become fully activated. The societal implications (People & Planet) and structural impacts (Economic Context) can only be truly measured once the system begins interacting with real citizens and integrating with existing public service infrastructures.

This distinction highlights the inherently dynamic and non-deterministic nature of Artificial Intelligence. As the OECD [55] notes, a system in the field can change significantly over time, whether through the ingestion of new datasets, an increase in technological maturity, or a broadened scale of deployment. Consequently, this exposes the Administration to the phenomenon of "model drift" (algorithmic degradation), where a system that operated flawlessly and fairly "in the lab" begins to produce biased or inaccurate outcomes when exposed to shifting real-world data.

This conceptual split highlights a profound governance challenge. For the Portuguese Public Administration, it implies that one-off, pre-deployment audits are legally and operationally insufficient. Because the classification of an AI system may change as it evolves, governance must be a continuous lifecycle process, subjected to regular reviews. This principle of continuous monitoring will constitute a foundational pillar of the operational guidelines proposed in Chapter 5.

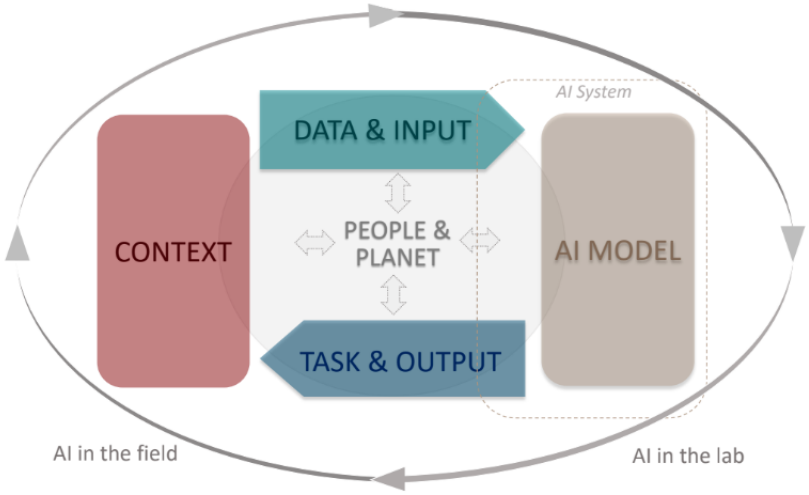


FIGURE 3.2. Application of an AI system "in the lab" or "in the field" (Source: OECD [55]).

3.3. Algorithmic Transparency: From Access Requests to Public Registers

The operationalization of algorithmic transparency in the public sector is not a monolithic concept, but rather a spectrum of governance practices. To diagnose the institutional maturity required in Portugal, it is fundamental to distinguish between traditional reactive approaches and the new global trend of proactive disclosure. The report *Algorithmic Transparency in the Public Sector*, developed by the Global Partnership on Artificial Intelligence (GPAI) [56], provides the essential taxonomy for this distinction, as summarized in Figure 3.3.

3.3.1. *Paradigm Shift: Demand-Driven vs. Supply-Driven*

Traditionally, administrative transparency operates on a "demand-driven" logic. In this model, the State acts only when prompted, responding to specific requests for information from citizens, journalists, or courts. While essential and implemented in Portugal through the *Lei de Acesso aos Documentos Administrativos (LADA)*, the right of access to administrative documents, this mechanism is insufficient for the scale and complexity of AI. It unfairly shifts the burden of oversight onto the citizen, who is often unaware of the very existence of the automated system evaluating them [56].

In contrast, the global trend points towards "supply-driven transparency". Here, the initiative originates from the State, which routinely discloses information about the Automated Decision-Making (ADM) systems it utilizes, without the need for external solicitation.

The primary instruments of this active transparency are Public Algorithmic Registries. These function as consolidated directories that serve as a "window" for the public to understand how decisions are made, what data is used, and how to contest the outcomes.

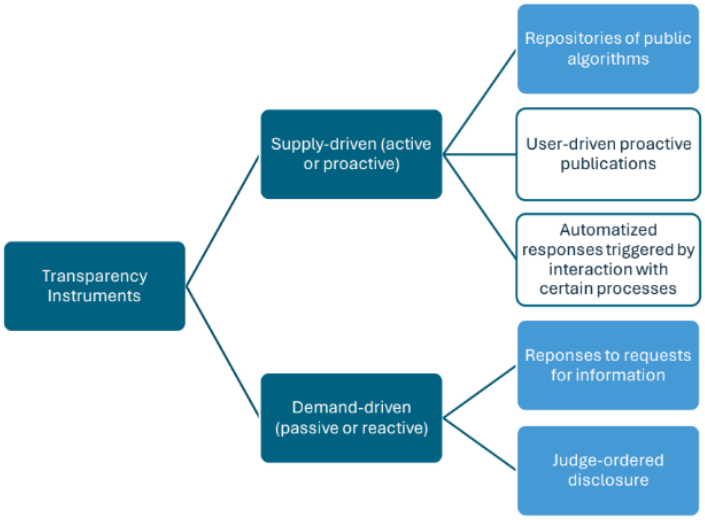


FIGURE 3.3. Classification of algorithmic transparency instruments (Source: Global Partnership on Artificial Intelligence (GPAI) [56]).

3.3.2. *The Global State of the Art: Fragmentation and Heterogeneity*

A 2024 mapping by the Global Partnership on Artificial Intelligence (GPAI) [56] identified 69 active registries worldwide. As shown in Figure 3.4, the geographical distribution reveals a significant concentration in the Global North, with the United States (37 registries) and the Netherlands (14 registries) leading implementation, followed by initiatives in France and the United Kingdom [56].

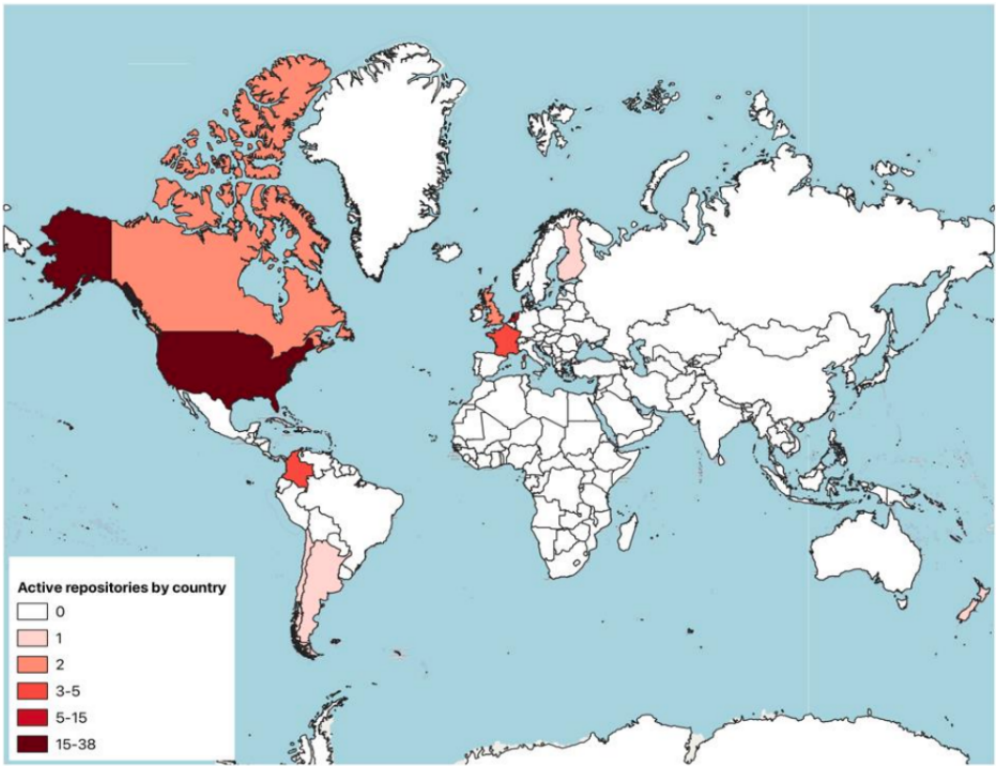


FIGURE 3.4. Geographical distribution of active algorithmic registries (Source: Global Partnership on Artificial Intelligence (GPAI) [56]).

However, the existence of these registries does not guarantee uniformity. The analysis reveals a significant dispersion in governance models: while some are centralized and managed by national bodies (62%), others are fragmented subnational initiatives (21%) or managed by civil society and academia. This fragmentation suggests that, in the absence of rigorous national standards, transparency becomes dependent on the political will of individual agencies or municipalities, creating information asymmetries within the same State.

3.3.3. *The Risk of "Performative Transparency"*

The phenomenon of "performative transparency" is the most pointed criticism made by the Global Partnership on Artificial Intelligence (GPAI) [56] report, and it should serve as a warning for implementation in Portugal. The practice of public entities publishing algorithmic registries to give the impression of efficiency and innovation without actually providing the technical information required for true scrutiny is described by this concept [56]. Therefore, the mere creation of an online portal does not solve the "black box" problem. Without binding standards requiring the disclosure of critical variables, like the model logic, training data, and error rates, registries risk becoming public relations exercises rather than tools for democratic accountability and legal efficiency. An example of meaningful transparency that counters this risk is the UK Algorithmic Transparency Recording Standard, shown in Figure 3.5.



FIGURE 3.5. Example of Meaningful Transparency: The UK Algorithmic Transparency Recording Standard (from gov.uk/algorithmic-transparency-records).

For the Portuguese reality, adopting such a robust registry is not merely an ethical measure, but an instrument of legal efficiency. It allows for systemic compliance with the duty to state reasons required by the CPA. By proactively disclosing the algorithmic logic, the Public Administration can prevent litigation and avoid overburdening administrative courts with repetitive requests for explanations.

3.4. Canada's Risk-Based Directive: The Administrative Compass

If the OECD framework provides the taxonomy and algorithmic registries provide the transparency, the Government of Canada offers the operational rulebook. In fact, the Government of Canada [57] stands as a pioneering, top-down instrument designed specifically for the public sector. Its primary objective is to ensure that ADM systems reduce risks to clients and society while leading to more efficient, accurate, consistent, and interpretable administrative decisions.

For the Portuguese Public Administration, the Canadian model can act as administrative guidance, demonstrating how to successfully implement a comprehensive risk-based approach as mandated by the EU AI Act. The Directive achieves this through two highly pragmatic mechanisms: Algorithmic Impact Assessment and a strict, four-tier intervention framework.

3.4.1. *The Algorithmic Impact Assessment as an ex ante Standard*

The Directive mandates that before any automated decision system is brought into production, the deploying institution must complete an Algorithmic Impact Assessment. This assessment is a mandatory questionnaire that quantifies the risk level of the system based on its design, the data used, and its potential impact on the rights, health, and economic well-being of citizens.

In the context of this dissertation, the Canadian Algorithmic Impact Assessment (AIA) serves as a practical model of how the Portuguese Government could implement the Fundamental Rights Impact Assessment required by Article 27 of the EU AI Act. Instead of leaving public managers to interpret abstract legal requirements, the Algorithmic Impact Assessment translates legal risks into a standardized and quantifiable score.

3.4.2. *Calibrating Administrative Discretion: The Four Levels of Impact*

The most significant contribution of the Canadian Directive is its Appendix C, which links the impact level generated by the AIA to mandatory governance requirements, particularly concerning human intervention. This mechanism directly addresses the tension between automation and administrative discretion identified in Chapter 2.

The Directive establishes four impact levels (I to IV), scaling the requirement for the "human-in-the-loop" proportionally to the risk:

- **Level I (Low Impact):** The system's decisions have reversible and brief impacts. The system may make decisions without direct human involvement.
- **Level II (Moderate Impact):** The decisions have temporary impacts. Human intervention is required primarily for quality assurance and overriding exceptional cases.
- **Level III (High Impact):** The decisions have lasting impacts on individuals. Human oversight must be integrated directly into the decision-making process.
- **Level IV (Very High Impact):** The decisions involve permanent, irreversible impacts on individuals or society. At this level, the final decision must be made

by a human. The algorithm may only serve as a supportive recommendation tool.

This tiered structure provides a definitive answer to the fear of full automation in critical public services. By adopting a similar matrix, the Portuguese Public Administration could clearly delineate where full algorithmic automation is legally permissible, saving administrative time, and where human sovereignty must remain absolute (protecting constitutional rights and ensuring due process).

3.5. The UK's Data Ethics Framework and ATRS: Tools for Technical Teams

While the Canadian Directive offers a "top-down" governance guidance for public managers, the successful implementation of AI also requires a "bottom-up" approach tailored for the data scientists, procurement officers, and project managers who actively build and deploy these systems. The United Kingdom provides the main benchmark for this technical tier through two complementary instruments: the Data Ethics Framework and the Algorithmic Transparency Recording Standard.

3.5.1. *Quantifying Ethics: The Scorecard System and Escalation Mechanisms*

The UK Government Digital Service [58] bridges the gap between abstract ethical principles and concrete project management. A standout feature of this framework is the implementation of self-assessment scorecards, as illustrated in Figure 3.7. Public sector teams are required to evaluate and score their projects on a scale of 0 to 5 across overarching principles, such as Transparency, Accountability, and Fairness, as well as specific operational actions.

This mechanism is applied repeatedly throughout the project lifecycle (Start, During, and After). To understand the practical rigor of this tool, it is helpful to analyze the first mandatory action of the framework, illustrated in Figure 3.6: Define public benefit and user need. Rather than serving as a generic compliance checkbox, the framework forces technical teams to actively map the societal footprint of their algorithm. On the scorecard, a score of 0 indicates that the public benefit is undefined, while a 5 requires that the need is documented, justified, and universally understood by all team members.

To achieve a high score, developers must answer targeted, uncomfortable questions during the design phase, such as:

- *Fairness:* What specific groups could be denied opportunities or face negative consequences because of the project?
- *Accountability:* How can the team demonstrate the value for money of your project?
- *Transparency:* Where can the team publish information on how the project delivers positive social outcomes for the public?

Crucially, the framework establishes a quantitative threshold for governance. If a project scores 3 or less in any of the principles, it triggers a mandatory escalation mechanism. The technical team cannot simply proceed, they are required to pause, document

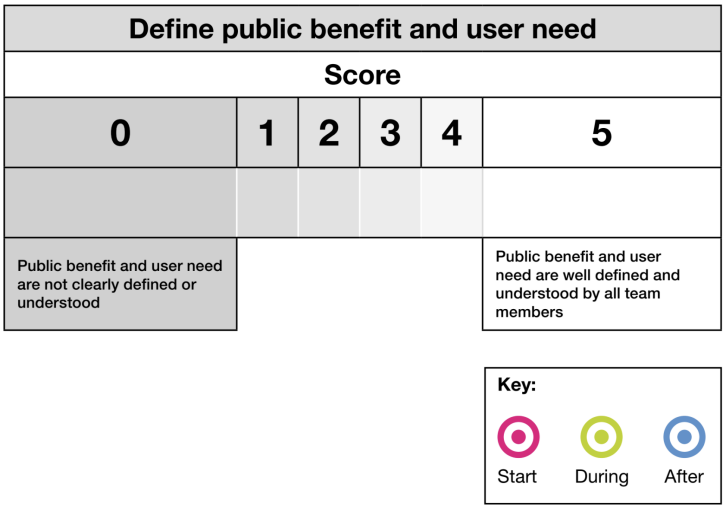


FIGURE 3.6. First Action Scorecard (Source: UK Government Digital Service [58]).

the reasons for the low score, and consult with an organizational ethics board or data ethics lead to address the system's ethical standards. This translates philosophical debates into quantifiable engineering metrics, ensuring that "ethics by design" becomes a strict and auditable criterion in project management.

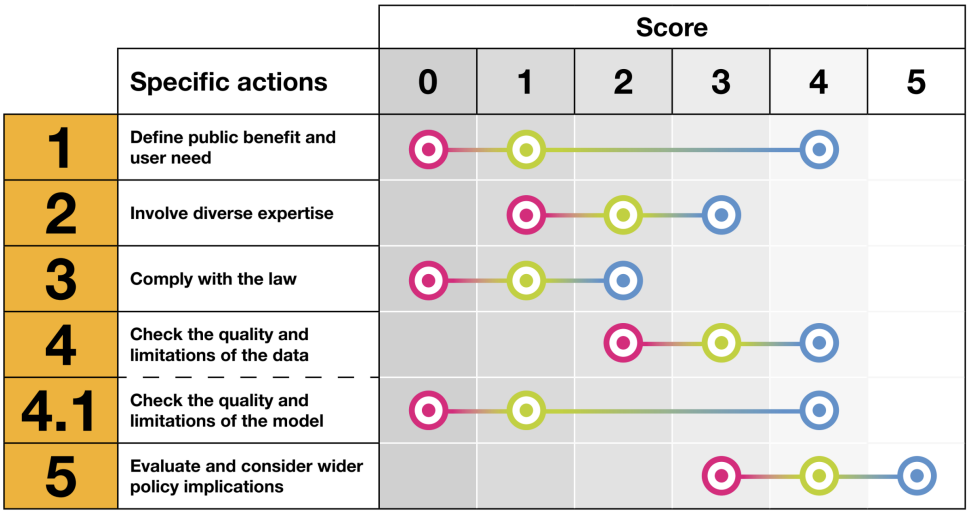


FIGURE 3.7. Example of the Self-Assessment Scorecard (Source: UK Government Digital Service [58]).

3.5.2. *Standardization: The Algorithmic Transparency Recording Standard*

Furthermore, the UK has recognized that internal self-assessment must be coupled with standardized public disclosure. To prevent the "performative transparency" and fragmentation previously discussed, the UK developed the Algorithmic Transparency Recording Standard (ATRS). This standard mandates that public sector organizations transparently disclose detailed information about their use of algorithmic methods, specifically targeting tools that significantly impact public decisions or directly engage with citizens.

As highlighted in the OECD and UNESCO [54], the use of the ATRS was made strictly mandatory for all UK government departments in February 2024, with plans for a gradual extension to the broader public sector.

This dual approach, internal scoring with escalation triggers, combined with mandatory, standardized external reporting, is highly instructive for the Portuguese Public Administration. Currently, Portugal risks a fragmented scenario where each ministry or municipality might document its algorithms differently, or worse, not document them at all. By adopting a national equivalent to the ATRS, coupled with ethical scorecards for technical teams, Portugal could enforce a uniform standard across the entire public sector. This would not only guarantee the "meaningful transparency" required to satisfy the CPA but also seamlessly integrate the stringent legal requirements of the EU AI Act into the daily workflows of Portuguese civil public servants.

3.6. The NIST AI Risk Management Framework: The Architecture of Trust

The United States, through the US National Institute of Standards and Technology [59], offers the most robust methodological architecture for global AI risk management, even though Canada and the United Kingdom provide excellent models of administrative governance and ethical project management.

The AI RMF 1.0's inclusion in this comparative analysis, is essential because it fills a critical gap left by the EU AI Act: the technical implementation of "trustworthiness." The NIST framework clearly acknowledges that AI risks are socio-technical by nature, which means software updates alone cannot manage them. They necessitate a thorough comprehension of the ways in which the system interacts with social dynamics and human behavior.

3.6.1. *Deconstructing Trust: Characteristics and Tradeoffs*

The NIST framework's greatest value for the Public Administration, which often deals with ambiguous concepts like "public interest", is its ability to break down the abstract idea of trust into seven quantifiable, socio-technical characteristics, as detailed in Figure 3.8:

(1) **Valid & Reliable:** The system must perform as intended, free from critical failure, and maintain its performance over its entire lifetime. Crucially, the NIST framework emphasizes that accuracy alone is insufficient, a trustworthy system must demonstrate robustness, meaning it minimizes harm even when operating in unexpected settings or processing unfamiliar data.

(2) **Safe:** The system must not endanger human life, health, property, or the environment under defined conditions. For public entities, this requires rigorous in-domain testing, the technical capacity for real-time monitoring and the ability to safely shut down or override the system if it deviates from expected functionality.

(3) **Secure & Resilient:** The system must withstand unexpected adverse events and adversarial attacks (e.g., data poisoning or model exfiltration). A resilient public

AI system is one that can maintain its core functions in the face of external change or, if necessary, degrade safely and gracefully rather than collapsing entirely.

(4) **Accountable & Transparent:** Transparency refers to the availability of information across the entire AI lifecycle, from training data provenance to the structure of the model. Accountability requires explicit governing structures for harm reduction, ensuring that responsibility mechanisms are tailored to the role of each actor within the system.

(5) **Explainable & Interpretable:** The framework establishes a vital conceptual distinction between these two elements. Explainability answers the question of *how* a decision was made by revealing the underlying mechanisms. Interpretability answers *why* that decision was made and what it means in its functional context. Together, they provide the technical foundation for public entities to fulfill the administrative duty to state reasons.

(6) **Privacy-Enhanced:** The system must safeguard human autonomy, identity, and dignity. This characteristic mandates the incorporation of Privacy-enhancing technologies (PETs) and data-minimization methods, ensuring that the State's use of AI does not lead to unwarranted surveillance or unauthorized inference of sensitive citizen data.

(7) **Fair - With Harmful Bias Managed:** Fairness transcends mere demographic balance. NIST requires the management of three major categories of bias: systemic (arising from institutional norms), computational (arising from unrepresentative data), and human-cognitive (arising from how humans perceive the AI output). Mitigating these biases is essential to prevent the State from amplifying societal inequalities at scale.

Importantly, the US National Institute of Standards and Technology [59] cautions that focusing on these characteristics separately will not guarantee trustworthiness because they do not function independently. In practice, public managers usually have to make difficult tradeoffs. For instance, optimizing explainability may oversimplify reality and compromise resilience against hostile cyberattacks; improving privacy through data anonymization or sparsity techniques may significantly lower the model's predictive accuracy. This suggests that a "perfect" AI system does not exist. However, for the Portuguese government, ethics and legality can be attained through deliberate governance, where tradeoffs between these seven attributes are clearly recorded, contextually justified, and openly shared with the public.

3.6.2. *The Core of the Framework: Four Functions for Public Engineering*

Moving away from static, linear checklists, the US National Institute of Standards and Technology [59] proposes a continuous, iterative lifecycle based on four central functions ("The Core"). The framework emphasizes that in order to identify blind spots and emergent risks, these functions must be executed by diverse, multidisciplinary teams, offering a

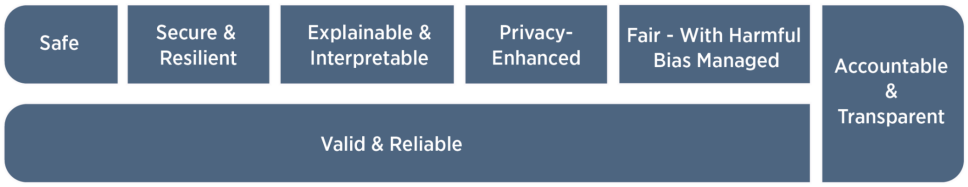


FIGURE 3.8. Characteristics of trustworthy AI systems (Source: US National Institute of Standards and Technology [59]).

clear roadmap to fill the operational void identified in this dissertation. The core functions of this lifecycle are illustrated in Figure 3.9.

- **GOVERN:** Acting as the cross-cutting foundation of the entire framework, this function dictates that risk management must be cultivated as an institutional culture. It defines roles, responsibilities, and risk tolerances before starting any project. For the Portuguese PA, this mandates defining explicit accountability structures and risk tolerances at the executive leadership level, ensuring that algorithmic governance is driven by senior management before any data is processed.
- **MAP:** This function establishes the socio-technical context. NIST emphasizes that AI risks are highly dependent on their deployment environment. The best technical intentions can be completely undermined by complex interdependencies in real-world settings. This phase forces public entities to engage diverse stakeholders to assess intended benefits, system limitations, and potential impacts on vulnerable communities. Crucially, the MAP function culminates in an initial "go/no-go" decision, preventing the blind deployment of "off-the-shelf" algorithms if the contextual risks are deemed unmanageable.
- **MEASURE:** Once mapped, risks must be rigorously quantified and tracked. This function employs quantitative, qualitative, and mixed-method tools to analyze system performance against the trustworthiness characteristics. It establishes robust Test, Evaluation, Verification, and Validation (TEVV) processes. To mitigate internal, human-cognitive biases and conflicts of interest within the Administration, NIST strongly advocates for independent reviews during this phase, which should be conducted in an open and transparent process. Ultimately, MEASURE provides the traceable, scientific data necessary to justify the ethical tradeoffs identified previously.
- **MANAGE:** The active execution phase, where public resources are systematically allocated to treat and mitigate the risks mapped and measured in the previous steps. It encompasses the creation of robust incident response, recovery, and communication plans.



FIGURE 3.9. The Core Functions of the AI RMF (Source: US National Institute of Standards and Technology [59]).

3.6.3. *The Imperative of Test, Evaluation, Verification, and Validation*

A vital contribution of NIST to the legal certainty of the Public Administration is the formalization of TEVV processes. The document clarifies that validation must not occur solely during the design phase but must be continuous throughout operation to detect phenomena such as "model drift" (performance degradation over time).

For Portugal, where public procurement traditionally views software as a finished product, adopting the NIST TEVV standard requires a paradigm shift: acquiring AI not as a static good, but as a continuous service of calibration, auditing, and monitoring.

3.7. The AIOLIA Project: AI Ethics in Practice

While the frameworks examined thus far were developed by regulatory bodies, standards organizations, and national governments, a complementary body of operational evidence emerges from applied research funded directly by the European Commission. The AIOLIA[1] project was designed specifically to translate established AI ethics principles into practitioner-validated measures through a bottom-up co-creation process conducted across six European deployment contexts. Its Deliverable D3.1 consolidates this process into a portfolio of 175 practical measures (101 technical and 74 organizational) co-created directly with the engineers, legal specialists, and domain experts responsible for deploying AI in real operational settings [60].

3.7.1. *Key Findings: Ethics Principles in Practice*

The most significant finding of the co-creation process, and the one most consequential for the design of governance frameworks, concerns the relationship between abstract ethics

[1]The AIOLIA Project website: `https://aiolia.eu`

principles and their practical components. The AIOLIA process revealed that, "for AI researchers and developers, the status and meaning of ethics principles and the exact nature of components may be much more fluid than in the abstract AI ethics frameworks" [60]. In applied contexts, the same ethical consideration, like transparency, can function simultaneously as an overarching principle and as a sub-component of other principles such as non-bias or accountability. Similarly, a single component such as auditability was found to appear across multiple distinct ethics principles.

A second finding concerns the motivational framing that makes ethics operationalization work in practice. The co-creation process revealed that practitioners approach operationalization most productively when it is framed as risk management rather than as ethical compliance. Use case partners viewed "the operationalization of ethics as a vital strategy to manage risks," and conversations about ethics proved more constructive when using "the more familiar framing around risks to establish common ground" [60]. This behavioral insight suggests that effective AI governance frameworks must speak the language of quantifiable impact and risk, not merely of ethical aspiration.

3.7.2. *Cross-Cutting Practical Measures*

The synthesis of practical measures across the six use cases identified a set of oversight mechanisms applicable regardless of deployment domain. These cross-cutting measures are directly actionable for the construction of the operational guidelines proposed in Chapter 5. The D3.1 identifies five resource categories that emerged consistently as prerequisites for ethical AI deployment [60]:

- **Oversight mechanisms:** Human-in-the-loop controls must function as universal safeguards in high-risk contexts, complemented by continuous monitoring through dashboards, thresholds, and feedback loops to detect model drift and feed lessons into training and process updates.
- **Transparency and traceability systems:** End-to-end traceability must be supported by versioning, immutable logs, and change management processes that allow for independent audits, incident reconstruction, and accountability across the AI lifecycle. User-facing transparency must be operationalized through confidence indicators, reasoning traces, and explanations that enable citizens to understand and contest system outputs.
- **Full lifecycle governance:** Structured workflows, periodic reviews, and role clarity must be embedded through standard operating procedures, ensuring governance is a continuous process rather than a one-off pre-deployment checkpoint.
- **Trained and competent personnel:** Structured onboarding, certification, and mentorship ensure that human oversight remains substantive rather than nominal, particularly relevant in contexts where automation bias may cause reviewers to defer uncritically to algorithmic outputs.

- **Security and system integrity:** Encryption, access controls, and cryptographic artifacts protecting data and infrastructure constitute foundational technical prerequisites for any responsible deployment.

3.7.3. *Eight Recommendations for Meaningful Operationalization*

Synthesizing the lessons of the co-creation process across all six use cases, the AIOLIA D3.1 formulates eight recommendations for any organization seeking to meaningfully operationalise AI ethics [60]:

(1) **Plan for a collaborative, multidisciplinary, and durable process:** Successful operationalizations require strong, long-term collaboration between practitioner and academic partners, with continuous communication and feedback loops ensuring that theoretical and practical perspectives remain aligned and can be refined over time.

(2) **Involve the right people:** A recurring failure mode in operationalization processes is the inclusion of only a narrow set of experts typically legal or technical specialists. Broad inclusion of actual and potential end-users with varying levels of expertise ensures that concerns relating to usability, interpretability, and trustworthiness are captured.

(3) **Prioritize practicality and accessibility:** Practical measures are only effective if they are feasible within the specific operational context and if the individuals responsible for implementation understand both the measures and their rationale. The core audience for implementation, assessment, and auditing must be kept in mind throughout the formulation process.

(4) **Accommodate early-stage AI systems:** Operationalizations may be conducted when systems are still in an evolving stage, before deployment experience is available. In such cases, some measures will necessarily be marked as "planned" rather than observable, and should be explicitly labeled as forward-looking, subject to revision once the system is fully operational.

(5) **Account for sensitive and classified contexts:** Successful operationalization requires detailed information about AI capabilities, data, deployment conditions, and business processes. In industrial and security settings, such information is frequently sensitive or legally restricted, requiring careful consideration of access controls and confidentiality requirements throughout the operationalization process.

(6) **Maintain an open communication stance:** Ethics conversations risk becoming defensive when stakeholders interpret them as fault-finding exercises. A practice-oriented and solution-based communication approach supports constructive engagement and increases the likelihood that operationalization produces genuine improvement rather than performative compliance.

(7) **Safeguard transparency and auditability of the process itself:** Documentation must cover not only the outcomes of the operationalization but also

its process, including disagreements and challenges encountered. This process-level transparency supports accountability and enables decisions to be revisited if circumstances change.

(8) **Ensure sustainability through planned reviews:** AI systems and their operational conditions change over time, whether through technical updates, shifts in data distributions, or new regulatory requirements. Practical measures must therefore be reviewed at regular intervals throughout the deployment lifecycle to ensure continued relevance and effectiveness.

4. National Diagnosis: The Operational Gap in Portugal

Comparing international frameworks reveals a global ecosystem in which AI governance has evolved from abstract ethical principles to tangible operational control mechanisms. Portugal, on the other hand, is in a different stage of maturity, according to an analysis of the most recent national strategic documents, particularly the Estratégia Digital Nacional (EDN) and its 2025-2026 Action Plan [61]. The country is experiencing a strong acceleration in digitalization and infrastructure, yet it suffers from an "operational void" regarding the specific management of algorithmic risk.

4.1. Strategic Ambition: Digitalization vs. Risk Governance

With a particular focus on administrative modernization, the new Portuguese strategy is organized around four pillars: People, Business, Government, and Infrastructure. Promising is the initiative #8 ("Digital Journey for Public Administration,"), which offers models for assessing "digital maturity" and a definition of a digital transformation journey. However, upon closer inspection, it becomes clear that this maturity is more commonly associated with cybersecurity, interoperability, and procedural efficiency than algorithmic safety. While jurisdictions such as Canada require an AIA before a system operates, the Portuguese plan focuses on dematerialization. Consequently, there is a severe risk of digitizing decision-making processes without safeguarding, *ex ante*, against the risks of bias and opacity identified by the EU AI Act.

4.2. The Technological Push vs. The Governance Void

The strategic pillars that are most closely related to the topic of this dissertation are Initiative #10 ("National Artificial Intelligence Agenda") and Initiative #13 ("Data Available for Everyone"). A comprehensive review of these initiatives, however, shows a significant conflict between the desire for technological innovation and an absence of governance, with a significant reliance on infrastructure that expressly leaves out risk management procedures.

With the eventual objective of incorporating this approach into products like the *gov.pt* portal, Action 10.1 specifically requires the creation of the first Portuguese LLM, trained on national computing infrastructure and curated data. Concurrently, Initiative #13 aims to develop a "National Public Administration Data Model," while Actions 10.2 and 10.3 concentrate on building an "AI Factory".

Even though these measures show a great dedication to interoperability and national technological sovereignty, they exemplify the "Operational Gap" in two crucial ways:

- **Deployment without safeguards (The Methodology Gap):** Integrating an LLM into the *gov.pt* portal means generative AI will interact directly with citizens seeking public services. A concrete current example is the deployment of AI-powered virtual assistants on the *ePortugal* portal to guide citizens through administrative procedures.[2] While the Action Plan frequently cites "ethics" as a guiding principle, it lacks the tools for practical translation. It provides no equivalent of the Canadian AIA to assess the risk of algorithmic hallucinations, nor does it mandate a NIST-aligned TEVV process before the model goes live. The risk is that ethics remains a decorative value rather than a hard engineering requirement.
- **The Taxonomy and Transparency Gap:** Despite the push for a national data model, there is no provision for a National Algorithm Registry. In stark contrast to the recommendations of the Global Partnership on Artificial Intelligence (GPAI) and the UK's ATRS, the Portuguese Government is building powerful predictive and generative models without a standardized mechanism to explain their logic or data provenance to the public. Without knowing where AI is deployed (Inventory) and how it is classified (OECD Taxonomy), the Administration cannot govern what it does not know.

4.3. The Institutional Opportunity: The New Digital Agency

This dangerous asymmetry, building state-owned AI without state-owned AI governance, leads directly to Initiative #11. The 2025-2026 Action Plan mandates the creation of the *Agência Nacional para o Digital*, tasked with ensuring a coordinated strategic vision across the Public Administration [61].

This represents the critical institutional opportunity identified by this dissertation. For this Agency to function effectively and not merely as another bureaucratic layer, it requires operational instruments. Currently, the plan assigns it execution responsibilities but fails to define the technical standards of algorithmic auditing it must enforce over projects like the new Portuguese LLM. This is the core of the problem: the Government is building the technology (Initiative #10) and building the institution (Initiative #11), but the operational guidelines and standards of practice have yet to be formalized.

4.4. Comparative Synthesis: What is Missing in Portugal?

The diagnostic results are summarized in Table 3.1, which compares the reality described in the national strategic documents with the international benchmarks examined in this chapter.

[2]The AI virtual assistant is currently active on the national digital public services portal, available at: `https://eportugal.gov.pt/`. Accessed on June 11, 2026.

TABLE 3.1. Gap Analysis: International Benchmarks versus Portuguese Strategy

Governance Component	International Benchmark (Best Practice)	Portuguese Reality	The Identified "Gap"
Definition & Taxonomy	**OECD Framework:** Multidimensional classification of AI systems based on their socio-technical context and operational impact.	**Initiative #13:** Focus strictly on data catalogs and interoperability.	Lack of an official taxonomy to classify AI systems by risk and impact on rights.
Risk Control	**Canada (AIA):** Binding directive with four risk levels and mandatory human intervention.	**Initiative #8:** Generic evaluation of "digital maturity" and efficiency.	Absence of a mandatory Algorithmic Impact Assessment.
Transparency	**UK (ATRS) / GPAI:** Mandatory algorithmic recording standard (Active Transparency).	**Initiative #15:** Monitoring of general ecosystem indicators (Digital Blueprint).	Lack of a centralized, public repository specific to algorithmic systems.
Engineering & Ethics	**UK (Scorecard) / NIST (AI RMF):** Ethical scorecards and continuous TEVV processes.	**General Principles:** Ethics cited as an abstract strategic value, without operational control mechanisms or actionable validation protocols.	Absence of objective engineering standards to validate "Trustworthiness" during public procurement.
Ethics Operationalization	**AIOLIA (D3.1):** 175 practitioner-validated measures and 8 recommendations co-created across six European deployment contexts.	**No equivalent:** Portugal has no co-creation process or practitioner-validated measures for public sector AI deployment.	Absence of empirically grounded operational measures calibrated for the specific realities of AI deployment in the Portuguese PA.

4.5. Conclusion: The Need for National Guidelines

Portugal possesses a rapidly growing infrastructure and the political will, as demonstrated by the National Digital Strategy. The governance "middleware" or the technical and procedural standards that direct the public manager on a daily basis, are crucially missing from it.

Without these tools, the Public Administration runs the risk of being imprudent (deploying "black boxes" without accountability controls) or paralyzed (not innovating for fear of violating the EU AI Act). This gap will be precisely filled by the Guidelines proposed in Chapter 5, which will give public leaders and the future National Agency for Digital the "instruction manual" they need to implement AI in a safe, lawful, and effective manner.

CHAPTER 4

Methodology for Guidelines Development

This chapter describes the methodological approach adopted to transform the theoretical findings and international benchmarks established in the preceding chapters into a practically deployable governance artifact. It explains how the proposed guidelines were designed and defines the evaluation criteria and demonstration methodology within the Portuguese Public Administration.

1. Research Approach: Design Science Research

This dissertation is based on the DSR methodology because it enables the rigorous creation and *ex ante* evaluation of artifacts designed to solve problems. The DSR framework is widely used in research in Information Systems and Computer Science. It focuses on designing and testing new IT artifacts intended to address organizational problems that have been found in organizations.

Unlike purely observational methodologies, which synthesize existing knowledge without producing an actionable intervention, DSR explicitly bridges the gap between theory and practice and tries to come up with specific answers to complicated social and technical problems. Equally, DSR was preferred over Action Research, which, while well-suited to iterative, participatory problem-solving within a real organizational client, requires sustained collaborative access to a participating institution over an extended period, a condition beyond the temporal and resource constraints of a master's research project. DSR, by contrast, enables rigorous design and *ex ante* evaluation without mandating live deployment, a recognized strength of the methodology for producing governance artifacts in emergent regulatory domains.

In the context of this study, the proposed artifact comprises a set of operational and practical guidelines specifically designed for the implementation of AI in the Portuguese PA. To maintain scientific rigor and reproducibility, this research adheres strictly to the six-phase DSR Process Model proposed by Peffers et al. [5], as illustrated in Figure 1.1 (Chapter 1). The following sections explain how to carry out these steps.

2. Phases 1 & 2: Problem Identification and Objectives Definition

The initial phases of the DSR methodology are designed to define the specific research problem and infer the objectives of a potential solution. In this dissertation, these phases were thoroughly addressed through the systematic review and the comparative framework analysis conducted in the preceding chapters.

Phase 1, focusing on Problem Identification and Motivation, was extensively covered in Chapter 2. Through an SLR following the PRISMA protocol, the study identified

a critical "Accountability Gap" in the public sector. The literature demonstrated that the deployment of complex, opaque ML models (often referred to as "black boxes") directly conflicts with the foundational principles of the CPA, specifically the strict duty to state reasons for administrative acts mandated by Articles 152 and 153. This collision between technological capability and constitutional guarantees provided the fundamental motivation for this research.

Building upon this problem statement, Phase 2, which aims to define the Objectives of a Solution, was addressed in Chapter 3. By comparing international best practices, including the EU AI Act, OECD principles, the Canadian Directive, the UK ATRS, the NIST AI RMF, and the practitioner-validated operational measures and recommendations produced by the AIOLIA project, with the current Portuguese digital strategy, an "Operational Gap" was clearly diagnosed (as summarized in Table 3.1). Consequently, the primary objective of the proposed artifact was defined: to create a governance "middleware" that effectively translates abstract regulatory requirements (Hard Law) and ethical principles (Soft Law) into measurable engineering and management requirements specifically tailored for Portuguese public entities.

3. Phase 3: Design and Development

The third phase constitutes the core creative process of developing the artifact. The proposed guidelines, which will be fully presented in Chapter 5, were developed through a deliberate synthesis of established international frameworks analyzed in Chapter 3, carefully calibrated for the Portuguese institutional and legal reality. To ensure a comprehensive governance model, the design of the guidelines followed a tripartite architecture.

The first dimension is related to Taxonomy and Risk Classification. To address the specific regulatory deficiencies identified in the Portuguese Public Administration, this dimension proposes a hybrid Algorithmic Impact Assessment specifically designed for Portugal (AIA-PT). The guidelines are based on the multidimensional ethical principles of the OECD, the practical algorithmic impact assessment of the Canadian Directive on ADM, and the strict compliance tiers of the EU AI Act. They are organized around a complete risk-tier system (Levels I to IV).

The second dimension addresses Engineering and Continuous Validation. To mitigate the unpredictable and changing nature of AI systems, which can lead to risks like model drift, data degradation, and algorithmic bias long after they are put into use, the artifact uses ideas from the NIST AI Risk Management Framework. This section turns the strict need for continuous TEVV and the NIST core functions (Govern, Map, Measure, Manage) into actionable procedural requirements for public procurement officers and IT project teams. This is because the Portuguese Public Administration often relies on outside vendors for technological solutions. The guidelines aim to prevent the procurement of algorithmically opaque systems by including these technical requirements directly in procurement contracts. They also require XAI mechanisms and ensure that the state

is not held responsible for algorithmic failures. These engineering requirements are further grounded in the cross-cutting practical measures identified by the AIOLIA project, whose co-creation process across six deployment contexts confirmed that full lifecycle governance with continuous monitoring constitute universal prerequisites for responsible AI deployment, regardless of sector or system complexity.

Lastly, the third dimension operationalises institutional governance and citizen accountability mechanisms. This dimension goes beyond internal technical validation to ensure "ethics by design". This is achieved through the proactive transparency requirements of the UK's ATRS and the OECD principles of human-centric values. The guidelines set up clear ways for algorithmic transparency, human oversight, and citizen redress to fight the widespread problems of automation bias and the breakdown of due process. This aspect actively fills the accountability gap found in the literature by requiring multidisciplinary teams to keep detailed records of their design choices, data sources, and risk tradeoffs. The design of this dimension is based on the AIOLIA project's recommendations for meaningful operationalization, specifically for technical specialists, to safeguard the transparency and auditability of the governance process itself and to ensure sustainability through planned reviews as systems and regulatory conditions evolve. In the end, it ensures that public sector AI deployments meet the strict administrative requirement for explainability, which is required by Portuguese administrative law.

4. Phases 4 & 5: Demonstration and Evaluation

Verifying the usefulness, effectiveness, and validity of the produced artifact is a crucial component of the DSR methodology. This research employs an *ex ante* qualitative evaluation strategy based on Theoretical Scenarios (Scenario-Based Evaluation) because implementing a new algorithmic governance framework throughout the Portuguese State is beyond the scope of a master's dissertation.

The artifact will go through Phases 4 (Demonstration) and 5 (Evaluation) in Chapter 6 by simulating its application across four distinct public sector scenarios and will be assessed against three primary evaluation criteria. First, Legal Resolution evaluates whether the guidelines demonstrably resolve the legal-technical deadlock between opaque algorithmic systems and Articles 152 and 153 of the CPA, specifically by ensuring that a legally valid *iter cognoscitivo* can be extracted for each deployment scenario. Second, Institutional Accountability examines whether the framework eliminates the current accountability gap by clearly assigning legal and political responsibility for algorithmic operations at a hierarchical level proportionate to the involved risk. Finally, Regulatory Robustness determines whether the artifact exceeds the baseline established by the EU AI Act by actively capturing high-stakes sovereign functions that fall outside the Act's predefined high-risk categories.

While the first three scenarios represent high-stakes domains mapped during the SLR, a fourth scenario was deliberately introduced to test the framework's agility and proportionality in low-risk contexts:

- **Public Health Scenario:** The deployment of an AI-driven predictive triage system in a Serviço Nacional de Saúde (SNS) – the Portuguese Government health service – Emergency Department.
- **Justice Scenario:** The implementation of a recidivism risk assessment algorithm by the *Direção Geral de Reinserção e Serviços Prisionais*, the Portuguese reintegration and prison services.
- **Public Finance Scenario:** The use of unsupervised ML models by the *Autoridade Tributária* – Portuguese tax authority – to detect complex tax evasion networks.
- **Local Administration Scenario:** The deployment of an NLP-driven municipal FAQ chatbot, serving as a baseline to demonstrate the framework's capacity to facilitate innovation without inducing bureaucratic asphyxiation.

The guidelines will be applied step-by-step for every hypothetical situation. The assessment will critically determine if adhering to the suggested artifact effectively demands the public entity to record its algorithmic decisions, reduce bias, and preserve human sovereignty over crucial choices. The ultimate goal of this scenario-based walkthrough is to demonstrate that the guidelines guarantee adherence to the EU AI Act and end the legal-technical impasse imposed by the CPA.

5. Phase 6: Communication

The final phase of the Peffers et al. [5] DSR model requires the communication of the identified problem, the designed artifact, and its evaluation to the relevant scientific and professional communities. This phase is implemented through two complementary outputs. The first is this dissertation, which is the primary scholarly artifact of the research cycle and provides the complete theoretical, methodological, and evaluative account of the AIA-PT framework.

The second output is a peer-reviewed research article, which presents a condensed version of the framework and its scenario-based evaluation to the broader academic and policy community. Together, these two outputs complete the DSR communication cycle by disseminating the AIA-PT to both the academic community and the public sector practitioners for whom the framework is designed.

CHAPTER 5

Guidelines for Portuguese Public Administration

1. Introduction and Architecture of the Guidelines

As determined by the diagnostic in Chapter 2 and Chapter 3, there is a critical "Operational Gap." Although the Portuguese national strategic ambition is clearly high, as demonstrated by the Estratégia Digital Nacional and the approaching implementation of the *Agência Nacional para o Digital* (Initiative #11), public entities presently lack the standardized, practical tools necessary to balance the deployment of complex AI systems with the rigorous administrative guarantees required by Portuguese law. In particular, there is a legal-technical deadlock regarding the obligation under Articles 152 and 153 of the CPA to provide the rationales (*dever de fundamentação*).

The main product of this Design Science Research is presented in this chapter as a comprehensive set of practical guidelines intended to function as a governance "middleware" in order to break this impasse. These guidelines are specifically designed to be used by public managers and adopted by the upcoming *Agência Nacional para o Digital*. Three interrelated pillars constitute the framework's structure, as illustrated in Figure 5.1:

- **Pillar I: The Portuguese Algorithmic Impact Assessment (AIA-PT)**, establishing a four-tier risk classification system and a required socio-technical taxonomy.
- **Pillar II: Engineering Requirements and Continuous TEVV**, defines technical standards for public procurement and lifecycle monitoring, deeply established in the NIST AI Risk Management Framework.
- **Pillar III: Team Ethics and Proactive Transparency**, setting external reporting standards and internal self-assessment scorecards to guarantee citizen accountability.

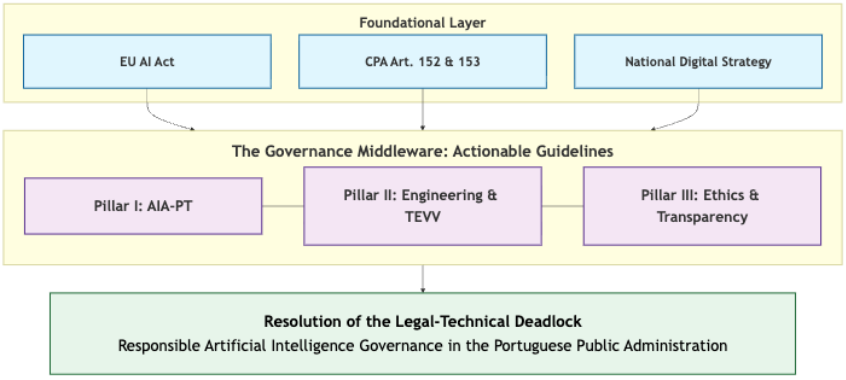


FIGURE 5.1. High-level architecture of the proposed Guidelines

2. Pillar I: The Portuguese Algorithmic Impact Assessment (AIA-PT)

Before a Portuguese public entity can use any AI system, whether it was developed internally or procured through public procurement, it must go through the AIA-PT. The AIA-PT establishes a required two-step *ex ante* evaluation based on the algorithmic impact metrics of the Canadian Directive on Automated Decision-Making and the multidimensional OECD Classification Framework. This is meant to assist with administrative discretion. The complete workflow of this evaluation process is represented in Figure 5.2.

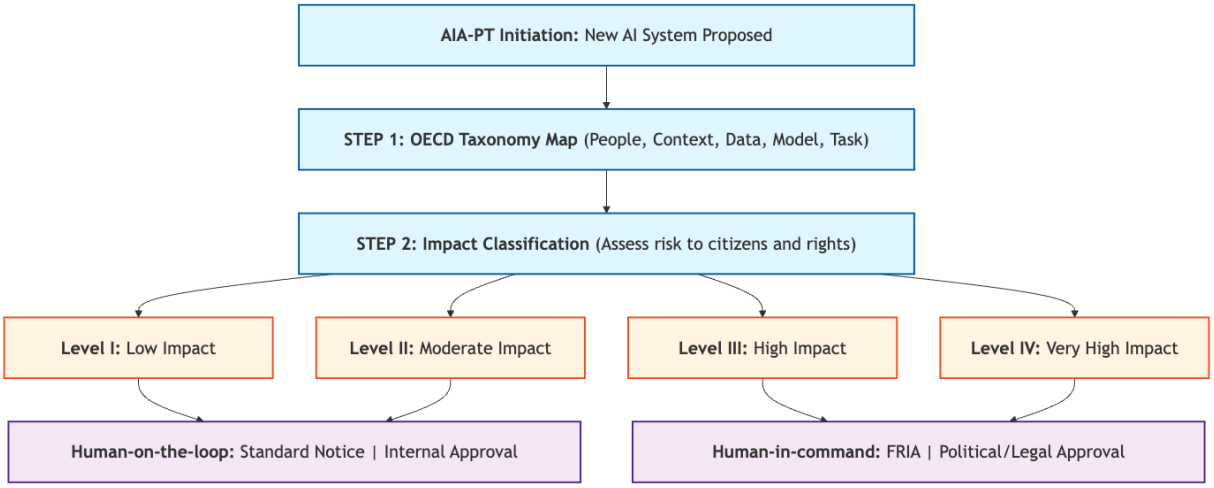


FIGURE 5.2. The AIA-PT workflow: Taxonomic Mapping, Risk Classification, and Mitigation Requirements.

2.1. Guideline 1.1: Standardized System Taxonomy

The deploying institution must systematically document the system across the five dimensions of the OECD framework to ensure that the public entity fully understands the socio-technical context of the proposed tool:

(1) **People and Planet:** This dimension evaluates the system's capacity to support human-centered and reliable AI. It does not just identify the direct users (e.g., public employees), it also includes a wider mapping of everyone and passive stakeholders who use or are affected by the applied AI system. For the public sector, it is imperative that the system's social footprint be rigorously assessed, including how it affects basic human rights, the well-being of citizens, the environment, and the job market.

(2) **Economic Context:** This dimension shows the exact sectoral and organizational context in which the AI system works. In the field of Public Administration, it is necessary to specify the particular public service area (for example, healthcare triage, criminal justice, or tax fraud detection) and the administrative role that the system serves. It is important because it tells decision-makers to evaluate the "critical nature" of the deployment, which means assessing whether a system failure would affect important public services and the magnitude of the resulting impact.

(3) **Data and Input:** This dimension examines the data and expert inputs used to construct the system's computational representation of the environment, with an emphasis on the raw materials of algorithmic decision-making. It requires a thorough examination of data provenance, collection techniques, and structural formats. The difference between the data used to train the system ("in the lab") and the live data it processes during real-world operation ("in the field") is a crucial contribution. This thorough data profiling is crucial for public organizations in order to identify past biases at the source and ensure compliance with privacy regulations.

(4) **AI Model:** The computational core, which represents the external environment and internal decision-making logic, is disassembled in this dimension. Public managers must categorize the system's technical architecture, differentiating between models created using data-driven ML, models created using explicitly programmed expert knowledge and hybrid approaches. The system's explainability, which determines how its performance goals can be audited by regulatory bodies and how its logic can be made understandable to the citizens it affects, is closely linked to an understanding of this underlying architecture.

(5) **Task and Output:** The last dimension evaluates the system's operational execution, its functional outputs. It categorizes the specific tasks performed by the algorithm. Above all, it assesses the system's "action autonomy," determining whether the AI merely recommends an action to a human official or independently executes an administrative decision. Establishing suitable performance metrics and determining the required degree of human oversight depend on this assessment.

2.2. Guideline 1.2: The Algorithmic Impact Scoring System and Four-Tier Matrix

To eliminate personal bias from risk assessment, the AI system must be given a numerical score based on the taxonomic mapping done in Guideline 1.1, public managers must give each of the five OECD dimensions a score from 1 (Lowest Risk) to 4 (Highest Risk). This method aims to ensure that the final classification explicitly and consistently defines the procedural duties of the public entity.

By summing the points across all five dimensions, the system generates a Total Impact Score (ranging from 5 to 20). This score determines the system's placement within the Four-Tier Impact Matrix. The proposed risk matrix distributes the total scoring range (5 to 20 points) equally into four exact quartiles of 4 values each. This uniform distribution ensures that no arbitrary weight is assigned to one level at the expense of another, as detailed in the rubric presented in Table 5.1:

Level I (Low Impact) - Score 5 to 8 points: The automated decision has minimal to none or easily reversible impacts on citizens' rights, equality, dignity, privacy, economic interests, and individual autonomy. The data used by the system pose

TABLE 5.1. AIA-PT Risk Scoring Rubric

Dimension	Low - 1 Point	Moderate - 2 Points	High - 3 Points	Very High - 4 Points
People & Planet	Minimal / reversible impact. No vulnerable groups.	Moderate, short-term impact on administrative rights.	High, difficult to reverse impact on fundamental rights.	Irreversible impact on human dignity, liberty, or life.
Economic Context	Internal back-office administration.	General public service delivery.	Essential services.	Sovereign high-stakes functions.
Data & Input	Open, anonymized, or non-personal data.	Structured, non-sensitive personal data.	Sensitive data or unstructured data.	Highly classified, incomplete, or scraped data.
AI Model	Deterministic, rule-based systems.	Simple and interpretable ML models (e.g., Logistic Regression, Decision Trees).	Complex models applying validated post-hoc XAI techniques (e.g., Random Forests, Boosted Trees, Neural Networks).	Complex models with absent post-hoc XAI techniques or continuous learning models.
Task & Output	Information sorting only.	Drafts recommendations for human review.	Makes decisions with nominal human oversight.	Fully autonomous execution without human intervention.

low risk (e.g., non-personal or unclassified information), and the operational context itself is also low risk. *Example:* An internal AI tool for sorting generic public inquiry emails or IT helpdesk chatbots.

Level II (Moderate Impact) - Score 9 to 12 points: The automated decision has moderate, likely reversible, and short-term impacts on fundamental rights, equality, and economic interests. The data utilized present moderate risks (e.g., personal, unclassified or protected information), and the operational context implies a moderate level of risk. *Example:* An algorithm optimizing the routing of municipal waste collection trucks or scheduling administrative appointments.

Level III (High Impact) - Score 13 to 16 points: The automated decision has high, difficult to reverse, and potentially ongoing impacts on individuals' rights, health, well-being, or economic interests. The system relies on high-risk data (e.g., sensitive personal information or unstructured data) and operates in an inherently high-risk context. *Example:* Predictive triage systems in public hospitals (Health) or algorithms filtering applications for social housing (Finance).

Level IV (Very High Impact) - Score 17 to 20 points: The automated decision has very high, irreversible, and perpetual impacts on fundamental rights, equality, dignity, or human autonomy. The data involved present very high levels of risk (e.g., highly sensitive, classified, or incomplete data), and the system operates in a critically high-risk context. *Example:* AI risk assessment tools for criminal recidivism (Justice) or automated models predicting complex tax evasion leading to immediate asset freezing (Finance).

The Legal Safeguard: The EU AI Act Override Rule and Pre-Screening Filter

To prevent administrative redundancy and ensure procedural efficiency, the scoring system must not override supranational legal obligations. For that reason, the AIA-PT establishes the EU AI Act as a mandatory *ex ante* pre-screening filter. This filter is applied after the taxonomic mapping (detailed in Section 2.1) and prior to calculating the algorithmic impact score in Guideline 1.2.

If a proposed AI system falls under the prohibited practices (Article 5) of the EU AI Act, its procurement or deployment is immediately blocked. Furthermore, if the system aligns with any of the "High-Risk" use cases outlined in Annex III of the EU AI Act, the scoring process is entirely bypassed, and the system is automatically classified under Level IV protocol.

Only when an AI application does not trigger these explicit European "Hard Law" thresholds should public workers proceed to evaluate the system using the AIA-PT scoring matrix. This sequential workflow ensures that the Portuguese government can optimize administrative resources by avoiding unnecessary evaluations for systems that are already classified by supranational law, while quickly operating lower-risk applications in full regulatory compliance.

2.3. Guideline 1.3: The Human Oversight Rule for CPA Compliance

The AIA-PT creates mandatory levels of human oversight that are intrinsically connected to the established Impact Levels in order to address the direct conflict with the human-centric guarantees of Articles 152 and 153 of the CPA. The legal parameters of "Hybrid Administration" in Portugal are effectively defined by this guideline:

- **For Levels I and II (Human-on-the-loop)** If continuous quality assurance and combined oversight mechanisms are maintained, the AI system may carry out automated decisions without direct, case-by-case human involvement. Under Portuguese Administrative Law, this delegation of direct human validation is strictly limited to bound administrative acts *(atos administrativos vinculados)*. In these situations, the law is very clear about what the premises and consequences are, so there is no room for administrative discretion, personal value judgments, or weighing conflicting public interests. By restricting autonomous execution solely to bound acts, the public entity preserves the legal guarantees of the CPA while making operations more efficient.
- **For Levels III and IV (Human-in-command)** The final administrative decision *must* be made by a human public worker. The algorithm is restricted to an auxiliary role (decision support). Human operators must actively review the system's recommendations, retain the absolute legal authority to override them, and bear sole responsibility for providing the legal justification (*iter cognoscitivo*) to the citizen.

2.4. Guideline 1.4: Procedural Escalation and Approval Matrix

The Four-Tier Matrix classification is not merely descriptive but it also acts as a rigorous procedural trigger. Adapting the operational rigor of the Canadian Directive on Automated Decision-Making to the Portuguese legal and institutional landscape, the AIA-PT imposes an escalating set of procedural requirements. As the risk to citizens' rights increases, so does the administrative burden of due diligence, transparency, and institutional approval.

Notice and Explanation (The Duty to State Reasons)

The foundation of administrative law is transparency. To guarantee adherence to the right to information and the obligation to state reasons (*dever de fundamentação*), the AIA-PT sets detailed requirements:

Levels I and II: All delivery channels in use (internet, in person, mail or telephone) must post a plain-language notice from the public entity indicating that the decision is supported by an AI system. The public must have access to a general description of the system's operation, the data it processes, and its function in the administrative workflow.

Levels III and IV: Any decision that results in the denial of a public service, benefit, or regulatory action must be accompanied by a personalized, meaningful explanation in addition to general notices. This explanation should detail the essential factors that influenced the system's output while also emphasizing the timely options available to citizens for contesting the automated decision.

Fundamental Rights Impact Assessment (FRIA)

Replacing the Canadian requirement for a Gender-Based Analysis Plus (GBA+), the AIA-PT tightly integrates the European regulatory framework by mandating an Fundamental Rights Impact Assessment (FRIA). As established in Article 27 of the EU AI Act, this acknowledges the unique power asymmetry between the State and the citizen.

Levels I and II: A simplified impact mapping is required. While a full FRIA is not legally mandated for lower-risk tiers under the EU AI Act, this proactive step ensures that the system does not inadvertently rely on discriminatory proxies or violate the GDPR principles of data minimization and purpose limitation.

Levels III and IV: A comprehensive FRIA is strictly mandatory prior to deployment, aligning with the obligations for high-risk systems. The public entity that is deploying the system must map the system's specific temporal and procedural context, explicitly identifying vulnerable groups that could be negatively impacted. By methodically recording mitigation techniques for algorithmic bias and socioeconomic discrimination, the State is forced to show that the promised efficiency gains do not come at the expense of constitutional rights.

Institutional Approval to Operate (Accountability)

Addressing the accountability gap identified in the literature, the AIA-PT clearly assigns legal and institutional responsibility for the system's deployment. Approval authority escalates hierarchically:

Level I: Approval is granted by a Middle Manager (e.g., *Diretor de Serviços* or *Chefe de Divisão*) or a specifically delegated Senior Technician (*Técnico Superior*) responsible for the execution of the administrative procedure.

Level II: Approval requires authorization from the maximum governing body of the respective public institution (e.g., the Board of Directors of a public institute, the President of a Municipality, or the Rector of a Public University).

Level III: Given the high risk to fundamental rights, deployment can only be authorized at the political level by the respective Minister or Member of Government holding the portfolio (e.g., Minister of Health, Minister of Justice), assuming ultimate political accountability.

Level IV: Due to the extreme, perpetual, and potentially irreversible impacts on human liberty and equality, often bordering on prohibited practices under the EU AI Act, these systems cannot be deployed via standard administrative procedures. Approval strictly requires a mandate from the Council of Ministers (*Conselho de Ministros*) or specific authorizing legislation passed by the Portuguese Parliament (*Assembleia da República*).

A summary of these procedural escalation requirements is provided in Table 5.2.

TABLE 5.2. AIA-PT Procedural Escalation Matrix by Impact Level

Impact Level	**Human Oversight (CPA)**	**Transparency & FRIA**	**Institutional Approval**
Level I (Low)	Human-on-the-loop (Aggregated oversight)	General Notice & Simplified Mapping	*Diretor de Serviços* / *Técnico Superior*
Level II (Moderate)	Human-on-the-loop (Aggregated oversight)	General Notice & Simplified Mapping	Maximum Governing Body (e.g., Rector, Board)
Level III (High)	Human-in-command (Active review required)	Meaningful Explanation & Comprehensive FRIA	Minister / Portfolio Member of Government
Level IV (Very High)	Human-in-command (Absolute override)	Meaningful Explanation & Comprehensive FRIA	Council of Ministers / Parliament Legislation

3. Pillar II: Engineering Requirements and Continuous TEVV

Pillar II specifies *how* an AI system must be developed, purchased, and maintained, while Pillar I determines *when* it can be implemented based on its risk to citizens. Software is treated as a static, final product under Portugal's traditional public procurement paradigm, which is regulated by the Public Contracts Code (*Código dos Contratos Públicos*). However, due to their inherent dynamic nature, ML models are vulnerable to performance degradation over time, a phenomenon known as model drift, as their behavior changes as

they interact with real-world data in the field. Figure 5.3 illustrates this necessary shift from a static software procurement model to a continuous AI TEVV lifecycle.

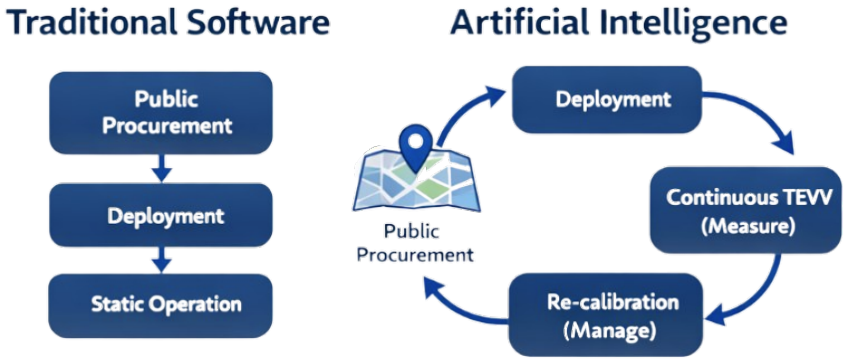


FIGURE 5.3. Shifting from a static software procurement model to a continuous AI TEVV lifecycle.

The following guidelines adapt the fundamental features of the NIST AI Risk Management Framework (Map, Measure, Manage) in order to address this technical reality and keep the State from acquiring opaque "black box" models and creating an accountability gap. Additionally, they incorporate the rigorous requirement for ongoing TEVV into the lifecycle of the Portuguese public sector.

3.1. Guideline 2.1: AI Public Procurement and the "Map" Function

In keeping with the recently created *Iniciativa #6* (*Polo Colaborativo para o Digital*) of the *Estratégia Digital Nacional*, which places a strong emphasis on "Public Procurement of Innovation," public organizations need to fundamentally revise the way they purchase AI solutions from third-party vendors. The contracting entity must systematically "Map" the operational context and demand verifiable proof of algorithmic robustness before completing any procurement contract (*Caderno de Encargos*) for AI systems, particularly for Levels II, III, and IV.

To enforce this, public entities must integrate the following normative requirements into their procurement documentation:

Requirement 2.1.1 - Data Provenance and Quality Testing: Vendors must transparently disclose the composition of the training datasets. The public entity must verify if the data is representative of the Portuguese demographic context to prevent imported historical bias (e.g., a triage algorithm trained on US healthcare data may not accurately reflect the realities of the Portuguese *SNS*).

Requirement 2.1.2 - Transparency and XAI Capabilities: To fulfill the CPA duty to state reasons (established in Pillar I), contracts must explicitly mandate Explainable AI (XAI) capabilities. Vendors cannot lock the State out of auditing the model's logic under the pretext of intellectual property.

Requirement 2.1.3 - Performance Metrics Baseline: Rather than allowing individual contracting entities to arbitrarily define acceptable performance, procurement contracts must enforce the minimum thresholds for accuracy, precision, recall, and false-positive rates explicitly dictated by the technical baselines of the upcoming *Agência Nacional para o Digital*. This centralization ensures a standardized, nationwide benchmark for algorithmic robustness prior to deployment.

A summary of these mandatory verification steps is provided in Table 5.3.

TABLE 5.3. AI Public Procurement Mandatory Requirements Checklist

Phase	Requirement Focus	Vendor Obligation	Public Entity Verification
Pre-Tender	Demographic Context (Data)	Disclose training data provenance and composition.	Verify alignment with Portuguese population.
Contracting	Explainability (XAI)	Incorporate interpretable logic and waive "black-box" intellectual property claims for audits.	Test if operators can extract an *iter cognoscitivo* for the CPA.
Delivery	Algorithmic Robustness	Deliver technical documentation with tested precision/recall rates.	Establish the baseline metrics required for future TEVV phases.

3.2. Guideline 2.2: Continuous TEVV Lifecycle and Drift Mitigation

The idea of "Trustworthy AI" is an ongoing engineering effort rather than a one-time certification. In order to dynamically "Measure" the continuous safety of all active AI systems in the public sector, the future *Agência Nacional para o Digital* must impose the following mandatory TEVV protocols:

Requirement 2.2.1 - Monitoring: Public entities must implement continuous monitoring to measure the system's live outputs against the baseline established during procurement.

Requirement 2.2.2 - Proactive Drift Detection: Engineering and IT teams must specifically monitor for *Data Drift* (changes in the input data distribution over time, such as new economic realities affecting a tax fraud model) and *Concept Drift* (changes in the statistical relationship between input data and the target variable).

Requirement 2.2.3 - Automated Re-calibration Triggers: If a system's performance metrics drop below the legally acceptable baseline, an automatic technical alert must be generated. This must trigger mandatory human intervention for the re-calibration or retraining of the model.

3.3. Guideline 2.3: Red Lines and the Decommissioning Protocol

The final step of the NIST AI RMF is the "Manage" function, which allocates resources to mitigate prioritized risks. However, the most crucial, and often overlooked, administrative capability in the public sector is the absolute authority to stop an automated system.

Requirement 2.3.1 - The "Do Not Deploy" Rule: If, during the pre-deployment TEVV phases, the multidisciplinary engineering team determines that the algorithmic risks (e.g., discriminatory bias against a vulnerable group) cannot be sufficiently mitigated through technical controls, the public manager must exercise the decision to not deploy the system.

Requirement 2.3.2 - Mandatory Decommissioning and Fall-back Plans: For systems already in production (particularly Level III and IV systems), if continuous monitoring detects irreversible model drift that consistently violates the principles of fairness, or compromises the entity's ability to fulfill the duty to state reasons (Article 152 of the CPA), the system must be immediately suspended. Public entities must ensure service continuity by implementing tested fail-safe fallback plans. For Level III and IV systems, procurement contracts must mandate a secondary, human-led or conventional system to which the entity can immediately revert in case of algorithmic failure or decommissioning. Efficiency must never supersede administrative legality.

4. Pillar III: Team Ethics and Proactive Transparency

While Pillars I and II address the administrative and technical prerequisites for AI deployment, algorithmic governance ultimately relies on human agency and public trust. The last pillar of the proposed framework shifts the Portuguese Public Administration from a model of passive information access to one of active algorithmic transparency and offers the operational tools to enforce "ethics by design" within multidisciplinary development teams.

4.1. Guideline 3.1: Multidisciplinary Ethical Self-Assessment Scorecards

As noted in Chapter 3 and confirmed by practitioner evidence from the AIOLIA project [60], defining ethical principles is not enough without a way to quantify them. Teams working in engineering and public policy frequently lack a common vocabulary to assess how a system will affect society. To put ethics into action, the future *Agência Nacional para o Digital* must require the use of Ethical Self-Assessment Scorecards for all AI projects, which are based on the UK Data Ethics Framework.

This guideline proposes a hybrid methodological approach: it adopts the operational mechanics and lifecycle tracking of the UK Data Ethics Framework, but replaces its national criteria with the European Commission's Assessment List for Trustworthy Artificial Intelligence (ALTAI) [62] to ensure absolute alignment with the EU AI Act. Project teams must quantitatively score their progress on a scale of 0 to 5 for each of the seven European requirements for Trustworthy AI:

(1) **Human Agency and Oversight:** Ensuring AI empowers citizens and respects fundamental rights.
(2) **Technical Robustness and Safety:** Securing resilience to attacks, accuracy, and fail-safe plans.
(3) **Privacy and Data Governance:** Guaranteeing GDPR compliance and data quality integrity.
(4) **Transparency:** Providing traceability and explainability (XAI) to fulfill the CPA.
(5) **Diversity, Non-discrimination and Fairness:** Actively avoiding unfair bias and ensuring accessibility.
(6) **Societal and Environmental Well-being:** Evaluating broader ecological and democratic impacts.
(7) **Accountability:** Establishing auditability and risk management.

Methodological Reference: The UK Scoring Mechanics
While the seven ALTAI [62] dimensions dictate *what* is evaluated, the AIA-PT establishes that *how* they are evaluated follows the UK Government Digital Service [58]. Instead of a binary "compliant/non-compliant" checklist, teams evaluate the ALTAI auditable questions using a 0 to 5 maturity scale. This approach avoids methodological redundancy by leveraging a pre-validated project management framework while subjecting the system to European-standard ethical scrutiny.

Importantly, this scorecard should not be viewed as a static administrative barrier but rather as a dynamic tool. Teams are required to perform and record this self-evaluation using the Actions Scorecards at three different stages of the project lifecycle, aligning seamlessly with the continuous TEVV cycles established in Guideline 2.2:

- **At Start:** During initial project scoping and procurement.
- **During:** Iteratively, during development and continuous monitoring.
- **After:** Post-deployment review and auditing.

The Escalation Mechanism (The $\leq$ 3 Rule)
If the project scores three or below on any of the seven ALTAI principles, this indicates a critical compliance risk. The technical team is subjected to a mandatory pause. The project cannot proceed to the next lifecycle phase until the team formally documents the reasons for the low score and consults the team leader or the *Agência Nacional para o Digital* to advise on the specific next steps to improve the ethical standards of the project.

4.2. Guideline 3.2: The National Algorithmic Transparency Registry

At the moment, the *Lei de Acesso aos Documentos Administrativos* is primarily responsible for the transparency of Portuguese public data. Freedom of information requests, however, are a "Demand-Driven" (passive) approach that is insufficient for AI, according to the GPAI. Citizens cannot formally request information about automated systems if they are entirely unaware of their existence.

Portugal must adopt a "Supply-Driven" (active) strategy in order to address the risk of "performative transparency" and fully adhere to the constitutional duty to state reasons. The *Estratégia Digital Nacional* makes clear how crucial the *dados.gov.pt* portal and the *participa.gov.pt* platform are to improving civic engagement. Leveraging this existing infrastructure, the State must establish a National Algorithmic Transparency Registry.

The registry ensures that information is available to both technical auditors and the general public through a two-tier standardized format that was inspired by the UK's ATRS. Importantly, the Portuguese registry creates a "partial disclosure" rule based on the UK's Scope and Exemptions Policy. The public entity cannot withhold the complete record if certain algorithmic details pose a serious risk to cybersecurity, intellectual property, or algorithmic gaming (e.g., exposing precise thresholds for tax fraud detection). Rather, in accordance with the LADA, the entity must publish the registry and specifically designate the sensitive field as "Exempt," providing an effective legal explanation.

By centralizing this data, the Portuguese State will not only comply with possible European regulations but also create a culture of trust by allowing civil society to independently verify that public algorithms serve the public interest. The complete structure of this two-tier transparency registry is detailed in Table 5.4.

TABLE 5.4. Structure of the National Algorithmic Transparency Registry (Two-Tier Model)

Data Field	**Description / Requirement**
Tier 1: General Public	
System Name	Plain-Portuguese title of the AI application.
Description & Purpose	Clear, non-technical explanation of the intended civic benefit and scope.
Entity Responsible	The specific government department accountable for the system.
Impact Level	The official AIA-PT classification (Levels I, II, III, or IV).
Public Contact & Recourse	Official contact email and URL directing citizens on how to interact with or contest the system.
Tier 2: Auditors / Experts	

Table 5.4 – continued from previous page

Data Field	Description / Requirement
Owner and Responsibility	Identity and tax number (*NIPC*) of external suppliers, procurement procedure, and explicit data access terms granted to the vendor.
Description and Rationale	Detailed technical scope, expected operational benefits, and non-algorithmic alternatives previously considered.
Decision Making Process	Exact integration into the administrative workflow, frequency of use, and explicit "Human-in-the-loop" protocols (linking to CPA compliance).
Tool & Model Specification	System architecture, lifecycle phase (e.g., Pilot, Production), model type (e.g., LLM, Random Forest), and performance metrics (Precision/Recall).
Data Specification	Source, modality, and demographic composition of training datasets, including explicit mention of sensitive attributes processed.
Risks and Mitigations	Direct link to the Fundamental Rights Impact Assessment and documentation of technical/operational mitigations applied against algorithmic bias.
Applied Exemptions	Clear legal justification (under LADA) if any Tier 2 data is withheld to protect cybersecurity or trade secrets.

[This page is intentionally left blank.]

CHAPTER 6

Discussion

1. From Theory to Practice

The previous chapter introduced the principal artifact of this Design Science Research: a comprehensive set of Guidelines for Artificial Intelligence Systems for the Portuguese Public Administration that are based around three strategic pillars. However, the theoretical premise of an artifact is insufficient without rigorous methodological validation. This chapter is entirely dedicated to carrying out Phases 4 (Demonstration) and 5 (Evaluation) of the DSR methodology.

The central objective of this discussion is to demonstrate that the proposed architecture effectively closes the "Operational Gap" and overcomes the legal-technical deadlock identified at the beginning of this research. Show how the artifact balances the utilization of predictive algorithms with the strict duty to provide justification (*dever de fundamentação*) required by Articles 152 and 153 of the CPA, while guaranteeing anticipatory compliance with the EU AI Act.

This validation uses a Scenario-Based Evaluation, as described in Chapter 4. The artifact will be used in a diverse spectrum of theoretical scenarios within the Portuguese public sector, allowing for an *ex ante* qualitative evaluation of its operational mechanics and legal safeguards prior to empirical implementation.

2. Demonstration: Multi-Scenario Application

The guidelines are applied step-by-step to four distinct public sector use cases. The first three scenarios represent the most important areas that were mapped out during the systematic literature review, while the fourth introduces a low-risk municipal application. Each scenario has its own set of regulatory and technical stress-test conditions, demonstrating that the framework can enforce strict oversight when it needs to without stopping low-risk digital innovation.

2.1. Scenario A: Public Health (Predictive Triage in the SNS)

2.1.1. *Scenario Presentation*

An SNS hospital intends to procure an AI algorithm designed to analyze vital signs and cross-reference patient clinical history during emergency room triage. The system's objective is to predict the probability of hospital admission and the impact on access to essential healthcare.

2.1.2. *Applying Pillar I: Algorithmic Impact Assessment (AIA-PT)*

Step 1: Taxonomic Mapping (Guideline 1.1)

Following Guideline 1.1, the hospital's multidisciplinary team mapped the system according to the five OECD dimensions:

(1) **People and Planet:** The primary stakeholders are the patients awaiting urgent care. Passive stakeholders may be the medical and nursing staff, whose clinical judgment may be influenced by the system. One important social footprint identified is the possibility of discrimination, where the model might not give the same importance to patients from lower socioeconomic backgrounds due to historical health disparities in those areas.

(2) **Economic Context:** The system operates within the "Healthcare" sector of the Serviço Nacional de Saúde. Its administrative role involves triage support and resource allocation in emergency healthcare.

(3) **Data and Input:** The system uses two main streams: (i) Retrospective clinical data from the Registo de Saúde Eletrónico (RSE) for training; and (ii) Real-time field data, including vital signs and the MTS color code assigned by the nurse. The data provenance is internal to the SNS.

(4) **AI Model:** The core can be a Random Forest Classifier, which is a type of supervised ML model. To comply with the "Right to Explanation" under the Portuguese CPA, a SHAP module is integrated to give local explainability for each prediction.

(5) **Task and Output:** The task is "Binary Classification" (Admission vs. Discharge). The system has "Limited Autonomy" and works as a recommender system. The final decision remains with the human (Human-in-the-loop), who must validate the algorithmic recommendation before any hospital bed is officially allocated.

Step 2: Risk Scoring and Categorization (Guideline 1.2)

Based on the methodological workflow described in Chapter 5, the risk evaluation must begin with the *Legal Safeguard* before initiating any mathematical calculation.

The hospital's evaluation team first checks the proposed system against supranational legislation. This AI tool decides who can access essential public healthcare services and its failure could directly threaten patient life. This is why it aligns with the "High-Risk" classification criteria of the EU AI Act (Annex III). This immediately triggers the *EU AI Act Override Rule* under the AIA-PT framework. So, the system is automatically escalated to high-level governance protocols, preventing administrative redundancy.

Despite the legislative override regarding this evaluation but to validate the internal consistency of the AIA-PT Matrix against the European directive, the following section and using the mapping from Step 1, provides a simulated application of the numerical scoring based on Table 5.1:

(1) People & Planet score classification is Very High (4 points): The consequences of an algorithmic error (a false negative leading to an unjustified discharge) carry an irreversible impact on human life. Furthermore, the identified risk of discrimination actively threatens vulnerable demographic groups.
(2) Economic Context score classification is Very High (4 points): The system operates within the public health sector, specifically in the management of Emergency Department and patient prioritization, and therefore performs a high-risk, critical function.
(3) Data & Input score classification is High (3 points): The algorithm continuously processes highly protected and sensitive personal health data, including historical clinical records and real-time vital signs.
(4) AI Model score classification is High (3 points): The architecture utilizes complex ML models (e.g., Random Forests) augmented with post-hoc explainability modules (SHAP).
(5) Task & Output score classification is Moderate (2 points): The system does not make the final triage decision. It only drafts recommendations for human review, maintaining a strict "Human-in-the-loop" safeguard where the physician holds the ultimate authority.

By summing the dimension scores, the system achieves a Total Impact Score of 16 points. Looking at the Four-Tier Matrix, this places the system squarely within Level III (High Impact - Score 13 to 16 points).

Step 3: Human Oversight and Escalation (Guideline 1.3 & Guideline 1.4)

This requires a "Human-in-command" protocol in order to mitigate automation bias and follow Guideline 1.3. The algorithm can only suggest a priority, and the final admission decision must be always validated by a human doctor. Also, Guideline 1.4 says that if a patient's admission is denied or put on hold, the hospital must provide a personalized, meaningful explanation detailing the technical and clinical factors that influenced the AI's recommendation, as well as clear channels for the citizen to appeal the decision. Additionally, the deployment requires explicit political approval from the Minister of Health, who assumes the ultimate political and legal accountability for its operation within the SNS. There must also be a mandatory Fundamental Rights Impact Assessment (FRIA) to ensure that no demographic group is historically discriminated against by the triage data.

2.1.3. *Applying Pillar II: Engineering Requirements and Continuous Test, Evaluation, Verification, and Validation*

Public Procurement (Guideline 2.1)

When drafting the *Caderno de Encargos*, the hospital includes Requirement 2.1.1, which says that the vendor must show that the training data reflects the real demographic situation in Portugal. This reduces the risk of bringing in historical biases from other

healthcare systems. Under Requirement 2.1.2, the hospital requires XAI capabilities, and does not accept any vendor that claims "black-box" intellectual property clauses. This ensures that medical staff can always get the reasoning behind an admission prediction to meet their legal duty to give reasons. The hospital's procurement team also does not just arbitrarily define what the triage system's acceptable error margins are. The contract makes it clear that the vendor must meet the minimum standards for precision, recall, and false-negative rates set by the *Agência Nacional para o Digital*. This prevents the procurement of mathematically fragile systems and ensures a standardized, nationwide benchmark for clinical algorithmic robustness.

Drift Detection and Decommissioning (Guideline 2.2 & Guideline 2.3)
Six months after deployment, a severe winter flu outbreak changes the way patient symptoms are distributed statistically. The continuous monitoring protocols (Requirement 2.2.1) detect both a *Data Drift* and a critical *Concept Drift*, as the relationship between standard triage variables and the actual need for admission changes, causing the algorithm's precision to drop below the established baseline.

This deviation triggers an automated technical alert (Requirement 2.2.3), which requires immediate human intervention. Recognizing that the degraded performance compromises patient safety and the hospital's ability to guarantee legally valid clinical justifications, the administration triggers Guideline 2.3.

The AI system is immediately suspended under the *Mandatory Decommissioning* protocol (Requirement 2.3.2). The hospital can seamlessly revert to its traditional, purely human-led MTS, ensuring that public service continues without interruption. This fail-safe fallback plan was required during the procurement phase, which proved a key point of the AIA-PT: administrative efficiency should never come before constitutional rights and clinical safety.

2.1.4. *Applying Pillar III: Team Ethics and Proactive Transparency*

Ethical Scorecards (Guideline 3.1)
The technical project team must do the required UK-based ethical self-assessment during the initial project scoping and procurement. This is a requirement of the *Agência Nacional para o Digital*.

The team gives the system a score from 0 to 5 on each of the seven ALTAI dimensions during this first evaluation. The strict SNS infrastructure gives high scores to categories like *Technical Robustness and Safety* and *Privacy and Data Governance*. However, the *Transparency* dimension starts out with a score below 3. This low score means that the vendor's first proposal was based on an opaque architecture that did not have a clear XAI interface for the medical staff.

Because this score falls below the acceptable threshold, the AIA-PT *Escalation Mechanism (The $\leq$ 3 Rule)* is immediately triggered. The project has to stop for a mandatory

administrative pause. It cannot proceed to the "During" phase until the technical problem is fixed and documented. The hospital consults the *Agência Nacional para o Digital* to break the deadlock and legally forces the vendor to integrate a SHAP module to ensure local interpretability. After this architectural change, a second assessment raises the Transparency score to a number above 3.

This situation shows that the hybrid methodological approach is not a static bureaucratic checklist, but a dynamic "hard-stop" governance tool that actively prevents unethical AI deployments before they reach the clinical environment.

Algorithmic Transparency Registry (Guideline 3.2)

Once approved and put into use, the system is published in the National Registry. Patients can freely consult the general tool's details. Simultaneously, external auditors can access Tier 2 to review who developed it, the rationale behind it and all of its technical specifications.

2.2. Scenario B: Justice (Recidivism Risk Assessment by DGRSP)

2.2.1. *Scenario Presentation*

The *Direção Geral de Reinserção e Serviços Prisionais* intends to implement an AI risk assessment tool to evaluate the probability of criminal recidivism, helping judges in parole and early release decisions.

2.2.2. *Applying Pillar I: Algorithmic Impact Assessment (AIA-PT)*

Step 1: Taxonomic Mapping (Guideline 1.1)

Following Guideline 1.1, the DGRSP's multidisciplinary team mapped the proposed recidivism risk assessment system across the five OECD dimensions:

(1) **People and Planet:** The primary stakeholders are the inmates eligible for parole. Passive stakeholders may be judges, victims, and society as a whole. The social footprint is very high because of the risk of algorithmic bias. As demonstrated by international precedents (e.g., the COMPAS system in the United States), algorithms trained on historical criminal justice data can inadvertently punish racial minorities and people from lower socioeconomic backgrounds, perpetuating systemic cycles of incarceration.

(2) **Economic Context:** The system operates within the "Justice" sector of the Public Administration. Its administrative role is "Parole and Early Release Decision Support" and the outcome can directly dictate a citizen's physical liberty.

(3) **Data and Input:** The system uses two main streams: (i) Retrospective penal data from the historical DGRSP databases for training, including past criminal records and socio-demographic variables; and (ii) Real-time field data, including psychological evaluations and disciplinary assessments conducted by prison staff. The data provenance is strictly internal to the DGRSP and the Portuguese judicial system.

(4) **AI Model:** The core can be a Deep Neural Network, which is a type of complex predictive model. To comply with the judicial duty to state reasons under the Portuguese CPA, robust XAI mechanisms are integrated to guarantee that the *iter cognoscitivo* is transparent and comprehensible to the human magistrate.
(5) **Task and Output:** The task is "Risk Scoring" (e.g., classifying recidivism probability as Low, Medium, or High). The system has "Limited Autonomy", functioning strictly as an advisory tool. However, the risk of "automation bias" is extreme in this context, as judges might systematically defer to the algorithm's score to avoid being held personally responsible, which would undermine the human-in-the-loop safeguard.

Step 2: Risk Scoring and Categorization (Guideline 1.2)
The DGRSP evaluation team must use the *Legal Safeguard* first, according to the methodological workflow, before using the numerical rubric. When comparing the proposed recidivism assessment tool with European legislation, the system explicitly matches the criteria in Annex III for law enforcement and the administration of justice. The EU AI Act clearly calls the algorithm a "High-Risk" system because it assesses the risk of a person committing a crime or reoffending. This triggers the *EU AI Act Override Rule*, which skips over lower-level categories.

Despite the legislative override regarding this evaluation but to validate the internal consistency of the AIA-PT Matrix against the European directive, the following section and using the mapping from Step 1, provides a simulated application of the numerical scoring based on Table 5.1:

(1) People & Planet score classification is Very High (4 points): The denial of parole directly results in the deprivation of physical liberty. The impact on human dignity is irreversible, and the historical risk of algorithmic bias against marginalized demographic groups is severe.
(2) Economic Context score classification is Very High (4 points): The justice system and the penal execution apparatus represent "Sovereign high-stakes functions".
(3) Data & Input score classification is Very High (4 points): The system requires highly classified and deeply sensitive personal data, including criminal histories, disciplinary records within the prison system, and psychological profiles.
(4) AI Model score classification is High (3 points): The architecture involves complex predictive models (e.g., Deep Neural Networks).
(5) Task & Output score classification is High (3 points): While theoretically restricted to drafting recommendations (which would suggest a moderate risk), the extreme operational risk of "automation bias" fundamentally alters this classification, because judges may systematically defer to the algorithm's score to avoid personal liability.

By summing the dimension scores, the system achieves a Total Impact Score of 18 points. Looking at the Four-Tier Matrix, this places the system squarely within Level IV (Very High Impact - Score 17 to 20 points).

Step 3: Human Oversight and Escalation (Guideline 1.3 & Guideline 1.4)
This requires a "Human-in-command" protocol in order to mitigate automation bias and follow Guideline 1.3. The algorithm can only suggest a risk score, and the final parole decision must be always validated by a human judge. Also, Guideline 1.4 says that if an inmate's parole is denied, the court must provide a personalized, meaningful explanation detailing the technical and criminological factors that influenced the AI's recommendation (guaranteeing the *iter cognoscitivo*), as well as clear channels for the citizen to appeal the decision. Additionally, due to its Level IV classification, the deployment requires explicit political approval from the Council of Ministers (*Conselho de Ministros*) or the Portuguese Parliament (*Assembleia da República*), who assume the ultimate political and legal accountability for its operation within the judicial system. There must also be a mandatory Fundamental Rights Impact Assessment (FRIA) to ensure that no demographic or socioeconomic group is historically discriminated against by the penal data.

2.2.3. *Applying Pillar II: Engineering Requirements and Continuous TEVV*

Public Procurement (Guideline 2.1)
When drafting the *Caderno de Encargos*, the DGRSP includes Requirement 2.1.1, which says that the vendor must show that the training data reflects the real demographic situation in the Portuguese penal system and is free from historical biases against socioeconomic or ethnic minorities. This reduces the risk of bringing in historical biases from foreign criminal justice data. Under Requirement 2.1.2, the DGRSP requires XAI capabilities, and does not accept any vendor that claims "black-box" intellectual property clauses. This ensures that magistrates can always get the reasoning behind a risk score prediction to meet their legal duty to give reasons. The DGRSP's procurement team also does not just arbitrarily define what the risk assessment tool's acceptable error margins are. The contract makes it clear that the vendor must meet the minimum standards for precision, recall, and false-positive rates set by the *Agência Nacional para o Digital*.

Drift Detection and Decommissioning (Guideline 2.2 & Guideline 2.3)
Eight months after deployment, a major legislative change combined with shifting socioeconomic dynamics changes the way recidivism patterns are distributed statistically. The continuous monitoring protocols (Requirement 2.2.1) detect both a *Data Drift* and a critical *Concept Drift*, as the relationship between standard socio-demographic variables and actual reoffending changes, causing the algorithm's false-positive rate to spike above the established baseline.

This deviation triggers an automated technical alert (Requirement 2.2.3), which requires immediate human intervention. Recognizing that the degraded performance compromises the principle of equality and the court's ability to guarantee legally valid judicial justifications, the political authorities trigger Guideline 2.3.

The AI system is immediately suspended under the *Mandatory Decommissioning* protocol (Requirement 2.3.2). The DGRSP can go back to its traditional, purely human-led multidisciplinary psychological evaluations without any problems, which means that judicial due process will continue without interruption.

2.2.4. *Applying Pillar III: Team Ethics and Proactive Transparency*

Ethical Scorecards (Guideline 3.1)

The technical project team must do the required UK-based ethical self-assessment during the initial project scoping and procurement. This is a requirement of the *Agência Nacional para o Digital*.

The team gives the system a score from 0 to 5 on each of the seven ALTAI dimensions during this first evaluation. The strict DGRSP infrastructure gives high scores to categories like *Technical Robustness and Safety* and *Privacy and Data Governance*. However, the *Diversity, Non-discrimination and Fairness* dimension starts out with a score below 3. This low score means that the vendor's first proposal was trained on unmitigated historical penal data that disproportionately flagged individuals from lower socioeconomic backgrounds as high-risk.

Because this score falls below the acceptable threshold, the AIA-PT *Escalation Mechanism (The $\leq$ 3 Rule)* is immediately triggered. The project has to stop for a mandatory administrative pause. It cannot proceed to the "During" phase until the ethical and technical problem is fixed and documented. The DGRSP consults the *Agência Nacional para o Digital* to break the deadlock and legally forces the vendor to apply algorithmic bias mitigation techniques and prove its fairness metrics. After this data recalibration, a second assessment raises the Fairness score to a number above 3.

Algorithmic Transparency Registry (Guideline 3.2)

Once approved and put into use, the system is published in the National Registry. Legal representatives and the general public can freely consult Tier 1 to understand the tool's general purpose and scope. Simultaneously, external auditors and judicial oversight bodies can access Tier 2 to review who developed it, the rationale behind it, and all of its technical specifications, including the applied bias mitigation strategies.

Crucially, under the *Applied Exemptions* field, the team and DGRSP invokes the LADA to legally withhold the specific model parameters and mathematical weights from public disclosure. This exemption is legally justified to prevent algorithmic gaming by inmates who might otherwise artificially alter their behavior or responses in psychological evaluations to artificially lower their risk scores, thereby protecting the security and integrity of the penal execution system.

2.3. Scenario C: Public Finance (Tax Evasion Detection by Autoridade Tributária (AT))

2.3.1. *Scenario Presentation*

The AT plans to deploy an unsupervised ML model to detect complex, emerging tax evasion networks by analyzing corporate financial transfers in real-time.

2.3.2. *Applying Pillar I: Algorithmic Impact Assessment (AIA-PT)*

Step 1: Taxonomic Mapping (Guideline 1.1)

Following Guideline 1.1, the AT mapped the proposed tax evasion detection system across the five OECD dimensions:

(1) **People and Planet:** The primary stakeholders are corporate entities and individual taxpayers. Passive stakeholders include the national economy and the employees of the audited companies. The most important social footprint is the chance of algorithmic false positives. Unjustified asset freezes or aggressive tax audits based on flawed algorithmic profiling can bankrupt legitimate businesses, severely impacting economic livelihoods and violating the principle of presumption of innocence in tax matters.

(2) **Economic Context:** The system works in the "Public Finance" sector of the State. Its administrative role is to perform tax auditing and fraud network detection.

(3) **Data and Input:** The system uses two main streams: (i) Retrospective corporate financial data from domestic e-invoicing databases (SAF-T) and cross-referenced international asset registries for training; and (ii) Real-time field data, including continuous massive volumes of corporate bank transfers. The data provenance combines internal AT records with external banking information, all inherently sensitive and protected by strict tax and banking secrecy laws.

(4) **AI Model:** The core can be a Graph Neural Network paired with clustering algorithms (e.g., DBSCAN), to discover hidden, non-obvious "carousel fraud" networks, which is a type of unsupervised ML model. To comply with the administrative duty to state reasons under the Portuguese CPA, complex surrogate models and post-hoc XAI mechanisms are integrated to extract explainability from notoriously opaque clusters, guaranteeing that the *iter cognoscitivo* is transparent to the tax inspector.

(5) **Task and Output:** The task is "Anomaly Detection and Profiling" (e.g., classifying corporate networks as legitimate or potentially fraudulent). The system has "Limited Autonomy", functioning strictly as an advisory tool that automatically drafts audit priority lists. However, the risk of "automation bias" is severe in this context due to the sheer complexity of the multi-layered financial graphs generated by the AI, tax inspectors might systematically validate the algorithm's

findings, which would undermine the human-in-the-loop safeguard prior to any asset freezing.

Step 2: Risk Scoring and Categorization (Guideline 1.2)

Based on the methodological workflow described in Chapter 5, the risk evaluation must begin with the *Legal Safeguard* before initiating any mathematical calculation.

The AT team first checks the proposed system against supranational legislation. However, unlike the Health and Justice scenarios, a critical regulatory gap emerges here. While the EU AI Act's Annex III regulates systems used to evaluate creditworthiness, it explicitly exempts *"AI systems used for the purpose of detecting financial fraud"* (Annex III, point 5(b)). Because this AI tool targets corporate tax evasion networks, it falls squarely into this regulatory exception, effectively bypassing the "High-Risk" classification. Therefore, the *EU AI Act Override Rule* is not triggered.

Because of this legislative gap, applying the AIA-PT scoring becomes strictly mandatory to ensure the system does not operate without adequate oversight. Using the mapping from Step 1, the following section provides the numerical scoring based on Table 5.1:

(1) People & Planet score classification is High (3 points): While it does not directly threaten human life, the consequences of an algorithmic error (a false positive resulting in an unjustified asset freeze) carry a high, difficult-to-reverse impact on core economic rights, potentially bankrupting legitimate businesses.

(2) Economic Context score classification is Very High (4 points): The system operates within the public finance sector. Executing tax collection and prosecuting fraud networks represent a "Sovereign high-stakes function" of the State.

(3) Data & Input score classification is High (3 points): The algorithm continuously processes highly protected and sensitive financial data, including domestic e-invoicing (SAF-T) and banking transfers protected by strict tax secrecy laws.

(4) AI Model score classification is Very High (4 points): The architecture utilizes extremely complex, inherently opaque Unsupervised ML models (e.g., Graph Neural Networks). However, the AT actively integrated post-hoc XAI mechanisms to comply with the CPA.

(5) Task & Output score classification is High (3 points): While theoretically designed with "Limited Autonomy" as an advisory tool, the extreme operational risk of "automation bias" fundamentally alters this classification. Due to the sheer complexity of the multi-layered financial graphs, tax inspectors might systematically validate the algorithm's findings.

By summing the dimension scores, the system achieves a Total Impact Score of 17 points. Looking at the Four-Tier Matrix, this places the system at Level IV (Very High Impact - Score 17 to 20 points).

This result successfully illustrates the AIA-PT framework's necessity, proving that the mathematical matrix acts as a critical regulatory safety net to capture complex, sovereign systems that slip through the cracks of the EU AI Act.

Step 3: Human Oversight and Escalation (Guideline 1.3 & Guideline 1.4)

This requires a "Human-in-command" protocol in order to mitigate automation bias and follow Guideline 1.3. The algorithm can only suggest suspicious networks or draft audit priority lists, and any final coercive action (such as an asset freeze) must be always validated by a human tax inspector. Also, Guideline 1.4 says that if a corporate entity's assets are frozen or a formal audit is launched, the AT must provide a personalized, meaningful explanation detailing the technical and financial anomalies that influenced the AI's recommendation (guaranteeing the *iter cognoscitivo*), as well as clear channels for the business to appeal the decision. Additionally, due to its Level IV classification, the deployment requires explicit political approval from the Council of Ministers (*Conselho de Ministros*) or the Portuguese Parliament (*Assembleia da República*), who assume the ultimate political and legal accountability for its operation within the public finance system. There must also be a mandatory Fundamental Rights Impact Assessment (FRIA) to ensure that no legitimate business sector or specific regional economy is systematically discriminated against by the financial data.

2.3.3. *Applying Pillar II: Engineering Requirements and Continuous TEVV*

Public Procurement (Guideline 2.1)

When drafting the *Caderno de Encargos*, the AT includes Requirement 2.1.1, which says that the vendor must show that the training data reflects the real corporate financial landscape in the Portuguese economy. This reduces the risk of bringing in incompatible economic models from foreign tax jurisdictions. Under Requirement 2.1.2, the AT requires XAI capabilities, and does not accept any vendor that claims "black-box" intellectual property clauses or uses the complexity of unsupervised clustering as an excuse. This ensures that tax inspectors can always get the reasoning behind an audit recommendation (via surrogate models) to meet their legal duty to give reasons. The AT's procurement team also does not just arbitrarily define what the fraud detection tool's acceptable error margins are. The contract makes it clear that the vendor must meet the minimum standards for algorithmic precision and maximum false-positive rates set by the *Agência Nacional para o Digital*.

Drift Detection and Decommissioning (Guideline 2.2 & Guideline 2.3)

Twelve months after deployment, an unprecedented macroeconomic shock (e.g., a global supply chain crisis and subsequent state subsidy injections) changes the way corporate financial transfers are distributed statistically. The continuous monitoring protocols (Requirement 2.2.1) detect both a *Data Drift* and a critical *Concept Drift*, as the relationship between normal financial transactions and actual tax evasion changes. The algorithm misinterprets legitimate emergency subsidies as carousel fraud, causing the false-positive rate to spike above the established baseline.

This deviation triggers an automated technical alert (Requirement 2.2.3), which requires immediate human intervention. Recognizing that the degraded performance compromises economic stability and the AT's ability to guarantee legally valid audit justifications, the political authorities trigger Guideline 2.3.

The AI system is immediately suspended under the *Mandatory Decommissioning* protocol (Requirement 2.3.2). The AT can go back to its traditional, deterministic rule-based auditing frameworks without any problems, which means that tax administration will continue without interruption.

2.3.4. *Applying Pillar III: Team Ethics and Proactive Transparency*

Ethical Scorecards (Guideline 3.1)

The technical project team must do the required UK-based ethical self-assessment during the initial project scoping and procurement. This is a requirement of the *Agência Nacional para o Digital*.

The team gives the system a score from 0 to 5 on each of the seven ALTAI dimensions during this first evaluation. The strict AT infrastructure gives high scores to categories like *Technical Robustness and Safety* and *Privacy and Data Governance*, given the historical maturity of tax secrecy protocols. However, the *Human Agency and Oversight* dimension starts out with a score below 3. This low score means that the vendor's first proposal for the unsupervised clustering model lacked an intuitive interface for tax inspectors to challenge or override the complex network graphs, risking severe automation bias.

Because this score falls below the acceptable threshold, the AIA-PT *Escalation Mechanism (The $\leq$ 3 Rule)* is immediately triggered. The project has to stop for a mandatory administrative pause. It cannot proceed to the "During" phase until the ethical and technical problem is fixed and documented. The AT consults the *Agência Nacional para o Digital* to break the deadlock and legally forces the vendor to redesign the user interface, integrating a mandatory "human-override" workflow with clear audit logging before any asset freeze can be recommended. After this architectural recalibration, a second assessment raises the Human Agency score to a number above 3.

Algorithmic Transparency Registry (Guideline 3.2)

Once approved and put into use, the system is published in the National Registry. Corporate taxpayers and the general public can freely consult Tier 1 to understand the tool's general purpose and scope in combating tax evasion. Simultaneously, external auditors and sovereign oversight bodies (such as the *Tribunal de Contas*) can access Tier 2 to review who developed it, the rationale behind it, and all of its technical specifications, including the integrated surrogate explainability models.

Crucially, under the *Applied Exemptions* field, the AT invokes the LADA to legally withhold the specific clustering thresholds, network weights, and exact anomaly detection parameters from public disclosure. This exemption is legally justified to prevent algorithmic gaming by malicious corporate entities who might otherwise artificially structure

their financial transfers to stay just below the detection radar, thereby protecting the financial security and investigative integrity of the State.

2.4. Scenario D: Local Administration (Municipal FAQ Chatbot)

2.4.1. *Scenario Presentation*

A Local Municipality wants to put a NLP chatbot on its website. The goal of the system is to answer frequently asked questions from citizens, such as waste collection schedules, municipal pool hours, and basic administrative procedures. This will make it easier for people to get everyday local information 24/7.

2.4.2. *Applying Pillar I: Algorithmic Impact Assessment (AIA-PT)*

Step 1: Taxonomic Mapping (Guideline 1.1)

Following Guideline 1.1, the municipality's team mapped the proposed chatbot across the five OECD dimensions:

(1) **People and Planet:** The primary stakeholders are local citizens. The impact of an algorithmic failure (e.g., the bot providing the wrong pool schedule) is minimal and perfectly reversible. There are no vulnerable groups being specifically targeted or profiled.

(2) **Economic Context:** The system operates within Local Administration. Its role is strictly "General public service delivery" focused on civic convenience.

(3) **Data and Input:** The system uses two main streams: (i) Retrospective administrative data from official municipal regulations, public schedules, and static PDFs for the knowledge base; and (ii) Real-time field data, including continuous text queries inputted by citizens on the public portal. The data provenance combines internal municipal repositories with external public interactions, all exclusively open, non-personal, and strictly devoid of sensitive citizen profiling.

(4) **AI Model:** The core architecture is a simple Intent-Matching NLP model or an LLM strictly constrained by Retrieval-Augmented Generation (RAG) tied to official municipal documents. This architecture ensures high explainability and traceability of the generated text.

(5) **Task and Output:** The system does not have an impact in administrative decisions, it merely retrieves and displays public information to help people with their daily routines.

Step 2: Risk Scoring and Categorization (Guideline 1.2)

Based on the methodological workflow described in Chapter 5, the risk evaluation must begin with the *Legal Safeguard* before initiating any mathematical calculation.

The Municipality's team first checks the proposed system against supranational legislation. A chatbot designed solely to answer general FAQs does not determine access to essential services, nor does it involve biometrics, employment, or law enforcement. Therefore, it does not fall under the "High-Risk" categories of Annex III of the EU AI Act.

Consequently, the *EU AI Act Override Rule* is not triggered, and the evaluation proceeds to the scoring.

To determine the proportionate level of governance required, the team using the mapping from Step 1, calculates the Total Impact Score based on Table 5.1:

(1) People & Planet score classification is Low (1 point): The system has a minimal and perfectly reversible impact. Providing an incorrect municipal schedule does not threaten fundamental rights, and no vulnerable groups are targeted.
(2) Economic Context score classification is Moderate (2 points): The system operates within Local Administration, focusing strictly on general public service delivery and civic convenience.
(3) Data & Input score classification is Low (1 point): The chatbot relies exclusively on open, anonymized, and non-personal data (e.g., public municipal regulations and static PDFs).
(4) AI Model score classification is Moderate (2 points): The architecture utilizes a constrained NLP model with RAG, functioning as a simple ML system with high explainability and traceability.
(5) Task & Output score classification is Low (1 point): The system is restricted to information sorting and basic interactions. It has zero autonomy over administrative decisions.

By summing the dimension scores, the system achieves a Total Impact Score of 7 points. Looking at the Four-Tier Matrix, this places the system at Level I (Low Impact - Score 5 to 8 points).

This result is a vital validation of the AIA-PT artifact. It demonstrates that the framework is flexible, scalable and proportionate. By accurately classifying this system as Level I, the framework protects the Local Administration from bureaucratic asphyxiation, allowing the municipality to innovate and implement the chatbot quickly without requiring the heavy administrative burden necessary for higher-risk systems.

Step 3: Human Oversight and Escalation (Guideline 1.3 & Guideline 1.4)

Classified as a Level I system under the AIA-PT, the municipal chatbot benefits from the most streamlined administrative procedures, avoiding unnecessary bureaucratic bottlenecks. Under Guideline 1.3, the system operates under a "Human-on-the-loop" protocol. The chatbot is allowed to interact with citizens on its own without case-by-case human validation because it only retrieves public information and does not perform any binding administrative acts (*atos administrativos vinculados*).

Regarding transparency and the duty to state reasons (Guideline 1.4), the municipality is not required to provide personalized explanations for the chatbot's answers. It must simply display a clear, plain-language notice on the chat interface informing the citizen that they are interacting with an AI system. There is also no need for a full, detailed FRIA, the team only needs to conduct a simple impact mapping to ensure basic GDPR

compliance, such as configuring the system to automatically discard any unexpected personal data a citizen might accidentally type into the chat.

Regarding the accountability matrix, it is important to note that institutional approval does not require Ministerial escalation. The system's deployment can be swiftly and legally authorized by a Middle Manager, like the *Chefe de Divisão* or the *Diretor de Serviços*. This perfectly illustrates the core philosophy of the AIA-PT: making sure that high-risk systems have strong protections while also encouraging and speeding up low-risk digital innovation in the Public Administration.

2.4.3. *Applying Pillar II: Engineering Requirements and Continuous TEVV*

Public Procurement (Guideline 2.1)

When drafting the *Caderno de Encargos*, the municipality adapts the engineering requirements proportionally for a Level I system. Under Requirement 2.1.1, the vendor must simply demonstrate that the NLP model works well with European Portuguese and local administrative language. This means that complex socio-demographic bias testing is not needed because no profiling is done. Under Requirement 2.1.2, the RAG architecture must legally cite and link the exact source document (e.g., the specific municipal PDF) next to its conversational answer in order to be explainable. Also, under Requirement 2.1.3, the contract sets basic, manageable performance baselines focused on conversational accuracy and low hallucination rates, avoiding the heavy mathematical rigidness required for high-risk sovereign systems.

Drift Detection and Decommissioning (Guideline 2.2 & Guideline 2.3)

Two years after deployment, a major overhaul of local waste management regulations renders several of the municipality's older documents obsolete. The continuous monitoring protocols (Requirement 2.2.1) detect that the chatbot is providing outdated collection schedules to citizens based on the old data, representing a localized *Data Drift* and this deviation triggers an automated technical alert (Requirement 2.2.3). This is a Level I system, so the response is quick and only operational, the chatbot is immediately suspended under the *Mandatory Decommissioning* protocol (Requirement 2.3.2). The municipality can quickly go back to its simplest backup plan: temporarily hiding the chat widget from the website and redirect people to the regular search bar and traditional phone lines, means that civic service will continue without any breaks. This perfectly demonstrates how Pillar II ensures that low-risk local innovations are technically reliable without making the administrative process more difficult.

2.4.4. *Applying Pillar III: Team Ethics and Proactive Transparency*

Ethical Scorecards (Guideline 3.1)

The technical project team conducts the required UK-based ethical self-assessment during the initial project scoping phase. During this evaluation, the team scores the system across the seven ALTAI dimensions but because the chatbot relies exclusively on open municipal data, does not profile citizens, and employs a highly traceable RAG architecture, unlike the

high-risk sovereign systems analyzed in previous scenarios, no dimension scores 3 or lower and the *Escalation Mechanism (The $\leq$ 3 Rule)* is not triggered. There are no required administrative breaks, bureaucratic deadlocks, or forced technical redesigns between the "Start" and "During" phases of the project. This is a great example of how the ethical scorecard works as a fair way to govern. It helps local innovation happen safely and with little risk, rather than being a static barrier to administration.

Algorithmic Transparency Registry (Guideline 3.2)

Once approved and put into use, the system is published in the National Registry. Local citizens can freely consult Tier 1 to understand the chatbot's general purpose and its role in municipal service delivery. Also, there is no risk of algorithmic gaming by malicious because the system only retrieves basic FAQs and does not pose any threat to law enforcement, financial security, or sensitive personal data. So, the municipality does not need to invoke any *Applied Exemptions* under the LADA. Both Tier 1 and Tier 2 (including the specific NLP architecture and public data sources) are available to everyone. This full-disclosure approach maximizes civic transparency and builds public trust in Local Administration technologies without compromising state security.

2.5. Demonstration Summary

To consolidate the findings from the multi-scenario demonstration, Table 6.1 provides a comparative overview of how the AIA-PT framework scales across different public sector contexts. This synthesis highlights the artifact's core strengths: its ability to seamlessly integrate with European regulations (acting as an immediate override in Scenarios A and B), its capacity to act as a regulatory safety net bridging critical European legislative gaps (Scenario C), and its agility in ensuring proportionate governance for low-risk systems, actively preventing bureaucratic asphyxiation (Scenario D).

TABLE 6.1. Summary of the AIA-PT Multi-Scenario Demonstration

Scenario	Public Sector	OECD Task	EU AI Act Status	Impact Score	Final Level
A	Health (SNS)	Binary Classification	High-Risk (Override)	16	**Level III** (High)
B	Justice (DGRSP)	Risk Scoring	High-Risk (Override)	18	**Level IV** (Very High)
C	Finance (AT)	Anomaly Detection	Exempt (Financial Fraud)	17	**Level IV** (Very High)
D	Local Admin	Information Sorting	Low-Risk (Transparency)	7	**Level I** (Low)

3. Evaluation

The Design Science Research methodology dictates the evaluation phase measures how effectively the proposed artifact addresses the explicated problem. Since the guidelines constitute a governance framework that has not yet been empirically deployed, this section conducts an *ex ante* evaluation through logical argumentation, drawing directly from the empirical stress tests conducted in the multi-scenario demonstration. The evaluation focuses on three core dimensions of artifact utility: legal resolution, institutional accountability, and regulatory robustness.

3.1. Resolving the Administrative Law Deadlock (CPA Art. 152/153)

The main operational gap found at the start of this research was the inherent conflict between opaque ML models ("black boxes") and the strict administrative duty to state reasons (*dever de fundamentação*), required by Articles 152 and 153 of the CPA. The demonstration scenarios provide empirical evidence that the AIA-PT effectively addresses this legal-technical impasse.

The artifact actively prevents the State from procuring opaque systems for high-risk functions by enforcing Requirement 2.1.2 (Transparency and XAI Capabilities) during the procurement phase. As shown in Scenarios A (Public Health) and B (Justice), mandating XAI modules ensures that medical staff and human magistrates can systematically extract a legally valid *iter cognoscitivo* to justify a hospital discharge or a parole denial.

On the other hand, the artifact shows its flexibility by not stopping low-risk innovation. As shown in Scenario D (Local Administration), the framework legally allows a "Human-on-the-loop" architecture and this releases the Public Administration from the burden of case-by-case human validation and personalized explanations, proving that the guidelines can optimize administrative efficiency while remaining strictly within constitutional bounds.

3.2. Closing the Accountability Gap

The literature review revealed a persistent "accountability gap" in public sector AI, where the State frequently leaves sovereign decision-making to private vendors, leading to a diffusion of responsibility in the event of algorithmic failures. The framework eliminates this ambiguity through its Procedural Escalation Matrix (Guideline 1.4) and proactive transparency requirements.

The artifact makes the level of algorithmic risk legally linked to the hierarchical level of institutional accountability. While a Middle Manager can swiftly authorize a low-risk municipal chatbot (Scenario D), extreme-risk sovereign systems (Scenarios B and C) structurally demand explicit political authorization from the Council of Ministers or the Parliament. This ensures that the ultimate legal and political responsibility for high-stakes algorithmic deployment remains exclusively with the elected sovereign body, rather than being pushed to IT departments or third-party vendors. Also, the required Ethical Scorecards (Guideline 3.1) and continuous procurement documentation (Pillar II) meticulously trace accountability throughout the entire development lifecycle, ensuring it is clear who decided what, and when.

3.3. Regulatory Robustness Beyond the EU AI Act

Perhaps the most critical validation of the artifact is its capacity to complement and exceed the supranational baseline established by the EU AI Act. While the European regulation relies heavily on predefined lists of "High-Risk" sectors (Annex III), the Portuguese Public Administration requires a dynamic mechanism capable of evaluating the contextual severity of sovereign functions not covered by those lists. This necessity was

categorically illustrated in Scenario C (Public Finance). The EU AI Act explicitly exempts systems used for detecting financial fraud from its highest regulatory scrutiny. If the Portuguese State relied solely on European hard law, a complex, opaque clustering algorithm capable of bankrupting legitimate businesses through unjustified asset freezes would operate with minimal oversight.

The AIA-PT's scoring rubric captured this precise regulatory blind spot. By evaluating the sovereign context, the data sensitivity, and the severe risk of automation bias, it escalated the tax evasion system to a Total Impact Score of 17 points (Level IV - Very High Impact).

The practical necessity of such a national safety net is further corroborated by documented failures in comparable European jurisdictions. In the Netherlands, the System Risk Indication (SyRI) was deployed by the Dutch government to detect welfare fraud by cross-referencing data from multiple public registries. In February 2020, the District Court of The Hague ruled the system unlawful, finding that it violated Article 8 of the European Convention on Human Rights due to its opacity and the absence of adequate safeguards against discriminatory profiling [63]. Applying the AIA-PT retrospectively to this case is useful, the system's cross-referencing of sensitive personal data from multiple sovereign registries, its unsupervised classification logic targeting economically vulnerable citizens, and its direct triggering of coercive state action would have produced a Total Impact Score placing it firmly within Level IV. This classification would have mandated a comprehensive FRIA, mandatory XAI capabilities in the procurement contract, and political authorization at the level of the *Conselho de Ministros*, procedural safeguards that, had they existed, would likely have prevented the system's deployment in its original form.

This provides preliminary evidence that the AIA-PT is not merely a bureaucratic translation of European law, but a robust, standalone governance architecture that actively protects the Portuguese economy and its citizens from localized algorithmic risks that slip through the cracks of the EU AI Act.

3.4. Limitations of the Evaluation

While the multi-scenario demonstration provides structured evidence that the AIA-PT resolves the legal-technical deadlock identified at the outset of this research, the evaluation methodology carries inherent limitations that must be acknowledged.

First, the scenarios were constructed theoretically by the author and validated against the framework's own internal logic, rather than through independent empirical testing. This *ex ante* approach, while recognized in DSR as a valid evaluation stage, means that the framework has not yet encountered the institutional friction, political resistance, or resource constraints of real deployment. Vendor disputes over XAI requirements or budgetary limitations affecting Pillar II compliance are foreseeable challenges that the scenarios do not model.

Second, the reproducibility of the AIA-PT scoring rubric has not been empirically tested. Two independent evaluators applying 5.1 to the same system may arrive at different scores in some dimensions.

Third, the framework's operationalization depends critically on the full establishment and technical capacity of the *Agência Nacional para o Digital*. The performance baselines required by Requirement 2.1.3, the TEVV protocols of Guideline 2.2, and the escalation mechanisms of the Guideline 3.1 all presuppose a central body with the authority and expertise to enforce them. If this institutional capacity is delayed or partial, these guidelines risk remaining aspirational rather than operational.

These limitations do not undermine the framework's conceptual validity or its contribution to resolving the identified operational gap. They do, however, define the empirical research agenda that must follow this work before the AIA-PT can be adopted as binding national policy, and are addressed directly in the directions for future work outlined in Chapter 7.

[This page is intentionally left blank.]

CHAPTER 7

Conclusion

This dissertation was motivated by a concrete and constitutionally consequential problem. The integration of AI into the Portuguese Public Administration creates a legal-technical deadlock between the operational logic of modern ML systems and the foundational guarantees of administrative law. Opaque and probabilistic algorithms increasingly deployed to allocate healthcare resources, assess criminal risk, and detect fiscal fraud are structurally incapable of producing the *iter cognoscitivo* required by Articles 152 and 153 of the CPA. Without a mechanism that translates regulatory obligations into operational engineering and management requirements, the Portuguese State risks deploying systems that are either unlawful by omission (generating administrative acts liable to annulment for lack of justification) or paralyzed by uncertainty (refusing to innovate for fear of violating the EU AI Act).

1. Summary of Contributions

This dissertation makes three distinct and interrelated contributions to the fields of IT, public administration governance and AI regulation.

1.1. The Operational Gap Diagnosis

Addressing RQ1, the SLR conducted in Chapter 2 mapped the primary applications, benefits, and risks of AI across Public Health, Justice, and Finance. The review confirms the predictive superiority of modern ML architectures in all three domains, however, it simultaneously identifies some structural and cross-domain risks that directly condition the lawfulness of AI in the Portuguese PA.

Addressing RQ2, the comparative framework analysis conducted in Chapter 3, produced the first structured diagnosis of Portugal's institutional maturity in AI governance relative to the international state of the art. By systematically contrasting the EU AI Act, the OECD classification framework, the Canadian Directive on Automated Decision-Making, the UK's ATRS and Data Ethics Framework, and the NIST AI Risk Management Framework against the *Estratégia Digital Nacional* and its 2025-2026 Action Plan, this dissertation identified four concrete governance gaps: the absence of an official AI system taxonomy, the lack of a mandatory algorithmic impact assessment instrument, the non-existence of a centralized, public algorithmic transparency registry and the absence of technical metrics to validate "trustworthiness" during public procurement. This diagnostic, synthesized in Table 3.1, provides the empirical foundation for the guidelines proposed in Chapter 5 and constitutes a standalone contribution to the literature on comparative AI governance in Civil Law jurisdictions.

1.2. The AIA-PT Framework

Addressing RQ3, the central artifact of this DSR is the AIA-PT framework, a tripartite governance middleware designed to operationalise abstract regulatory requirements into measurable engineering and management obligations for Portuguese public entities.

Pillar I, introduces the AIA-PT scoring instrument, which combines the multidimensional OECD classification framework with the impact-level logic of the Canadian Directive to produce a standardized, four-tier risk classification (Levels I to IV). The instrument is anchored by a mandatory EU AI Act pre-screening filter, ensuring that supranational "Hard Law" obligations are never superseded by national scoring logic. The resulting Procedural Escalation Matrix directly resolves the accountability gap by linking the level of algorithmic risk to a proportionate level of institutional approval authority, from a Middle Manager for Level I systems to the *Conselho de Ministros* or the *Assembleia da República* for Level IV deployments.

Pillar II, transforms the static Portuguese public procurement paradigm into a continuous AI lifecycle model, adapting the NIST AI Risk Management Framework (RMF) core functions (Map, Measure, Manage) into actionable contractual requirements. By mandating XAI capabilities, demographic data provenance verification, and continuous TEVV protocols, including automated drift detection triggers and mandatory decommissioning. These guidelines prevent the procurement of algorithmically opaque systems and ensure that the State retains the technical basis to fulfill its CPA duty to state reasons throughout the operational lifetime of each deployed system.

Pillar III, operationalises "ethics by design" through two complementary instruments, the Ethical Self-Assessment Scorecards that adopt the UK Data Ethics Framework mechanics applied to the seven European ALTAI dimensions, with a mandatory escalation trigger for any score at or below 3, and a National Algorithmic Transparency Registry structured in a two-tier disclosure model (general public and technical auditors) with a legally grounded partial-disclosure rule under the LADA. Together, these instruments shift Portugal from a reactive, "demand-driven" model of administrative transparency to a proactive, "supply-driven" one.

1.3. The Regulatory Safety Net for EU AI Act Blind Spots

One of the most original contribution of this dissertation lies in a finding that emerged from the scenario-based demonstration in Chapter 6. The EU AI Act, the most comprehensive AI regulatory instrument in use, explicitly exempts systems used for detecting financial fraud from its highest tier of oversight (Annex III, point 5(b)). Scenario C demonstrated that a Graph Neural Network deployed by the *Autoridade Tributária* to detect corporate tax evasion networks falls squarely within this exemption, and would therefore operate with minimal European regulatory oversight despite generating Total Impact Scores of 17 points under the AIA-PT rubric score, a Level IV classification mandating FRIA, XAI procurement requirements, and approval from the *Conselho de Ministros*.

This finding proves that the AIA-PT, is not a mere transposition instrument of European law into national practice. It is a standalone governance architecture with independent classification logic, capable of identifying and governing high-risk sovereign functions that the EU AI Act's predefined sector lists fail to capture. This result has direct implications beyond Portugal, any EU Member State deploying AI in tax enforcement, welfare fraud detection, or analogous sovereign functions faces the same regulatory blind spot, and a national governance instrument of the type proposed here is a structural necessity rather than a policy option.

2. Theoretical and Practical Implications

2.1. Theoretical Implications

This dissertation advances the academic literature on algorithmic governance in three directions. First, it demonstrates that the "Hybrid Administration" model, in which AI functions strictly as an auxiliary tool subordinate to human judgment and constitutional due process, is not merely a normative aspiration but can be operationalized through a structured, tiered scoring instrument grounded in existing international frameworks. The AIA-PT provides a concrete engineering specification for what "Hybrid Administration" looks like in procurement contracts, oversight protocols and transparency registries.

Second, this work contributes to the emerging literature on the transferability of AI governance frameworks across legal traditions. The predominance of Common Law contexts in algorithmic governance research creates a significant methodological gap for Civil Law jurisdictions, where the duty to substantiate administrative acts is codified rather than case-derived. The AIA-PT's explicit anchoring in Articles 152 and 153 of the CPA demonstrates how governance instruments can be adapted to Civil Law requirements without sacrificing the technical rigour of their engineering provisions.

Third, by documenting a specific and consequential blind spot in the EU AI Act through empirical scenario analysis, this dissertation contributes to the critical literature on the limitations of risk-based AI regulation. The finding that financial fraud detection systems, among the most consequential sovereign AI applications in terms of coercive state power and economic rights, fall outside the Act's highest oversight tier challenges the assumption that sector-specific exemption lists are a sufficient regulatory mechanism for dynamic high-stakes public sector deployments.

2.2. Practical Implications

For public managers and procurement officers, the AIA-PT provides an immediately actionable decision instrument: a standardized scoring rubric, a procedural escalation matrix with defined approval authorities, a mandatory procurement checklist and a two-tier transparency registry template. These tools are designed to be used without specialized AI expertise, translating complex regulatory obligations into concrete administrative procedures.

For the *Agência Nacional para o Digital*, the framework provides the "instruction manual" that the *Estratégia Digital Nacional* currently lacks. A set of technical and procedural standards that can be adopted as binding guidance for all public entities deploying AI systems, ensuring a uniform national standard rather than a fragmented, agency-by-agency approach.

For legislators and policymakers, the Scenario C finding warrants direct attention in the context of the EU AI Act's national implementation because the transposition of the Regulation into Portuguese law should consider whether additional national provisions are required to govern sovereign AI applications in tax enforcement, welfare administration, and analogous domains where the Act's Annex III exemptions create unacceptable oversight gaps.

3. Limitations

The limitations of this dissertation are documented in full in Section 3.4 of Chapter 6 and are acknowledged here in summary. The evaluation of the AIA-PT is *ex ante* and scenario-based, meaning the framework has not yet encountered the institutional friction, political resistance, or resource constraints of real public sector deployment. The reproducibility of the scoring rubric under independent evaluators has not been empirically tested, leaving inter-rater reliability as an open question. Finally, the framework's operationalization presupposes the full institutional capacity of the *Agência Nacional para o Digital*, a dependency that introduces implementation risk if that capacity is delayed or partial.

These limitations are not incidental, they are structural consequences of applying a DSR methodology to a governance artifact in a rapidly evolving regulatory domain where empirical deployment requires political authorization and institutional infrastructure that lies beyond the scope of a single research project. They define, in precise terms, the empirical validation agenda that must follow this work.

4. Directions for Future Work

Three directions for future research are identified, ordered by priority.

4.1. Empirical Validation with the Portuguese Public Administration

The most urgent extension of this work is a structured pilot study applying the AIA-PT framework to one or more real AI deployments within the Portuguese Public Administration, conducted in partnership with the *Agência Nacional para o Digital* or a sectoral ministry. Such a study would test the framework's feasibility under real institutional constraints, measure the time and resource burden of the scoring and procurement procedures, and generate empirical evidence on whether the escalation thresholds produce legally defensible outcomes in practice. This pilot would constitute the Phase 4 and Phase 5 empirical validation that this dissertation's *ex ante* approach was unable to conduct.

4.2. Inter-Rater Reliability of the AIA-PT Scoring Rubric

A dedicated psychometric study should test the inter-rater reliability of the AIA-PT scoring instrument by having independent multidisciplinary teams (combining legal, technical, and administrative profiles) apply the rubric to the same set of AI systems and measuring the degree of agreement across evaluators. This study would identify which dimensions produce the greatest scoring variance, informing targeted refinements to the rubric (e.g., *People & Planet* or *Task & Output* dimensions) which involve the most contextual judgment. The results would also provide the basis for supplementary guidance materials that reduce subjectivity without constraining the rubric's applicability across diverse public sector contexts.

4.3. Extension to Other EU Civil Law Jurisdictions

The AIA-PT was designed for the specific legal and institutional context of Portugal. However, its foundational logic, anchoring algorithmic governance in the administrative duty to substantiate decisions, and complementing the EU Hard Law with national scoring instruments for regulatory blind spots, is directly applicable to other Civil Law jurisdictions within the EU. A comparative adaptation study would identify the legislative provisions in each system that are functionally equivalent to Articles 152 and 153 of the Portuguese CPA, assess whether the scoring rubric's thresholds require jurisdiction-specific calibration, and produce a generalized "Civil Law AI Governance Template" that can inform AI governance reform across the continental European administrative tradition.

5. Closing Remarks

The deployment of AI in the public sector is not a question of whether but of how. The Portuguese State, like all EU Member States, is already integrating algorithmic systems into the functions that define the relationship between government and citizen: who receives healthcare, who is granted liberty, who is audited for fraud. The question this dissertation has addressed is not whether these systems should exist, but under what conditions their use can be simultaneously effective, lawful, and constitutionally legitimate.

The answer proposed here is a governance middleware: a structured set of instruments that translate the abstract language of regulation into the concrete language of procurement contracts, oversight protocols, and transparency registries. The AIA-PT does not promise to eliminate algorithmic risk, what it does is ensure that risk is identified before deployment, proportionately governed throughout operation and honestly disclosed to the citizens it affects.

In a state governed by the rule of law, the Public Administration cannot simply decide effectively, it must be able to explain. The AIA-PT provides the tools to do both.

[This page is intentionally left blank.]

References

[1] A. Završnik, "Criminal justice, artificial intelligence systems, and human rights", English, *ERA Forum*, vol. 20, no. 4, pp. 567–583, 2020, Publisher: Springer, ISSN: 1612-3093. DOI: 10.1007/s12027-020-00602-0.

[2] European Commission, "Digital Decade 2025 country reports: Portugal", European Commission, Brussels, Commission Staff Working Document SWD(2025) 294 final, Part 22/27, Jun. 16, 2025.

[3] B. Magalhães, R. Costa, and F. Matos, "Artificial Intelligence and Administrative Discretion in Disciplinary Proceedings: Between Technological Efficiency and Legal Limits", English, *Revista Juridica Portucalense*, vol. 1, no. 38, pp. 1–24, 2025, Publisher: Universidade Portucalense, ISSN: 2183-5799. DOI: 10.34625/issn.2183-2705(38)2025.ic-1.

[4] K. Crawford and J. Schultz, "AI systems as state actors", English, *Columbia Law Review*, vol. 119, no. 7, pp. 1941–1972, 2019, Publisher: Columbia Law Review Association, ISSN: 0010-1958.

[5] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research", *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. DOI: 10.2753/MIS0742-1222240302.

[6] A. M. Marin, "Development of e-government in the field of social services and benefits: Evidence from Romania", English, *Information Polity*, vol. 27, no. 4, pp. 455–471, 2022, Publisher: IOS Press BV, ISSN: 1570-1255. DOI: 10.3233/IP-220005.

[7] T. Bench-Capon, "The need for good old fashioned AI and Law", English, *Jusletter IT*, no. December, pp. 23–35, 2021, Publisher: Editions Weblaw, ISSN: 1664-848X. DOI: 10.38023/cefe7081-e6dd-49de-9592-9adbb6063fd6.

[8] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, "Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach", English, *Big Data and Cognitive Computing*, vol. 8, no. 1, 2024, Publisher: Multidisciplinary Digital Publishing Institute (MDPI), ISSN: 2504-2289. DOI: 10.3390/bdcc8010006.

[9] R. M. O'donnell, "Challenging racist predictive policing algorithms under the equal protection clause", English, *New York University Law Review*, vol. 94, no. 3, pp. 544–580, 2019, Publisher: New York University Law Review, ISSN: 0028-7881.

[10] M. M. Farayola, I. Tal, R. Connolly, T. Saber, and M. Bendechache, "Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review", English, *Information (Switzerland)*, vol. 14, no. 8, 2023, Publisher: Multidisciplinary Digital Publishing Institute (MDPI), ISSN: 2078-2489. DOI: 10.3390/info14080426.

[11] N. Yi, D. Baik, and G. Baek, "The effects of applying artificial intelligence to triage in the emergency department: A systematic review of prospective studies", eng, *Journal of Nursing Scholarship: An Official Publication of Sigma Theta Tau International Honor Society of Nursing*, vol. 57, no. 1, pp. 105–118, Jan. 2025, ISSN: 1547-5069. DOI: 10.1111/jnu.13024.

[12] M. A. Halwani, G. Merdad, M. Almasre, G. Doman, S. AlSharif, S. M. Alshiakh, D. Y. Mahboob, M. A. Halwani, N. A. Faqerah, and M. T. Mosuily, "Predicting triage of pediatric patients in the emergency department using machine learning approach", English, *International Journal of Emergency Medicine*, vol. 18, no. 1, 2025, Publisher: BioMed Central Ltd, ISSN: 1865-1372. DOI: 10.1186/s12245-025-00861-z.

[13] S. K. Patel and D. Panday, "Optimizing Credit Card Fraud Detection: A Genetic Algorithm Approach with Multiple Feature Selection Methods", English, *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 13, 2024, Publisher: University of Salamanca, ISSN: 2255-2863. DOI: 10.14201/adcaij.31533.

[14] M. Marques, A. Almeida, and H. Pereira, "The Medicine Revolution Through Artificial Intelligence: Ethical Challenges of Machine Learning Algorithms in Decision-Making", eng, *Cureus*, vol. 16, no. 9, e69405, Sep. 2024, ISSN: 2168-8184. DOI: 10.7759/cureus.69405.

[15] M. Savić, J. Atanasijević, D. Jakovetić, and N. Krejić, "Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method", English, *Expert Systems with Applications*, vol. 193, 2022, Publisher: Elsevier Ltd, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2021.116409.

[16] J. Nover, M. Bai, P. Tismina, G. Raut, D. Patel, G. N. Nadkarni, B. S. Abella, E. Klang, and R. Freeman, "Comparing Machine Learning and Nurse Predictions for Hospital Admissions in a Multisite Emergency Care System", English, *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 3, 2025, Publisher: Elsevier B.V., ISSN: 2949-7612. DOI: 10.1016/j.mcpdig.2025.100249.

[17] G. Gebrael, K. K. Sahu, B. Chigarira, N. Tripathi, V. Mathew Thomas, N. Sayegh, B. L. Maughan, N. Agarwal, U. Swami, and H. Li, "Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0", English, *Cancers*, vol. 15, no. 14, 2023, Publisher: Multidisciplinary Digital Publishing Institute (MDPI), ISSN: 2072-6694. DOI: 10.3390/cancers15143717.

[18] Z. Xu, "Human Judges in the Era of Artificial Intelligence: Challenges and Opportunities", English, *Applied Artificial Intelligence*, vol. 36, no. 1, 2022, Publisher: Taylor and Francis Ltd., ISSN: 0883-9514. DOI: 10.1080/08839514.2021.2013652.

[19] G. Peluso Lopes, "Bias in Adjudication and the Promise of AI: Challenges to Procedural Fairness", English, *Law, Technology and Humans*, vol. 7, no. 1, pp. 47–67, 2025, Publisher: Queensland University of Technology, ISSN: 2652-4074. DOI: 10.5204/lthj.3812.

[20] A. Nord-Bronzyk, J. Savulescu, A. Ballantyne, A. Braunack-Mayer, P. Krishnaswamy, T. Lysaght, M. E. H. Ong, N. Liu, J. Menikoff, M. Mertens, and M. Dunn, "Assessing Risk in Implementing New Artificial Intelligence Triage Tools—How Much Risk is Reasonable in an Already Risky World?", English, *Asian Bioethics Review*, vol. 17, no. 1, pp. 187–205, 2025, Publisher: Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, ISSN: 1793-8759. DOI: 10.1007/s41649-024-00348-8.

[21] M. Simmler, S. Brunner, G. Canova, and K. Schedler, "Smart criminal justice: Exploring the use of algorithms in the Swiss criminal justice system", English, *Artificial Intelligence and Law*, vol. 31, no. 2, pp. 213–237, 2023, Publisher: Institute for Ionics, ISSN: 0924-8463. DOI: 10.1007/s10506-022-09310-1.

[22] H. Benouachane, "ARTIFICIAL INTELLIGENCE IN SOCIAL SECURITY: OPPORTUNITIES AND CHALLENGES", English, *Zhurnal Issledovanii Sotsial'noi Politiki*, vol. 20, no. 3, pp. 407–418, 2022, Publisher: National Research University, Higher School of Econoimics, ISSN: 1727-0634. DOI: 10.17323/727-0634-2022-20-3-407-418.

[23] A. Z. Huq, "Racial equity in algorithmic criminal justice", English, *Duke Law Journal*, vol. 68, no. 6, pp. 1004–1134, 2019, Publisher: Duke University School of Law, ISSN: 0012-7086.

[24] Z. He and W. Li, "AI-Driven Management of Type 2 Diabetes in China: Opportunities and Challenges", English, *Diabetes, Metabolic Syndrome and Obesity*, vol. 18, pp. 85–92, 2025, Publisher: Dove Medical Press Ltd, ISSN: 1178-7007. DOI: 10.2147/DMSO.S495364.

[25] A. Sachoulidou, "Going beyond the "common suspects": To be presumed innocent in the era of algorithms, big data and artificial intelligence", English, *Artificial Intelligence and Law*, 2023, Publisher: Institute for Ionics, ISSN: 0924-8463. DOI: 10.1007/s10506-023-09347-w.

[26] Q. A. Almulihi, A. A. Alquraini, F. A. A. Almulihi, A. A. Alzahid, S. S. A. J. Al Qahtani, M. Almulhim, S. H. S. Alqhtani, F. M. N. Alnafea, S. A. S. Mushni, N. A. Alaqil, M. I. F. Assiri, and N. H. Maghraby, "Applications of Artificial Intelligence and Machine Learning in Emergency Medicine Triage - A Systematic Review", English, *Medical archives (Sarajevo, Bosnia and Herzegovina)*, vol. 78, no. 3, pp. 198–206, 2024, ISSN: 1986-5961. DOI: 10.5455/medarh.2024.78.198-206.

[27] K. Piliuk and S. Tomforde, "Artificial intelligence in emergency medicine. A systematic literature review", eng, *International Journal of Medical Informatics*, vol. 180, p. 105 274, Dec. 2023, ISSN: 1872-8243. DOI: 10.1016/j.ijmedinf.2023.105274.

[28] O. Adebayo, Z. A. Bhuiyan, and Z. Ahmed, "Exploring the effectiveness of artificial intelligence, machine learning and deep learning in trauma triage: A systematic review and meta-analysis", English, *Digital Health*, vol. 9, 2023, Publisher: SAGE Publications Inc., ISSN: 2055-2076. DOI: 10.1177/20552076231205736.

[29] V. Bellini, M. Russo, T. Domenichetti, M. Panizzi, S. Allai, and E. G. Bignami, "Artificial Intelligence in Operating Room Management", eng, *Journal of Medical Systems*, vol. 48, no. 1, p. 19, Feb. 2024, ISSN: 1573-689X. DOI: 10.1007/s10916-024-02038-2.

[30] S. Tyler, M. Olis, N. Aust, L. Patel, L. Simon, C. Triantafyllidis, V. Patel, D. W. Lee, B. Ginsberg, H. Ahmad, and R. J. Jacobs, "Use of Artificial Intelligence in Triage in Hospital Emergency Departments: A Scoping Review", eng, *Cureus*, vol. 16, no. 5, e59906, May 2024, ISSN: 2168-8184. DOI: 10.7759/cureus.59906.

[31] S. Molldrem, A. K. J. Smith, and A. McClelland, "Predictive analytics in HIV surveillance require new approaches to data ethics, rights, and regulation in public health", English, *Critical Public Health*, vol. 33, no. 3, pp. 275–281, 2023, Publisher: Routledge, ISSN: 0958-1596. DOI: 10.1080/09581596.2022.2113035.

[32] J. Yin, K. Y. Ngiam, and H. H. Teo, "Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review", eng, *Journal of Medical Internet Research*, vol. 23, no. 4, e25759, Apr. 2021, ISSN: 1438-8871. DOI: 10.2196/25759.

[33] Z. Li, H. Zhou, Z. Xu, and Q. Ma, "Machine learning and public health policy evaluation: Research dynamics and prospects for challenges", English, *Frontiers in Public Health*, vol. 13, 2025, Publisher: Frontiers Media SA, ISSN: 2296-2565. DOI: 10.3389/fpubh.2025.1502599.

[34] S. Tiribelli, A. Monnot, S. F. Shah, A. Arora, P. J. Toong, and S. Kong, "Ethics Principles for Artificial Intelligence–Based Telemedicine for Public Health", English, *American Journal of Public Health*, vol. 113, no. 5, pp. 577–584, 2023, Publisher: American Public Health Association Inc., ISSN: 0090-0036. DOI: 10.2105/AJPH.2023.307225.

[35] P. R. B. Fortes, "Paths to Digital Justice: Judicial Robots, Algorithmic Decision-Making, and Due Process", English, *Asian Journal of Law and Society*, vol. 7, no. 3, pp. 453–469, 2020, Publisher: Cambridge University Press, ISSN: 2052-9015. DOI: 10.1017/als.2020.12.

[36] M. Cavus, M. N. Benli, U. Altuntas, M. Sari, H. Ayan, and Y. F. Ugurluoglu, "Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal justice", English, *Applied Soft Computing*, vol. 176, 2025, Publisher: Elsevier Ltd, ISSN: 1568-4946. DOI: 10.1016/j.asoc.2025.113160.

[37] J. H. Carvalho, "Reincidência criminal em portugal: Um olhar através dos dados da direção geral de reinserção e serviços prisionais", M.S. thesis, ISCTE - Instituto Universitário de Lisboa, 2024.

[38] INESC-ID. "Iris project results presented at the supreme court of justice", Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa. [Online]. Available: https://www.inesc-id.pt/events/iris-project-results-presented-at-the-supreme-court-of-justice/.

[39] V. Chiao, "Fairness, accountability and transparency: Notes on algorithmic decision-making in criminal justice", English, *International Journal of Law in Context*, vol. 15, no. 2, pp. 126–139, 2019, Publisher: Cambridge University Press, ISSN: 1744-5523. DOI: 10.1017/S1744552319000077.

[40] P. Hartel, R. + Wegberg, and M. van Staalduinen, "Investigating sentence severity with judicial open data: A case study on sentencing high-tech crime in the Dutch criminal justice system", English, *European Journal on Criminal Policy and Research*, vol. 29, no. 4, pp. 579–599, 2023, Publisher: Springer Science and Business Media B.V., ISSN: 0928-1371. DOI: 10.1007/s10610-021-09503-5.

[41] Nuryani, A. B. Mutiara, I. M. Wiryana, D. Purnamasari, and S. N. W. Putra, "Artificial Intelligence Model for Detecting Tax Evasion Involving Complex Network Schemes", English, *APTISI Transactions on Technopreneurship*, vol. 6, no. 3, pp. 339–356, 2024, Publisher: Pandawan Sejahtera, ISSN: 2655-8807. DOI: 10.34306/att.v6i3.436.

[42] O. C. Xavier, S. R. Pires, T. C. Marques, and A. d. S. Soares, "Tax evasion identification using open data and artificial intelligence", English, *Revista de Administracao Publica*, vol. 56, no. 3, pp. 426–440, 2022, Publisher: Fundacao Getulio Vargas, ISSN: 0034-7612. DOI: 10.1590/0034-761220210256.

[43] D. Santschi, M. Grau, D. Fehrenbacher, and I. Blohm, "Artificial Intelligence to Improve Public Budgeting", Dec. 2024.

[44] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection", English, *Journal of Big Data*, vol. 9, no. 1, 2022, Publisher: Springer Science and Business Media Deutschland GmbH, ISSN: 2196-1115. DOI: 10.1186/s40537-022-00573-8.

[45] B. S. Kumar, P. P. Yadav, and M. R. Reddy, "An intelligent approach to detect and predict online fraud transaction using XGBoost algorithm", English, *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 3, pp. 1491–1498, 2024, Publisher: Institute of Advanced Engineering and Science, ISSN: 2502-4752. DOI: 10.11591/ijeecs.v35.i3.pp1491-1498.

[46] Z. Khan, D. Olivia, and S. Shetty, "A Machine Learning and Explainable Artificial Intelligence Approach for Insurance Fraud Classification", English, *Inteligencia*

Artificial, vol. 28, no. 75, pp. 140–169, 2025, Publisher: Asociacion Espanola de Inteligencia Artificial, ISSN: 1137-3601. DOI: 10.4114/intartif.vol28iss75pp140-169.

[47] B. Alshawi, "Utilizing GANs for Credit Card Fraud Detection: A Comparison of Supervised Learning Algorithms", English, *Engineering, Technology and Applied Science Research*, vol. 13, no. 6, pp. 12 264–12 270, 2023, Publisher: Dr D. Pylarinos, ISSN: 2241-4487. DOI: 10.48084/etasr.6434.

[48] E. Mill, W. Garn, N. Ryman-Tubb, and C. Turner, "Opportunities in Real Time Fraud Detection: An Explainable Artificial Intelligence (XAI) Research Agenda", English, *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 1172–1186, 2023, Publisher: Science and Information Organization, ISSN: 2158-107X. DOI: 10.14569/IJACSA.2023.01405121.

[49] J. Morrison, "Digital conditionality: Universal credit and its implications for social citizenship", English, *Journal of Social Welfare and Family Law*, vol. 47, no. 2-3, pp. 320–335, 2025, Publisher: Routledge, ISSN: 0964-9069. DOI: 10.1080/09649069.2025.2530877.

[50] S. Hofmann, Ø. Sæbø, and H. H. Rydén, "Implications of Digitalised Welfare Services from a Vulnerable Citizens' Perspective", English, *Nordic Welfare Research*, vol. 9, no. 2, pp. 127–141, 2024, Publisher: Scandinavian University Press, ISSN: 2464-4161. DOI: 10.18261/nwr.9.2.2.

[51] S. C. Morse, "GOVERNMENT-TO-ROBOT ENFORCEMENT", English, *University of Illinois Law Review*, vol. 2019, no. 5, pp. 1497–1526, 2019, Publisher: University of Illinois College of Law, ISSN: 0276-9948.

[52] Y. Wang and X. Wang, "Artificial intelligence in physical education: Comprehensive review and future teacher training strategies", English, *Frontiers in Public Health*, vol. 12, 2024, Publisher: Frontiers Media SA, ISSN: 2296-2565. DOI: 10.3389/fpubh.2024.1484848.

[53] European Parliament and Council of the European Union, *Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, Official Journal of the European Union, L Series, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj.

[54] OECD and UNESCO, "G7 toolkit for artificial intelligence in the public sector", OECD Publishing, Paris, Tech. Rep., 2024, Report prepared for the 2024 Italian G7 Presidency.

[55] OECD, "OECD Framework for the Classification of AI systems", OECD Digital Economy Papers, No. 323, OECD Publishing, Paris, Tech. Rep., 2022. DOI: 10.1787/cb6d9eca-en.

[56] Global Partnership on Artificial Intelligence (GPAI), "Algorithmic transparency in the public sector: A state-of-the-art report of algorithmic transparency instruments", GPAI, Paris, Tech. Rep., 2024.

[57] Government of Canada, *Directive on automated decision-making*, Treasury Board of Canada Secretariat, 2019. [Online]. Available: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592.

[58] UK Government Digital Service, *Data ethics framework*, GOV.UK, 2020. [Online]. Available: https://www.gov.uk/government/publications/data-ethics-framework.

[59] US National Institute of Standards and Technology, "Artificial intelligence risk management framework (ai rmf 1.0)", U.S. Department of Commerce, Tech. Rep. NIST AI 100-1, 2023. DOI: 10.6028/NIST.AI.100-1.

[60] P. S. Bayerl, E. Lawlor, M. Maris, D. Miorandi, G. Pekšys, M. Bjelica, K. Stojšin, M. Anastasova, O. Smith, A. Henestrosa, I. Yamshchikov, M. Bak, and B. Akhgar, "Operational ethics guidelines on use cases related to human behaviour and cognition", AIOLIA Project, Horizon Europe, Tech. Rep. Deliverable D3.1, 2026, Grant Agreement N° 101187937. https://aiolia.eu/wp-content/uploads/2026/02/AIOLIA-D3.1-certified.pdf.

[61] Governo de Portugal, *Estratégia digital nacional e plano de ação 2025-2026*, Resolução do Conselho de Ministros, 2024. [Online]. Available: https://portugaldigital.gov.pt/.

[62] High-Level Expert Group on Artificial Intelligence, "The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment", European Commission, Brussels, Tech. Rep., Jul. 2020. DOI: 10.2759/002360. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[63] District Court of The Hague, *Judgment of 5 february 2020, case c-09-550982-ha za 18-388 (SyRI case)*, ECLI:NL:RBDHA:2020:1878, Rechtbank Den Haag, Feb. 2020. [Online]. Available: https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:1878.