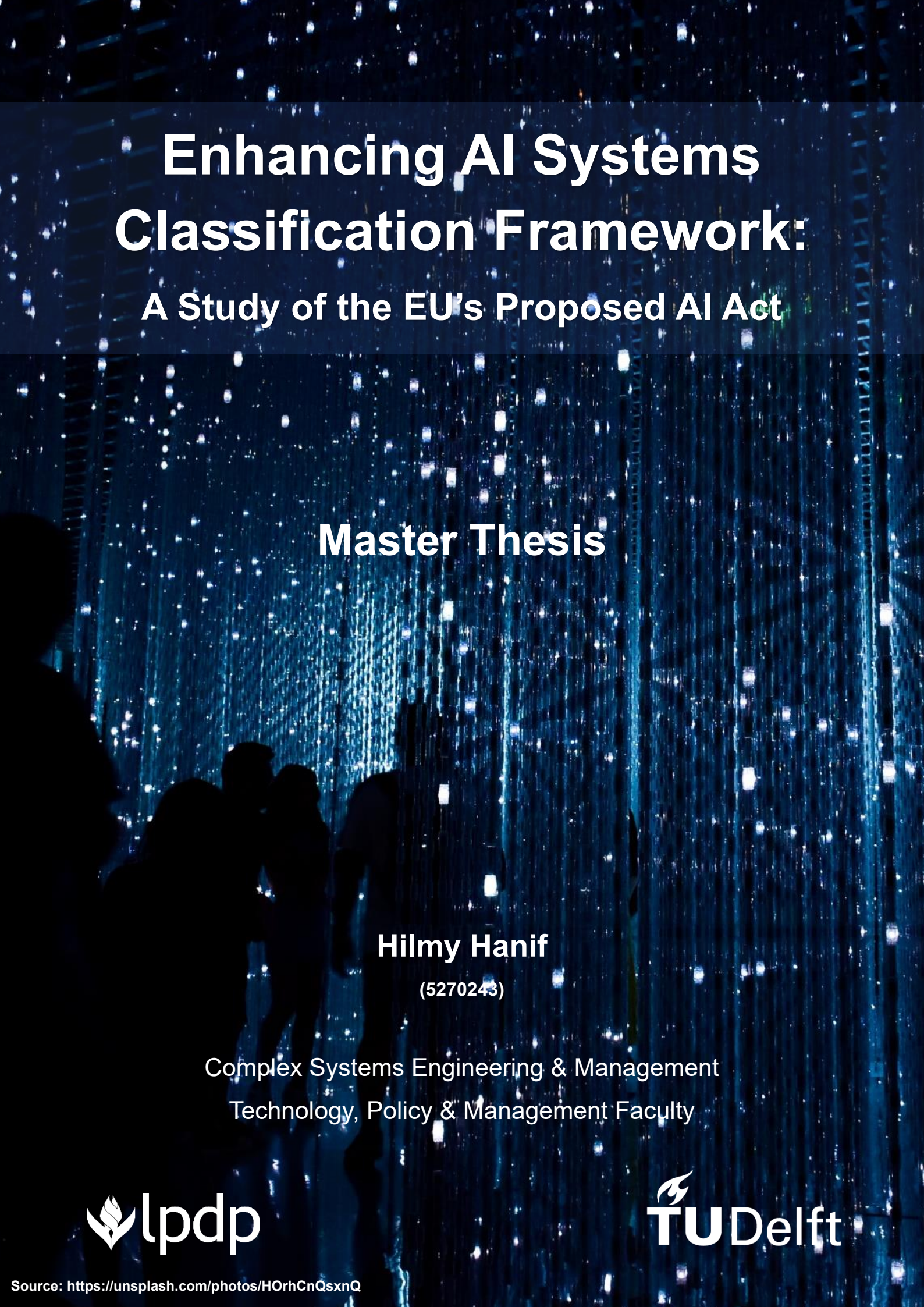

Enhancing AI Systems Classification Framework:

A Study of the EU's Proposed AI Act

Master Thesis

Hilmy Hanif

(5270243)

Complex Systems Engineering & Management

Technology, Policy & Management Faculty

lpdp

TU Delft

Source: https://unsplash.com/photos/HOrhCnQsxnQ

Enhancing AI Systems Classification Framework: A Study of the EU's Proposed AI Act

Master Thesis

by

Hilmy Hanif (5270243)

to obtain the degree of Master of Science

at the Complex System Engineering and Management, Delft University of Technology,

to be defended publicly on Wednesday, August 30, 2023 at 12:00 AM.

Student number:	5270243	
Project duration:	March 1, 2023 – August 30, 2023	
Thesis committee:	Prof. Michel van Eteen,	TU Delft, MAS/POLG, Chair
	Dr. Yury Zhauniarovich,	TU Delft, MAS/POLG, 1st Supervisor
	Dr. Jolien Ubacht,	TU Delft, ESS/ICT, 2nd Supervisor

This thesis is confidential and cannot be made public until August 30, 2024.

TU Delft

Preface

This thesis research marks the end of my master's journey, as I currently works on fascinating topic about AI regulation. Initially, I was unaware of how AI is seamlessly integrated into my everyday activities. From the transcription feature in platforms like Teams to helpful tools such as Grammarly, AI quietly assists us without fully realizing its widespread presence.

As discussions about regulating AI gained traction, I became increasingly intrigued and eager to delve deeper into this policy domain. Thus, the purpose of this research emerged—to understand how AI can be effectively classified.

I want to express my heartfelt gratitude to my supervisor, Dr. Yury Zhauniarovich, for believing in the significance of this topic and providing unwavering support throughout the research process. I am also thankful to Dr. Jolien Ubacht for her valuable insights into qualitative research methods and guidance in constructing a robust report. Additionally, I would like to thank Prof. Michel van Eteen for agreeing to chair the committee for this project and providing valuable feedback during the study.

I sincerely thank the many individuals who directly assisted me in this research, including Marie, Jorge, Aga, Antra, Puti, Uni Reni, and everyone who supported and motivated me to finish this project. Furthermore, I am grateful to those who contributed to enhancing my productivity, confidence, and overall capabilities in completing this thesis and my master's degree journey. I want to acknowledge the support I received from Agaby Masih, the Academic Counselor, and family, and friends, especially Lia and Hajar from the Thesis Writing Group. Our dedicated weekly writing sessions over two months proved immensely helpful in boosting my writing productivity.

I must admit that I commenced this project later than anticipated owing to personal issues that left me feeling unproductive and unsure of my direction. However, thanks to the passage of time, my genuine interest in this research topic and the support extended to me by countless individuals, especially from my families. I gradually found my path and regained my momentum.

I aspire that this study will not only contribute academically but also serve as a stepping stone for my personal growth, fostering the development of skills that will prepare me for the challenges beyond graduation.

In conclusion, I extend my deepest appreciation to all who have supported me throughout this endeavor, especially LPDP, with the financial support within these two years to experience a master's degree in a renowned university in the world, TU Delft. I hope this study will encourage further exploration as we continue to unravel the mysteries of AI.

Hilmy Hanif (5270243)
Delft, August 2023

Summary

The EU Artificial Intelligence Act (AI Act) proposed by the European Commission is a significant legislative effort to regulate AI systems. It is the first legal framework that specifically addresses the risks associated with AI systems, aiming to ensure their trustworthiness and alignment with the values enshrined in the Charter of the Fundamental Rights of the EU and the Union values. Thus, the draft of the AI Act emphasizes the importance of fundamental rights in Europe's AI approach.

The AI Act covers various AI applications, including machine learning, logical, statistical, and knowledge-based approaches. It provides a classification framework based on the purpose and risks posed by AI applications: Prohibited/Unacceptable risk, High-Risk, Limited-Risk, and Minimal/No risk. However, there are concerns about the clarity of the classification criteria mentioned in the AI Act. Some AI systems may fall into multiple classifications, leading to ambiguity. For example, a social robot used in patient treatment could be classified as High-Risk or Limited-Risk. This ambiguity is also observed in classifying AI systems in enterprise functions, where 40% of the classifications remain unclear.

Therefore, these challenges provide an opportunity to improve the classification process of AI systems under the AI Act, facilitating the classification process and accommodating emerging AI technologies. The main research question addressed in this thesis is: **"To what extent can the process of AI systems classification under the AI Act be improved?"**

The research focuses specifically on AI systems classification. It explores specific provisions of the AI Act, including Prohibited Risk, Classification Rules for High-Risk AI systems, Transparency Obligations, and Annexes II and III.

To achieve the objective of improving the classification accuracy of AI systems based on the AI Act, the study adopts the Design Science Methodology. This methodology involves systematically studying existing AI systems classifications and challenges, extracting themes to develop a framework, and evaluating the framework through feedback from AI experts.

A decision tree is designed as the proposed framework. It is evaluated on 16 respondents from two different backgrounds: legal and non-legal. In order to obtain comprehensive insights, the evaluation is designed to incorporate an experiment where respondents are tasked to classify AI systems to the risk level with the AI Act only. Then in the second experiment, they have to classify AI systems using the proposed decision tree framework. It is important to note that the study acknowledges the possibility of overestimating or underestimating respondents' ability to classify AI systems due to their diverse backgrounds and levels of understanding of the AI Act. Furthermore, a semi-structured interview is conducted to strengthen the analysis.

In general, the utilization of decision trees demonstrates a slight improvement in the classification of AI systems, particularly in terms of accuracy, reproducibility, and time-efficiency. Nevertheless, a comprehensive study analysis yields numerous valuable observations that can be utilized to enhance this classification.

Regarding the decision tree's performance for both obvious and non-obvious use cases, it became evident that the decision tree faced more significant challenges when classifying non-obvious cases compared to obvious ones, particularly in terms of reliability (reproducibility) and time efficiency. The lack of clarity within terms and definitions and limited contextual information posed notable difficulties in classifying non-obvious cases.

The performance of the decision tree also exhibited variations between legal and non-legal respondents. Legal experts demonstrated higher similarity agreement (reproducibility) and efficiency than non-legal respondents. These results indicate their familiarity with legal terminology and the nuances of the AI Act. However, it is noteworthy that both legal and non-legal respondents encountered challenges in classifying non-obvious cases, underscoring the need for enhancing decision tree frameworks or even reconsidering the creation of more effective tools/frameworks to enhance clarity and streamline the classification process.

Based on the analysis, several areas for improving AI systems classification under the AI Act have been identified. The current classification process faces challenges related to ambiguities in definitions, lack of contextual information, and difficulties in distinguishing between different risk levels.

To address these challenges and enhance the classification process, it is recommended to introduce clearer guidelines and refine the decision tree used for classification. The decision tree should incorporate additional criteria and features that provide more clarity and context. It is important to consider biases, subjective interpretations, clarity, and the dynamic nature of AI technologies in these improvements.

The study has certain limitations. The small sample size of respondents may impact the generalizability of the findings. The number of participants might not be representative of the entire population. Additionally, the limited number of use cases utilized in the research may limit the comprehensiveness of the classification framework. The study is based on the latest amendment of a policy proposal, and there is a potential for changes in the regulation's details, which may affect the effectiveness of the results. Finally, potential biases may exist in the development of the research, such as in making the decision tree and selecting the use cases.

Future research should explore the continuity of the decision tree's performance over time and its evaluation. There should be more research on non-obvious cases in specific domains or industries. It is crucial to focus on potential issues in classifying certain risk levels in the AI Act that hinder classification accuracy. Understanding the differences between legal and non-legal perspectives on the AI Act is also important to establish standardized understanding among stakeholders. Additionally, conducting quantitative research with larger and more diverse respondents from industrial backgrounds can further evaluate the proposed framework.

Contents

List of Tables

List of Figures

1 Introduction

The proposed EU Artificial Intelligence Act (AI Act) represents a significant legislative effort by the European Commission to govern AI systems [9]. The draft of this AI Act initially presented in April 2021 and revised in an amendment draft on May 2023, is the first legal framework on AI that specifically addresses the risk associated with AI systems. It aims to ensure the trustworthiness of AI systems while aligning their deployment with the values enshrined in the Charter of the Fundamental Rights of the EU and the Union values [5, 8]. Thus, the AI Act draft puts fundamental rights at the core of Europe's AI approach [5].

This proposal is the culmination of an EU-wide effort initiated by the political commitment of President von der Leyen, as stated in her political guidelines for the 2019-2024 Commission, which ultimately led to the establishment of the White Paper on AI - A European approach to excellence and trusts on 2020 [8]. Then, on May 2023, the latest amendment of this proposal was published to gather public feedback from the EU communities [11]. Similar to the implementation of the General Data Protection Regulation (GDPR), the AI Act's implementation is expected to require some time. Nevertheless, once enacted, the AI Act will significantly influence the development of AI systems in the EU for the next several decades.

The scope of the AI Act includes AI applications such as machine learning, logical, statistical, and knowledge-based approaches [9]. It provides criteria to categorize AI applications based on their purpose and the risks posed to fundamental rights: 1) Prohibited/Unacceptable risk (Title II); 2) High-Risk (Title III); 3) Limited-Risk (Title IV); and 4) Minimal/No risk (Title IX).

The Prohibited Risk prohibits AI practices that violate Fundamental Rights and Union values, such as AI systems using remote biometric identification or AI systems intended to harmful manipulation of a natural person's behaviour.

High-Risk AI systems are those intended to be used as a safety component of a product or are themselves a product. They require a third-party conformity assessment [10], and subjected to more stringent requirements than other categories. These include obligations around risk management, data quality, technical documentation, human oversight, transparency, robustness, accuracy, and security [9]. The European Commission may broaden the list of High-Risk AI systems used within specified pre-defined sectors by employing a set of criteria and risk assessment methods to ensure that the regulation can accommodate the emerging use cases and applications of AI.

For some AI systems, there are specific transparency obligations for the AI systems to comply with. For instance, the AI systems intended to interact with individuals, such as chatbots. For AI systems with such systems, the user has to be aware that they are interacting with the machine.

Meanwhile, the AI systems categorized as Minimal/No Risk, the AI Act does not impose any additional obligations but a voluntary code of conduct, with the same obligations as the High-Risk mandatory requirements.

However, scholars have noted that the classification criteria mentioned in AI Act are unclear. Moreover, certain AI systems may fall under one or more classification [18]. For instance, the use of social robot to assist in a patient's treatment could potentially fall under both High-Risk and Limited Risk. This use case may exceptionally entail two different risk levels, depending on the application area, because of exceptions stated in the legal text. Social robot with emotion recognition systems requires transparency obligations as they are associated with Limited Risk. Nevertheless, this use case also is obliged to High-Risk requirements since it is mentioned as part of medical devices.

The potential unclear classification is also assessed on the AI systems in enterprise functions where 40% of the AI systems classification in these areas remain ambiguous [25]. In this context, we assume that companies are more likely to choose the high-risk category in case of doubt in order to avoid potential risks. However, in discussions with European institutions, most of the unclear cases could fall into the lower risk category. The large proportion of unclear risk classifications creates a lot of uncertainty in all areas, which can further slow down investment in AI and the laready sluggish adaption of AI in Germany and Europe. A fear of mistakes or penalties in companies matters here too [25].

Furthermore, the current classification of AI systems may not effectively accommodate rapidly developing technologies like Generative AI, raising concerns about the necessity of a flexible framework to accommodate potential breakthrough technologies. Ongoing discussions regarding the classification according to AI Act have proposed that AI systems generating complex texts without human oversight should be included in the High-Risk AI list, to prevent AI systems from producing disinformation at scale [15]. This highlights the need for a flexible classification framework that can accommodate emerging AI systems.

Hence, these challenges provide a unique opportunity to enhance the AI systems classification under the AI Act, facilitating the classification process and accommodating emerging AI systems. The question to be answered in this thesis is the following:

To what extent can the process of AI systems classification under the AI Act be improved?

The focus of this research is specifically on the AI systems classification. Thus, the research will explore specific provisions of the AI Act, including Article 5 (Prohibited Risk), Article 6 (Classification Rules for High-Risk AI systems), Article 52 (Transparency Obligations) and Annexes II and III. It is essential to note that by limiting the number of participants and considering their diverse backgrounds and levels of understanding of the AI Act, there is a possibility of either overestimating or underestimating their ability to classify various AI systems during the validation phase.

This study aims to improve the the effectiveness and efficiency of the AI systems classification based on the AI Act. To achieve this objective, the Design Science Methodology is adopted. This scientific approach involves the systematic study and creation of artifacts to address practical problems of general interest [19]. The methodology encompasses identifying the requirements necessary to enhance AI systems classification, transforming these requirements into a decision tree framework that can further improve the classification process, and evaluating the framework by collecting feedback from AI experts.

By implementing the proposed decision tree framework, it is anticipated that the AI system classification will have better performance in terms of accuracy, reliability, and efficiency. This improvement will benefit organizations, AI providers, and AI experts by providing them with an initial screening tool for classifying their AI systems, as incorrect classification can result in fines [11].

Furthermore, the evaluation of the decision tree during this research will offer valuable insights to policymakers regarding the level of agreement among AI experts in classifying AI systems based on different use cases.

2

Problem Analysis

This chapter presents the problem analysis of the research by providing an overview of the AI Act, and some definitions of the AI Act and the structure of the regulation. It identifies the research gap that this study aims to address. The chapter also presents the research questions and provides the research flow diagram.

Section 2.1 briefly provides introduction of the AI Act and its classification. Sections 2.2 mentions some relevant definitions for this research. Section 2.3 and 2.4 discuss the research gap and research question, respectively. In Section 2.5, the research methods and the research sub-questions of this study are discussed. Finally, the research diagram and flow is visualized and explained in Section 2.6.

2.1. Overview of Artificial Intelligence Act

The Artificial Intelligence Act (AI Act) is a proposed regulation by the European Commission that was introduced in April 2021 to establish harmonised rules for Artificial Intelligence. It represents the first legal attempts to address the risks associated with AI systems [9]. This AI Act is still a proposal with the latest amendment was published on May, 2023 [11]. This regulation framework is proposed to ensure the trustworthiness of AI systems and align them with the Fundamental Rights and Union values [5, 8].

The AI Act has specific objectives outlined in the regulation [9]. Those objectives are:

- *ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values;*
- *ensure legal certainty to facilitate investment and innovation in AI;*
- *enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems;*
- *facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation [9].*

These objectives aim to foster the development, use, and uptake of AI in the EU while protecting safety, fundamental rights, and EU values. These objectives are distinct, but at the same time are compatible [5].

To achieve these objectives, the proposal defines harmonised rules for the development, placement on the market, and use of AI system in the EU. It follows a risk-based approach where AI systems that violate fundamental rights and union values are prohibited. High-Risk AI systems, which pose significant risks to the health, safety, fundamental rights and environment of natural person must comply with a set of horizontal mandatory requirements and follow conformity assessment procedures before being launched in the Union market. Additionally, AI systems with limited risk need minimum transparency obligations.

The regulation applies to AI providers placing or operating AI systems in the EU, regardless of their location, as well as users of AI systems located in the union and providers and users of AI systems located in third countries whose AI systems are used in the EU [9, 44].

2.2. Definitions and Structure in the AI Act

The AI Act proposal consists of 85 articles, with 75 articles in 11 sections devoted directly to AI regulation. The remaining 10 articles are amendments to several old legislation. In the latest amendment of the AI Act [11],

there are 67 definitions outlined in Title I, Article 3 (Definitions). According to the AI Act, an AI system is *a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions that influence physical or virtual environments.* Other definitions relevant to this research are provided in Table 2.1.

Table 2.1: Some Definitions in the AI Act (According to Title I, Article 3 of the AI Act)

Term	**Definitions in the AI Act**
AI System	a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions that influence physical or virtual environments
Risk	the combination of the probability of an occurrence of harm and the severity of that harm
Significant risk	a risk that is significant as a result of the combination of its severity, intensity, probability of occurrence, and duration of its effects, and its the ability to affect an individual, a plurality of persons or to affect a particular group of person
General purpose AI system	an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed
Intended purpose	the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation
Biometric data	biometric data as defined in Article 4, point (14) of Regulation (EU) 2016/679 (personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data
Biometric-based data	data resulting from specific technical processing relating to physical, physiological or behavioural signals of a natural person
Biometric identification	the automated recognition of physical, physiological, behavioural, and psychological human features for the purpose of establishing an individual's identity by comparing biometric data of that individual to stored biometric data of individuals in a database (one-to-many identification)
Biometric verification	the automated verification of the identity of natural persons by comparing biometric data of an individual to previously provided biometric data (one-to-one verification, including authentication)
Biometric categorisation	assigning natural persons to specific categories, or inferring their characteristics and attributes on the basis of their biometric or biometric-based data, or which can be inferred from such data
Emotion recognition system	an AI system for the purpose of identifying or inferring emotions,thoughts, states of mind or intentions of individuals or groups on the basis of their biometric and biometric-based data
Publicly accessible space	ny publicly or privately owned physical place accessible to the public, regardless of whether certain conditions for access may apply, and regardless of the potential capacity restrictions

In Table 2.2, the structure of the AI Act is presented. The regulation has eight specific provisions represented in 12 Titles. The first provision, Scope and Definitions, is listed in Title I, and some definitions are presented in Table 2.1. The second provision, Prohibited Artificial Intelligence Practices (Title II), establishes a list of

Table 2.2: Overview Structure of the AI Act

Specific Provision	Title	Chapter	Name	Articles
Scope and Definitions	I	N/A	Subject Matter, Scope, Definitions, and Amendment to Annex I	1-4
Prohibited Artificial Intelligence Practices	II	N/A	Prohibited Artificial Intelligence Practices	5
High-Risk AI Systems	III	1	Classifications of AI systems as High-Risk	6-7
		2	Requirements for High-Risk AI Systems	8-15
		3	Obligations of Providers and Users of High-Risk AI Systems and Other Parties	16-29
		4	Notifying Authorities and Notified Bodies	30-39
		5	Standards, Conformity Assessment, Certificates, Registration	40-51
Transparency Obligations for Certain AI Systems	IV	N/A	Transparency Obligations for Certain AI Systems	52
Measures in Support of Innovation	V	N/A	Measures in Support of Innovation (incl. AI Regulatory Sandbox)	53-55
Governance and Implementation	VI	1	European Artificial Intelligence Board	56-58
		2	National Competent Authorities	59
	VII	N/A	EU Database for Stand-Alone High-Risk AI Systems	60
	VIII	1	Post-Market Monitoring	61
		2	Sharing of Information of Incidents and Malfunctioning	62
		3	Enforcement	63-68
Codes of Conduct	IX	N/A	Codes of Conduct	69
Final Provisions	X	N/A	Confidentiality and Penalties	70-72
	XI	N/A	Delegation of Power and Committee Procedure	73-74
	XII	N/A	Final Provisions	75-85

establishes criteria that categorize AI systems deems as unacceptable as they violate fundamental rights or union values. The third provision, High-Risk AI Systems (Title III), contains specific rules for AI systems that pose risks to the health, safety, or fundamental rights of natural person. It defines classification rules for High-Risk AI systems and legal requirements they must comply with. Title III also mentions Annexes III, which lists use-cases of High-Risk AI systems. Transparency obligations for certain AI systems are presented in Title IV.

Title V, Measures in Support of Innovation, encourages the establishment of innovation-friendly regulatory sandboxes as a legal framework. Governance and Implementation provisions are discussed in Title VI, VII, and VIII, focusing on setting up governance systems at the Union and national levels, facilitating the monitoring work of the Commission and national authorities for High-Risk AI systems with implications for fundamental rights, and establishing monitoring and reporting obligations for providers of AI systems, respectively. Title IX, Codes of Conduct, creates a framework which encourage AI providers of non-high-risk AI to apply voluntarily application of the mandatory requirements outlined in Title III. Finally, the Final Provisions (Titles X, XI and XII) emphasize the obligations of all parties, outline rules for delegation and implementing powers, and establish the obligation to regularly assess the need for updating the list of use-cases for High-Risk AI Systems that mentioned in Annex III.

The AI Act also includes nine annexes that provide further provisions of the AI Act as listed in Table 2.3. Annex I discusses AI techniques and approaches, while Annex II sets list of union harmonisation legislation, derived from the New Legislative Framework and other Union harmonisation legislation. Annex III presents High-Risk AI systems that have significant harm to critical infrastructure outlined in Article 6 point 2. Annex IV and V discuss technical documentation and EU declaration of conformity, respectively. Annex VI sets the conformity assessment procedure based on internal control. Meanwhile, Annex VII discusses the conformity based on assessment of quality management system and assessment of technical documentation. Annex VIII specifies information to be submitted upon the registration of High-Risk AI Systems. Lastly, Annex IX presents Union Legislation on Large-Scale IT Systems in the Area of Freedom, Security and Justice.

Table 2.3: Overview Annexes of the AI Act

Annex	Title	Referred to Article
I	Artificial Intelligence Techniques and Approaches	Referred to in Article 3, point 1
II	List of Union Harmonisation Legislation	Referred to in Article 6
III	High-Risk AI Systems	Referred to in Article 6, point 2
IV	Technical Documentation	Referred to in Article 11, point 1
V	EU Declaration of Conformity	Referred to in Article 48
VI	Conformity Assessment Procedure Based on Internal Control	Referred to in Article 17, 61
VII	Conformity Based on Assessment of Quality Management System and Assessment of Technical Documentation	Referred to in Article 17
VIII	Information to be Submitted Upon the Registration of High-Risk AI Systems	Referred to in Article 51
IX	Union Legislation on Large-Scale IT Systems in the Area of Freedom, Security and Justice	-

2.3. State-of-the-Art Research

2.3.1. Selection Process

The literature related to the draft of the AI Act, published in 2021, is still in its early stages. To explore the existing body of work related to the AI Act, a literature search was conducted using the search keyword (“AI Act” or “Artificial Intelligence Act”) AND ("classification" or "categorization" or "classif*" or "categor*" or "assessment"). The search was done in April 2023 in the Scopus and Web of Science databases as visualized in PRISMA Flow Diagram depicted in Figure 2.1.

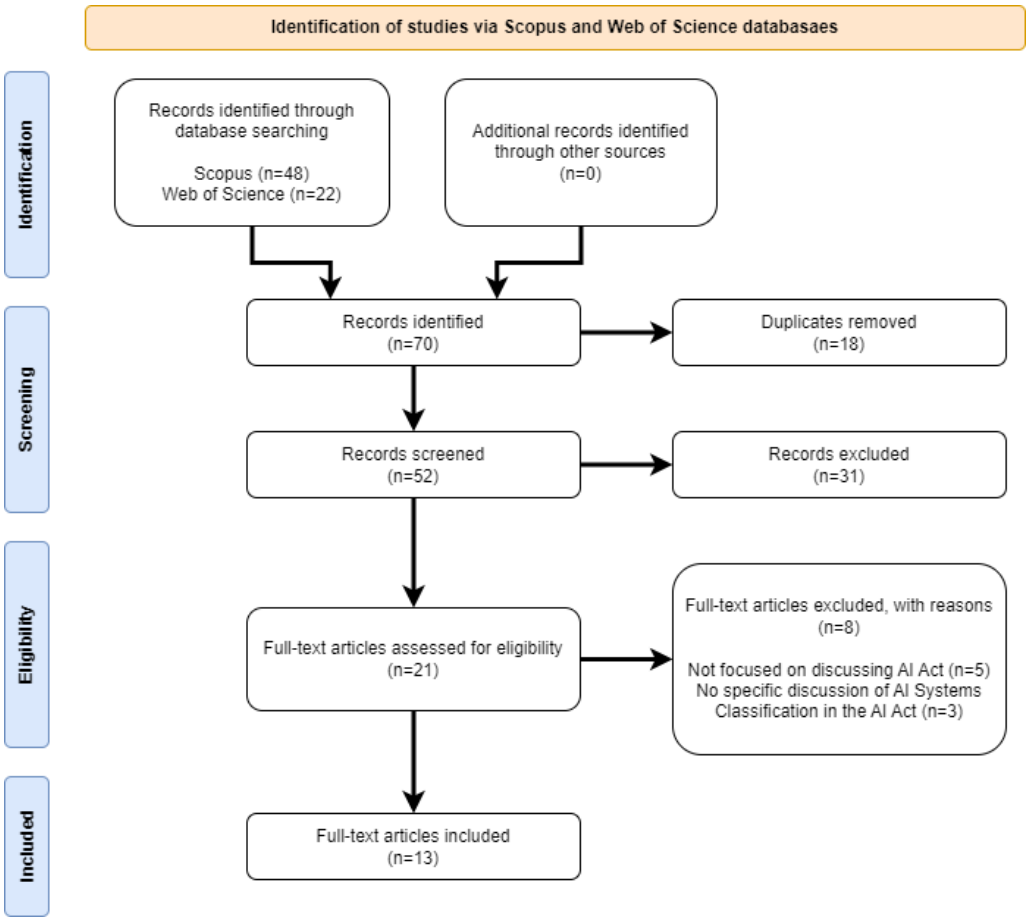

Figure 2.1: PRISM Flow Diagram of the Selection Process from Scopus and Web of Science Databases

The initial search yielded 48 articles from Scopus and 22 from Web of Science, which were subsequently screened for duplicates and exclusions. Selection criteria were refined based on factors such as Publication Status, Source Type, Language, Document Type, and Publication Year, the latter being set from 2021, coinciding with the initial proposal of the AI Act, to 2023. By applying these filters, 31 papers were excluded from the records.

The next step involved conducting title and abstract screening to determine the eligibility of papers discussing the AI Act, particularly in relation to AI Systems Classification, resulting in the identification of 13 papers for further analysis.

2.3.2. Selection Result

The shortlisted papers were then assessed by classifying relevant information related to the AI Act. Afterward, the data gathered during the classification process was synthesized to present the overview of academic discussion related to the body of the AI Act, specifically to the classification of AI systems under the AI Act. The findings will be discussed further in the following sub-section.

Table 2.4 summarized the discussion of the chosen 13 papers related to the AI Act.

Table 2.4: Literature Review of the AI Act Discussion

Author	Discussion of the AI Act
Lim et al. (2022) [26]	Apply the EU's AI regulatory legislation to cases such as accidents or incidents that emerged as social problems by AI, and analyze whether the legislative and regulatory measures of AI are valid and effective.
van Dijck (2022) [46]	Explores the impact the AI Act is expected to have on quantitative risk assessment deployed in the criminal justice system.
Barkane (2022) [5]	Argues that the proposed classification of AI biometric surveillance systems should be reconsidered to address risks to fundamental rights meaningfully.
Marano et al. (2023) [28]	Discuss that robo-advisors have the potential to pose substantial risks that should be regulated and corrected by legal instruments.
Kieseberg (2022) [21]	Provides guidelines for procuring AI-based systems that support the decision maker in identifying the critical elements for procuring secure AI systems, depending on the respective technical and regulatory environment.
Shumilo (2021) [42]	Identifies the core challenges for the EU policy on the use of AI, as well as the milestones of developing the holistic legislative proposal, and clarifying if the proposal indeed solves all the AI-related risks for future generations.
Hupont et al. (2022) [18]	Assess the risk level conveyed by each application according to the AI Act and reflect on current research, technical and societal challenges toward trustworthy facial processing systems.
Mökander et al. (2022) [30]	Discuss the two primary enforcement mechanisms proposed in the AIA: the conformity assessments that providers of high-risk AI systems are expected to conduct and the post-market monitoring plans that providers must establish to document the performance of High-Risk AI systems throughout their lifetimes.
Neuwirth (2023) [34]	Argues that the proposed regulatory approach appears problematic given the four categories' inherent interrelatedness and the numerous possibilities for their mutual combination and entwinement.
Orlando (2022) [35]	Assess whether the current regulatory framework will hold up against the increasingly widespread and disparate uses of AI systems in the field of sport.
Hacker (2021) [17]	Discusses concrete guidelines for re-using personal data for AI training purposes under the GDPR.
De Cooman (2022) [12]	Highlights even though the focus of the AI Act is regulating high-risk AI systems does not mean the residual category displays non-high-risk. This article argues that some AI systems still exhibit high risks, although excluded from this qualification.
Sovrano et al. (2021) [43]	Discuss the interplay between metrics used to measure the explainability of the AI systems and the proposed EU Artificial Intelligence Act.

2.3.3. Academic Discussions on the AI Act

This literature review aims to understand the current academic discourse related to the AI Act and identify key discussions relevant to this research. Four noteworthy discussions have emerged from the analysis. Firstly, scholars have explored the implications of the AI Act in various industries and use cases, shedding light on the challenges and opportunities it presents. Secondly, there is a pressing need for a clear differentiation between High-Risk and non-High-Risk AI systems. Additionally, lack of discussion for certain use cases within the AI Act has sparked debates regarding those particular AI systems classification. Lastly, scholars have highlighted the necessity of clear requirements and obligations to facilitate effective implementation and enforcement of the AI Act. By examining these discussions, this research contributes to the ongoing academic discourse on the need for a clear division between risk classification.

The AI Act implication in a variety of industries and use cases

Most academic discussions surrounding the AI Act focus on its potential impact on specific industries or use cases. Lim et al. [26] apply the AI Act to cases such as accidents or incidents that emerged as social problems by AI, while van Dijck [46] assesses its use in predicting recidivism risk in the criminal justice system. Barkane [5] questions the effectiveness of the AI Act in addressing risks posed by AI biometric surveillance systems. Similarly, Marano & Li [28] evaluate the potential risks associated with robo-advisors in insurance. Another specific case is the applicability of the AI Act in sports and whether it can keep up with the increasing use of AI systems [35].

Kieseberg [21] provides guidelines for the procurement of secure AI-based systems according to the AI Act. Hupont et al. [18] discuss the landscape of facial processing applications in the context of the AI Act and the development of trustworthy systems. Additionally, Shumilo & Kerikmäe [42] assesses the European approach to AI policy and its ability to address the future risks posed by AI. This research maps the core challenges for the EU policy on the use of AI, as well as the milestones of developing the holistic legislative proposal, and clarifies if the AI Act solves all the AI-related risks for future generations.

The need for a clear division between High-Risk and non-High-Risk

One of the research gaps is the need for a clearer division of High-Risk AI with other categories, as there are still many uncertainties in classifying AI systems. Lim et al. [26] argue that clear classification of AI is necessary to manage them through separate regulatory legislation. For instance, it would be reasonable to distinguish chatbot services from the risks of products such as autonomous vehicles. Barkane [5] suggests that the proposed classification of AI systems should be reconsidered since there are multiple exceptions and loopholes that should be closed in prohibited AI.

In addition, Hupont et al. [18] find that many High-Risk AI systems in the market fall under both High-Risk and Limited Risk categories, which implies unclear differentiation of current risk-level of AI systems classification. Another research also discovered proposed regulatory AI Act approach appears problematic given the four categories' inherent interrelatedness and the numerous possibilities for their mutual combination and entwinement [34]. This interrelatedness also mentioned by Orlando [35] who stated distinction between High-Risk and Low-Risk systems now seems quite rooted.

Some emerging use cases of AI systems are not categorized under the EU AI Act

According to Marano et al. [28], the AI Act draft need to address the emerging risks associated with robo-advisors in the insurance industry. They highlight the absence of concrete guidelines specifically tailored to the use of robo-advisors in insurance, emphasizing the need for adequate regulatory measures. While the AI Act has the potential to mitigate risks in insurance distribution when robo-advisors are deployed, the authors argue that the level of risk generated by these technologies should be proportionately considered. Therefore, they propose that robo-advisors equipped with a risk assessment function, when used in insurance distribution, should be classified as high-risk AI systems.

Additionally, there is a notable gap in the academic discourse regarding the flexibility of the AI Act classification to accommodate emerging technologies in the future. This flexibility is crucial for ensuring the protection of fundamental human rights while fostering innovation. Future discussions and considerations in this area are necessary for the effective and adaptive regulation of AI systems.

De Cooman [12] explains the concern of the AI Act regulation which solely focus on regulating High-Risk AI. De Cooman [12] emphasizes that while non-high-risk AI systems are not explicitly regulated by the AI Act, it does not negate the potential harm they can pose to individuals. The limited scope of the AI Act's application diminishes its overall effectiveness. Nonetheless, De Cooman [12] concludes by asserting that, overall, the proposal for the AI Act is reasonably satisfactory.

The need for clear requirements for the obligations

Barkane [5] argues that the AI Act should introduce stronger legal requirements such as third-party conformity assessment, fundamental rights impact assessment, and transparency obligations. Similarly, Mökander et al. [30] highlight the need to translate vague concepts into verifiable criteria and to strengthen institutional safeguards concerning conformity assessments based on internal checks. Hacker [17] discusses similar suggestions as the latter for a concrete specification of the criteria regarding reuse data as training data. Finally, Sovrano et al. [43] discuss the interplay between metrics used to measure the explainability of the AI systems and the proposed AI Act. This research, in the end, suggests more quantitative analysis of the metrics to

evaluate the AI systems to measure the explainability endorsed by the proposed AI Act.

2.4. Research Question and Scope

The previous section mentioned research gaps related to the AI Act, which mentions the need for clear distinctions between the classification criteria of the AI systems, including the emerging use cases not yet addressed by the AI Act. Therefore, in this study we want to propose a framework that would lead to the improvement of AI systems classification, according to the AI Act and other relevant literature. The main research question to be addressed is:

To what extent can the process of AI systems classification under the AI Act be improved?

The study is focused on the classification of AI systems within the scope of the AI definition mentioned in the AI Act. Additionally, the focus of this research is specifically on the classification process within the AI Act. Thus, the provisions of the AI Act that will be explored include Article 5 (Prohibited Risk), Article 6 (Classification Rules for High-Risk AI systems), Article 52 (Transparency Obligations), Article 69 (Codes of Conduct) and Annexes II and III. The definition of AI systems used in this research is based on the definition provided in Title I, Article 3 of the AI Act. This research is conducted based on the latest amendment to the AI Act proposal, published in May, 2023 [11], by the time this thesis report is made.

2.5. Research Methods and Sub-Questions

In order to address the main research question, more detailed sub-questions shall be elaborated. To formulate the sub-questions, a Design Science Research Methodology is used. The Design Science Methodology is the scientific study and creation of artifacts to solve practical problems of general interest [19]. The DSM (Design Science Methodology) approach is suitable for this study because this study aims to develop a decision tree framework (which can be referred to as an artifact) that will help to improve the classification accuracy of AI systems according to the AI Act. Moreover, the project design considers social and technological perspectives, addressing the differentiation of AI systems based on their risks to human rights and aiming to provide a proposed decision tree framework for classifying AI systems under the AI Act as the artifact.

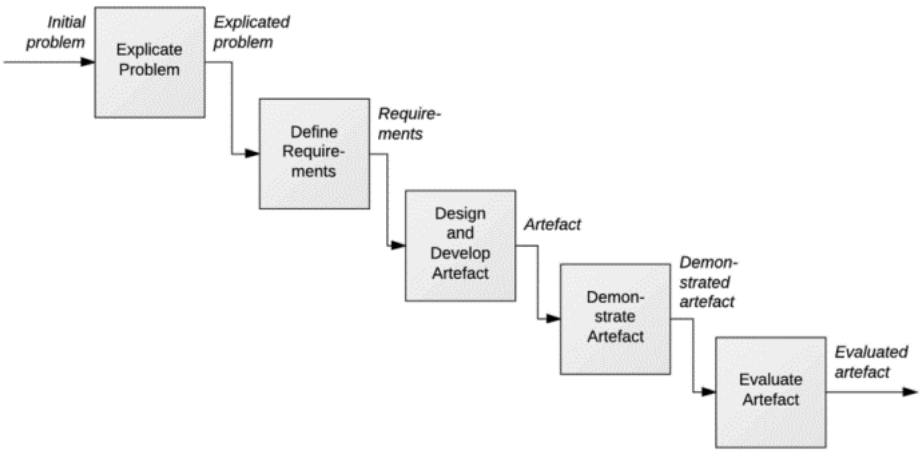

Figure 2.2: Overview of the Method Framework for Design Science Research Methodology (Johannesson & Perjons, 2021, p. 80) [19]

To meet the criteria of Design Science, as outlined by Johannesson & Perjons [19], this thesis project has to fulfill three conditions. First, this research has to establish research strategy to investigate the problem situation, elicit requirements, and employ appropriate data collection and results analysis methods. An evaluation strategy is also considered to improve the artefact. The overview of the Design Science Research Methodology is depicted in Figure 2.2. Second, this thesis has to relate to the existing knowledge, which in this case pertains to the AI systems classification in the AI Act. Lastly, the research findings are disseminated to the public.

The Design Science Methodologies comprises several steps, from problem identification to the communica-

tion of the result [19, 37]. Each stage of the DSM framework, as depicted in (Figure 2.2), corresponds to each sub-question as follows:

1. *Problem Explication.* This activity identifies specific research problem and justifies the value of a solution. This stage is the starting point for understanding the AI systems classification and its current state-of-the-art, exploring the existing classification and potential challenges. Therefore, SQ 1 is designed to know the existing AI systems classification in the AI Act and its challenges.

   SQ 1: What is the existing AI systems classification, and what are possible challenges to the current classification?.

   A literature review of all scientific articles revealed that there is a lack of research regarding the AI systems classification according to the AI Act. Therefore, this study will first provide AI systems classification based on the AI Act, followed by an analysis of potential challenges to the current classification. A desk research approach is used to answer this sub-question, including literature review from scientific and non-scientific sources.

2. *Defining requirements of the solution.* The requirements should be inferred rationally from the problem specification. In this stage, the research will focus on answering:

   SQ 2: What are features that differentiate each level of the AI Act classification?

   SQ2 of the research aims to gather abstraction of features that differentiate each level of AI systems classification in the AI Act. This step takes into account both the potential challenges (from SQ 1) and insights from discussion with legal experts.

3. *Design and Develop Artifact.* This step entails generating the artifact in an iterative process. Corresponding to the design and development of the artifact, SQ 3 is formulated as follows:

   SQ 3: What possible framework can be designed to improve the classification process of AI systems?

   This third sub-question focuses on providing a framework for better classifying AI systems for the current classification. The framework is designed based on the possible factors that differentiate the AI systems classification (the output of SQ 2) and take into consideration the current AI systems classification and potential challenges (the output of SQ 1) with a desk research approach.

4. *Demonstrate and Evaluate Artifact.* This step demonstrates the use of the artifact to solve the problem. The demonstration could involve experimentation, simulation, case studies, or other appropriate activity. The artifact is demonstrated during the interview process with respondents. Then, after the demonstration, an evaluation is conducted to measure how well the artifact supports a solution to the problem. SQ 4 is presented to be able to evaluate the proposed framework.

   SQ 4: How to evaluate the proposed framework and what improvements can be drawn from the evaluation?

   The fourth sub-question involves evaluating the proposed framework (SQ 3) through qualitative interviews with two group of experts (with legal and technical background). The results of the experiments will be analyzed to improve the proposed framework.

2.6. Research Flow Diagram

The research flow diagram, as depicted in Figure 2.3, outlines the application of the Design Science Research methodology to address each sub-question in a structured manner. The first three chapters introduce the research (Chapter 1), problem (Chapter 2), and research methodology (Chapter 3). Each sub-question corresponds to a specific chapter in this report. Chapter 4 is dedicated to answering SQ 1, while Chapter 5 provides a detailed elaboration on SQ 2. The development of the framework, which corresponds to SQ 6, is presented in Chapter 6. The subsequent chapters, Chapter 7 and Chapter 8, focus on the evaluation of the proposed framework. Finally, Chapter 9 concludes the study by discussing its limitations, providing insights gained from the research, and presenting the overall conclusions.

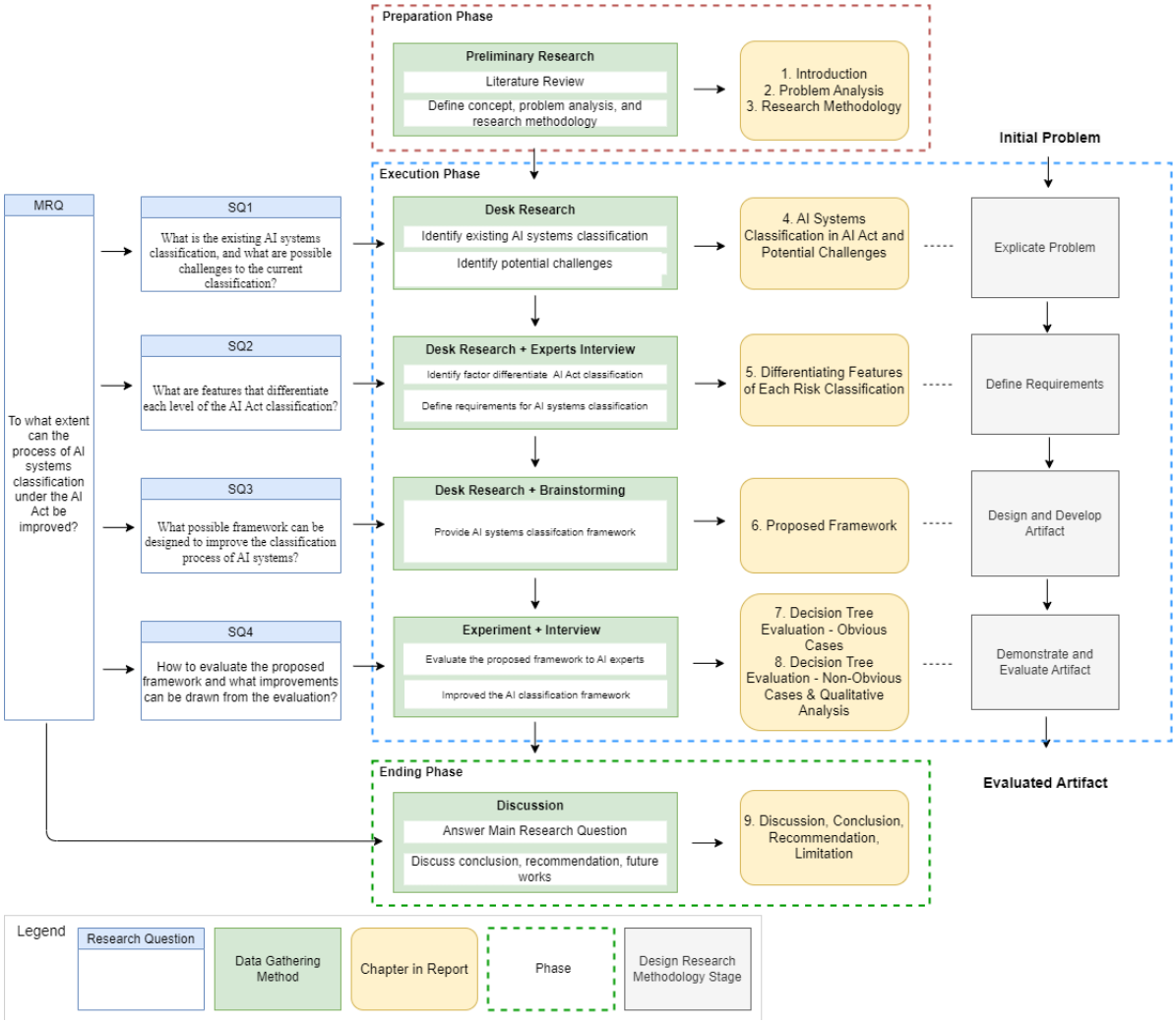

Figure 2.3: Research Flow Diagram

Table 2.5 summarizes the research sub-questions and the deliverable of each sub-questions. The research strategy, data collection methods, and expected deliverables for each sub-questions are listed in the table, with further explanations provided in subsequent chapter.

Table 2.5: Research Questions and Deliverables

	Question	Research Strategy	Sources	Data Collection Method	Deliverables
MQ	To what extent can the process of AI systems classification under the AI Act can be improved?				
SQ1	What is the existing AI systems classification according to the AI Act, and what are the possible challenges to the current classification?	Desk Research	Documents	Literature review on the AI Act and possible challenges to the current classification	The existing AI Act classification. The challenge of the existing AI Act.
SQ2	What are features that differentiate each level of the AI Act classification?	Desk Research, Interview	Documents, Expert Discussion	Literature on features to differentiate AI systems classification. Interview to legal experts to gather insight/feedback	List of factors to differentiate AI systems classification that will be used in the classification algorithm.
SQ3	What is the decision tree framework to improve classification of the AI systems in the AI Act?	Desk Research, Brainstorming	Documents	Literature review of AI systems classification, based on SQ1 and SQ2.	Framework to classify AI systems (e.g., algorithm, decision model, or taxonomy).
SQ4	How to evaluate the proposed framework and what improvements can be drawn from the evaluation?	Interview, Analysis	Expert Discussion, Result Analysis	Qualitative experiment with experts, conducted in an interview. The respondent will be given a set of cases where they must classify AI systems. Their answers in classifying AI systems with and without decision tree framework (SQ3) will be analyzed.	Improvement for AI systems classification. Feedback for the Decision Tree.

3

Research Methodology

This chapter describes the research methodology implemented in this study. This study applies Desk Research method to collect data, especially for SQ1 and SQ2. Then, interviews to several legal experts are organized in order to give more understanding on the AI Act Systems Classification, specifically to address SQ2. The decision tree framework for SQ3 is developed through Desk Research and brainstorming sessions in Miro. The framework undergoes several iterations based on insights obtained from the discussions with legal experts. Finally, the validation of the decision tree framework and the answer to the Main Research Question *"To what extent can the process of AI systems classification under the AI Act be improved?"*, are accomplished through an experiment conducted with AI experts having diverse background.

Section 3.1 explains the Desk Research methodology used to address SQ1. Section 3.2 describes the data collection for SQ2, involving Desk Research and Legal Experts Interview. Section 3.3 provides information on the process of generating the decision tree framework. Section 3.4 explains the organisation of the interview sessions conducted to validate the framework, including how the results from the interview sessions are analyzed to derive insights presented on Chapter 7 and 8.

3.1. SQ1 Research Method

The present study employs a qualitative research approach, specifically utilizing **desk research** as the primary approach. Desk research is a qualitative method that involves utilizing existing material without direct contact with the research object. The literature survey is one variant of desk research that are commonly used in research studies [48]. However, it should be noted that the quality of data obtained from a literature review may not be uniform [49]. To ensure that the data obtained is of sufficient quality, the primary search engine utilized in this study will be Scopus and Web of Science [38].

The desk research for SQ1 involves reviewing academic articles, policy briefs, and whitepapers from organizations discussing the AI Act. The output of SQ1 encompassess two main aspects (1) an understanding of the existing AI systems classification, and (2) identification of potential challenges related to the current AI Act. This SQ1 discusses each risk classification, including explanation and some examples. Furthermore, it explores the potential challenge of current AI systems classification as mentioned in Chapter 2.3. It is worth noting that the research also identifies borderline cases of AI systems that potentially fall under two or more categories. These borderline cases are employed in the interview session (SQ3) alongside obvious cases (cases which already clear falls within certain classification).

The result of this SQ1 research approach is presented in Chapter 4.

3.2. SQ2 Research Method

After gaining an understanding of the existing AI systems classification and its potential challenges, another **desk research** is conducted to comprehend specific criteria that allows one to classify AI applications in the AI Act. The main document in this research is the AI Act draft itself, with additional information from academic literature.

Then, a **thematic analysis** is resulted from the abstraction and interpretation process in understanding the criteria of each risk class in the AI Act. The process is facilitated by the Miro platform, enabling the visualization of specific criteria associated with each risk class. These criteria are crucial for generating the proposed framework (SQ3).

The analysis begins by selecting meaningful units, condensing and coding them, and organizing them into categories and themes [27]. Initially, all meaning units associated with each classification are selected. Subsequently, condensing and coding techniques are applied to eliminate repetitive and irrelevant words. Condensing the original text involves succinctly removing repetitive and non-essential words, maintaining the core content of the meaning unit. The subsequent coding process entails labeling the condensed meaning units with descriptive codes closely aligned with the original text at a low level of abstraction and interpretation.

Finally, categories are created by sorting and grouping related codes that distinguish themselves from other code groups. However, it is important to note that this process may be susceptible to misunderstandings, particularly when interpreting the AI Act, which primarily operates within a legal context.

In light of the legal terms and definitions found in the AI Act, which can sometimes be designed in an open context, discussions with legal experts are conducted to minimize misunderstandings and clarify any poorly understood contexts.

The discussion with the legal experts takes the form of **semi-structured interviews**, during which the following questions are posed:

1. What are the primary factors to determine AI system's classification as prohibited, high-risk, limited and no/minimal risk?
2. What are the key features of each class?
3. Why are high-risk AI systems categorized in specific domains? If the AI system is not mentioned in the Annexes, which class should this system belong to?
4. In case of prohibited risks, exemptions are made for certain biometric systems. To which class these systems should belong to?

These interviews contribute to a comprehensive understanding of the differentiation of each risk level in AI Act. The output of the SQ2 is discussed in Chapter 5.

3.3. SQ3 Research Method

The primary research strategy to address SQ3 is **desk research** and **brainstorming**. It begins with utilizing the output of SQ2, establishing design principle for the decision tree, and finally generating the decision tree framework. This process is facilitated using the collaborative platform, Miro. The overall process of generating decision tree is visualized in Figure 3.1.

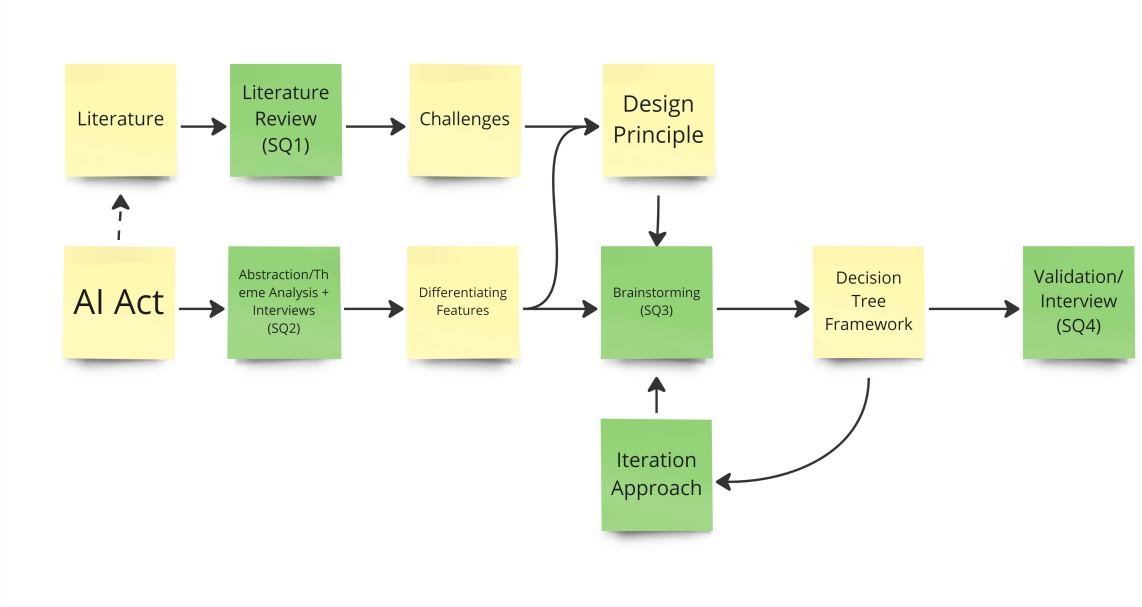

Figure 3.1: Flow of Generating Decision Tree

The design principles served as guiding principles in supporting the design of artifacts, specifically the decision tree, at a higher level [31]. This ensured that the design aligned with the research objectives and effectively addressed the identified knowledge gaps. These principles provided guidelines facilitating a structured approach to organizing the classification criteria, defining decision points, and determining the decision tree's flow. The design principle itself was derived from a suitable knowledge base [31], including an understanding of the existing AI systems classification from literature and insights obtained from expert interviews. The design principle in this research adheres to formulated design principles, which offer prescriptive knowledge regarding actions and the material properties of an artifact in terms of both form and function, within specified boundary conditions [7].

Once the design principles are established, the decision tree framework is generated through a creative brainstorming process within the Miro platform. This process takes into consideration the challenges identified in SQ1, as well as the distinctive features of each risk class outlined in SQ2, while adhering to the guideline of design principles. The resulting decision tree and its rationale are further explained in Chapter 6. The rationale is derived from the design principles and challenges that arise from the existing AI systems classification. An iterative approach is applied to ensure that the decision tree is ready for validation by the respondents. This iteration incorporates insight from some early respondents and experts to improve the proposed decision tree before doing the evaluation part.

3.4. SQ4 Research Method

The data collection process for SQ4 entails conducting **interviews** to selected respondents who possess expertise in the relevant fields. The interviews are structured in a semi-structured format, allowing for flexibility while ensuring key areas of interest are explored [16]. The primary objective of the interviews is to gain valuable insights from the respondents that can be used to validate the decision tree framework.

To gain a comprehensive understanding, the interview session is divided into three sections, which will be explained in the subsequent sub-sections. This structured approach allows a systematic exploration to evaluate the effectiveness of the decision tree and an in-depth understanding of respondents' perceptions related to the decision tree and classification process.

It is important to acknowledge that the limitation of interviews is the subjective nature of the data obtained. The responses provided by respondents may be influenced by their perceptions and biases [3]. However, employing a semi-structured interview format makes it possible to delve deeper into the interviewees' responses and gain a more in-depth understanding of their perspectives [23]. Furthermore, the expert feedback and comments will be analyzed and interpreted to develop more comprehensive recommendations for refining the decision tree framework.

3.4.1. Overview of Respondents

The respondents are chosen based on specific criteria, ensuring their ability to provide valuable insights into the classification of AI systems. We conducted 16 interviews with respondents from different backgrounds representing two groups: legal and technical expertise in AI. The respondents participating in the research and their backgrounds are listed in Table 3.1.

Table 3.1: Overview of Respondents

Respondents ID	Background
A	Legal
B	Technical (AI financial technology)
C	Legal
D	Technical (Cybersecurity/AI enthusiast)
E	Technical (Generative AI)
F	Legal
G	Technical (AI recommender system)
H	Technical (Cybersecurity/Ethics)
I	Legal (AI in content moderation)
J	Technical (ML/Ethics in Autonomous Vehicle)
K	Legal (Biometric Identification System)
L	Technical (AI/medical system)
M	Technical (AI/ethics & philosophy of technology)
N	Legal
O	Legal
P	Technical (NLP Researcher)

3.4.2. Selection Methodology

The selection of respondents for the interviews was conducted with careful consideration to ensure a comprehensive and diverse range of perspectives. The target number of respondents was 16 (multiple of 8), chosen to align with the total number of AI systems use-cases included in the interview session. This approach allowed for a balanced evaluation of each use case, ensuring adequate coverage and representation.

Multiple approaches were conducted to finalize the selection of the 16 respondents. First, it was essential that the respondents had expertise in AI-related fields in order to evaluate the decision tree framework for AI systems classification effectively. Second, respondents were chosen from legal and technical perspectives to gain a more comprehensive understanding of the classification process. This multi-disciplinary approach would contribute to a richer evaluation of the decision tree framework.

To reach potential respondents, an open invitation for the interview was posted on LinkedIn and promoted within relevant research groups and mailing lists. Additionally, a referral strategy was implemented to expand the network of AI experts, particularly those with a legal background. Over 40 personal invitations were sent via email and LinkedIn to individuals who met the aforementioned criteria.

The selection criteria for the respondents were as follows: (1) working in AI-related fields, (2) residing in the EU region, and (3) employed by an organization or company within the EU. By ensuring diversity in the respondents' backgrounds, their insights can contribute to a more comprehensive evaluation of the decision tree, as different perspectives can shed light on alternative ways of classifying AI systems. However, it is important to acknowledge that the respondents' expertise or background may include certain biases in their perception and understanding of the AI systems classification under the AI Act.

3.4.3. Interview Setup

The one-on-one interview session were conducted online, following the Interview Protocol (Appendix A), with a duration of approximately 60 to 90 minutes per respondent. The interviews were conducted using the Miro platform, which provided a collaborative and visual environment for effective communication and interaction between the interviewer and interviewee (respondent). Detailed visuals of the interview boards can be found in Figure A.1 and Figure A.2 in Appendix A.

The interview itself was structured into three sections, excluding the introduction, each serving a distinct purpose. In the first section, the respondent was presented with four AI systems use-cases and asked to classify them by referring to the relevant AI Act Article. The second section involved classifying the remaining four AI systems use-cases using the decision tree framework. Within these two sections, follow-up questions were asked to the respondents, seeking clarification and further insights into their classification choices. Finally, the third section consisted of open-ended questions aimed to gather additional insights and perspectives

from the respondents.

Table 3.2: Datacode Usecase

Use Cases	Risk Category	Case Category	Case
1	Minimal/No Risk	Obvious	AI system to filter unwanted mails and keep them separated from useful emails to reduce time and effort
2	High-Risk	Obvious	AI system use emotion recognition system to identify/recognize patient's emotion
3	Unacceptable Risk	Non-Obvious	AI system to measure a truck driver's fatigue and playing a sound to push them to drive longer
4	High-Risk/Limited Risk	Non-Obvious	AI systems designed for social robots for children with autism to capture their behavior to assist treatment
5	High-Risk/Minimal/No Risk	Non-Obvious	AI systems for automatic transcription or enhancement of speech
6	High-Risk	Non-Obvious	AI systems to assess recidivism risk by providing quantitative risk assessments
7	Unacceptable Risk	Obvious	AI systems using remote biometric identification of political protesters creates a significant chilling effect on the exercise of freedom of assembly and association
8	Limited Risk	Obvious	AI system that automatically converse with people in place for a human being and can interact with them

A set of use cases of AI systems is provided in Table 3.2. The rationale behind each use case is discussed in Chapter 4. Each respondent was required to classify a total of eight use cases, with four use cases being classified in the first section and the remaining four in the second section. To prevent potential bias, the order of the use cases was distributed as shown in Table 3.3. This ensured a balanced and unbiased approach to the classification process, as the respondents were not influenced by the order of the AI systems after classifying them in the first section.

Table 3.3: Respondents and Responding Use-Cases

Respondent ID	Use Cases							
	Without DT				With DT			
A	1	2	3	4	5	6	7	8
B	8	1	2	3	4	5	6	7
C	7	8	1	2	3	4	5	6
D	6	7	8	1	2	3	4	5
E	5	6	7	8	1	2	3	4
F	4	5	6	7	8	1	2	3
G	3	4	5	6	7	8	1	2
H	2	3	4	5	6	7	8	1
I	1	2	3	4	5	6	7	8
J	8	1	2	3	4	5	6	7
K	7	8	1	2	3	4	5	6
L	6	7	8	1	2	3	4	5
M	5	6	7	8	1	2	3	4
N	4	5	6	7	8	1	2	3
O	3	4	5	6	7	8	1	2
P	2	3	4	5	6	7	8	1

The question for the semi-structured interview in the third section are as follows:

a. Decision Tree

1. Is the given decision tree framework helpful to categorize the AI systems? Yes/No, why?
2. How did your perception or understanding of the AI systems change after using the decision tree framework?

3. In your opinion, what were the strengths or weaknesses (side effects) of the decision tree framework?
4. Is the step by step in the decision tree clear and easy to understand? Why?
5. Did the decision tree provide any additional insights or guidance in making your classification?
6. Do you have any suggestions to improve decision tree?

b. Use Cases

1. Were there any particular use cases that you found difficult to classify, either with or without framework? If so, why?
2. Which cases do you think is the easiest to classsify? Why?
3. What challenges or difficulties did you encounter while classifying the AI systems?
4. What are other potential use cases of AI systems within your expertises? Any concerns?
5. Where do you think Large Language Model use-cases should fall under? Why? What things do you need to consider in classifying this case?

These questions were designed to delve deeper into respondents' perspective and gain a comprehensive understanding of their insights. Spontaneous questions during the interviews were also asked to explore respondents' answers in greater detail.

3.4.4. Analysis of Interviews

The analysis of the interview results employs two different approaches, as listed in Table 3.4. The first approach focuses on quantitative analysis of the classification outcomes from the first and second sections of the interviews. The second approach is more qualitative analysis methodology.

Table 3.4: Analysis of Interview

Analysis	Case Type	Method/Measurement	Purpose
Quantitative Analysis	Obvious Case	Inter-rate Agreement Metrics (Krippendorff's Alpha)	To evaluate the performance of the decision tree framework, to assess the reliability of the decision tree framework, to assess the level of agreement between experts
		Accuracy and Confusion Matrix (Precision, Recall, F-1 score)	To evaluate the performance of the decision tree framework, in terms of accuracy, precision, F1-score, and recall)
		Time performance	To evaluate duration of classification with and without decision tree framework
	Non-Obvious Case	Inter-rate Agreement Metrics (Krippendorff's Alpha, Agreement Table)	To evaluate the performance of the decision tree framework, to assess the reliability of the decision tree framework, to address the level of agreement between experts
		Time performance	To evaluate duration of classification with and without decision tree framework
Qualitative Analysis	All Cases	Coding and categorizing	To gain qualitative insights from the experiment (interview)

3.4.4.1. Quantitative Analysis

The first approach involves aggregating interview responses and organizing them into an agreement table to assess the decision tree's performance in terms of accuracy and reliability.

In assessing a classification system's validity, as stated by [13], it is essential to examine the accuracy and reliability of the instrument. The classification framework should correctly classify use cases (accuracy) and do so consistently across multiple instances (repeatability) and among different individuals (reproducibility). However, for this study, only the reproducibility of the decision tree can be measured, as the experimentation results cannot accommodate repeatability measurement.

a. Accuracy

The accuracy of the decision tree performance was measured by comparing all the respondents' responses in the experiment to the ground truth as presented in Table 4.1.

Precision, recall, and **F-1 score** will complement the accuracy measurement to evaluate the decision tree's accuracy comprehensively.

Accuracy measurement can only be applied to Obvious cases, assuming they possess clear ground truth as specified in the AI Act. Conversely, accuracy measurement can not be measured for Non-Obvious cases due to the absence of ground truth for comparison in the evaluation.

b. Reproducibility

The reproducibility of the decision tree framework can be assessed for both Obvious and Non-Obvious cases using Inter-Rater Reliability (IRR) measurement, specifically Inter-rate Agreement metrics computed through Krippendorff's Alpha.

Krippendorff's Alpha coefficient is an efficient tool for assessing reliability among raters [41] [50]. This metric is suitable for this experiment as it accommodates multiple respondents, multiple subjects (case studies), and missing ratings (as not all participants classified all eight case studies using the Decision Tree framework) [13].

The Inter-Rater Reliability (IRR) measurement evaluates the decision tree framework's reproducibility (ensuring consistent results from different respondents) [13]. Furthermore, it provides insights into the level of agreement among experts, offering valuable guidance for improving the decision tree. The high inter-rater agreement indicates that the raters closely agree, instilling confidence in the classifications' reliability and the decision tree's effectiveness. These metrics are calculated for both Obvious and Non-Obvious case categories.

c. Time-efficiency

Additionally, to deepen the analysis, the time efficiency of the decision tree is also measured by comparing the time differences when respondents classify AI systems with and without the decision tree.

Through these three measures: accuracy, reliability (reproducibility), and time efficiency, the classification of AI systems aims to be more effective and efficient.

Acknowledging that these quantitative analyses may not achieve statistical significance due to the limited sample size of 16 respondents is crucial. Nevertheless, they offer valuable additional insights for evaluating the decision tree and comprehending the respondents' perspectives on AI systems classification. The individual backgrounds of the respondents are also considered in the analysis. A more detailed exploration of this initial approach is found in Chapters 7 and 8.

3.4.4.2. Qualitative Analysis

The second approach employs a qualitative analysis methodology involving coding each interview transcript. Despite varying perspectives expressed by the respondents, specific themes that span multiple interviews emerge. This thematic analysis is facilitated by using Atlas.ti software, aiding in encoding different topics discussed by the respondents. The interview transcripts are examined, allowing the identification of recurring themes/topics, patterns, and valuable insights derived from the discussions.

The synthesis of the analysis conducted through the first and second approaches will be further discussed in Chapters 7 and 8. Combining these two approaches, the research seeks to comprehensively evaluate the interview data, leading to refined recommendations and enhancements for the decision tree framework.

4

AI Systems Classification in AI Act and Potential Challenges

This chapter focuses on addressing the first sub-question: "What is the existing AI systems classification, and what are the possible challenges to the current classification?". This sub-question aligns with the 'Problem Explication' stage of Design Science Methodology. The chapter is structured into two main sections.

The first section provides an overview of the existing AI systems classification under the AI Act, with a focus on the latest amendment in May 2023. This section explains each risk level and presents relevant use cases to illustrate the classification.. The second section discusses the challenges that may arise from the current classification of AI systems. One primary challenge discussed in this section is the need for more clarity in the classification of AI systems under the AI Act. It also explores borderline cases (non-obvious cases) where AI systems may fall into multiple categories due to the lack of clear distinctions between risk levels within the AI Act. Alongside these non-obvious cases, the chapter also presents obvious cases. Selected use cases from both categories will facilitate the evaluation stage in addressing SQ4.

Section 4.1 provides an overview of risk-based classification of AI systems, with detailed explanations of each risk level covered in separate sub-sections (Sub-section 4.1.1 to Sub-section 4.1.4). Section 4.2 presents the challenges that emerge from the current classification of AI systems, drawing insights from various references. Sub-section 4.2.1.2 and Sub-section 4.2.1.1 outline potential non-obvious and obvious use cases of AI systems, including the selection of several use cases for the evaluation part of this research. Section 4.3 analyzes the output of SQ1.

4.1. AI Act Risk Classification

The importance of adopting a multi-layered risk-based approach for AI was emphasized during the public consultations following the release of the 2018 EU Draft Ethics Guidelines for Trustworthy AI and the 2019 White Paper on AI [12]. In response to these needs, the European Commission advocates for a sector-by-sector and case-by-case approach to regulate AI, rather than a one-size-fits-all or blanket approach [12]. Consistent with this risk-based approach, the proposed AI Act introduces a categorization of AI systems according to four different risk levels: 1) unacceptable risk (Title II); 2) high risk (Title III); 3) limited risk (Title IV); and 4) minimal risk (Title IX) [5] [14].

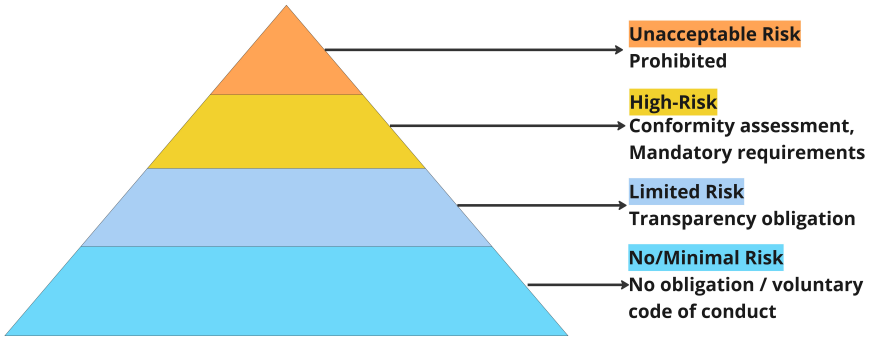

Figure 4.1: Level of Risk (Kop, 2021, p. 3) [22]

To gain a better understanding of the risk categorization in the proposed AI Act, a "pyramid of criticality" [22] is visualized in Figure 4.1. Stricter rules apply for each increasing level of risk associated with AI systems. Starting from the lowest level, the no/minimal risk category encompasses the majority of existing AI systems, falling outside the scope of the regulation. Moving up the pyramid, a larger number of systems fall into the limited risk category, where the only obligations imposed are to provide certain information to users. At the higher level of risk, the high-risk category includes a smaller subset of AI systems subject to various restrictions. Finally, at the top of the pyramid, AI systems with unacceptable risks are prohibited for use in the EU [6]. Further explanations of each level of risk will be provided in subsequent sub-sections.

The AI Act aims to safeguard the Fundamental Rights and Union values enshrined in the EU Charter of Fundamental Rights [9]. These rights and values encompass various aspects, such as the right to human dignity (Article 1), the respect for private life and protection of personal data (Articles 7 and 8), non discrimination (Article 21) and equality between women and men (Article 23), freedom of expression (Article 11), freedom of assembly (Article 21), the right to an effective remedy and to a fair trial, the rights of defence and the presumption of innocence (Articles 47 and 48), as well as the general principle of good administration. Additionally, the Act extends this protection to specific vulnerable groups, including worker's rights to fair and just working conditions (Article 31), a high level of consumer protection (Article 28), the rights of the child (Article 24), the integration of persons with disabilities (Article 26) and the right to a high level of environmental protection and the improvement of the quality of the environment (Article 37), including in relation to the health and safety of individuals [9].

4.1.1. Unacceptable Risk

At the top of the pyramid of criticality, AI systems with unacceptable risks entail the most severe consequences compared to other categories. These prohibited AI systems with unacceptable risks contravene the Fundamental Rights and European Union values. Any AI system falling under this category is strictly prohibited from being placed on the market, put into service, or used, with the exception of military purposes (Article 2.3). The prohibitions, outlined in Article 5 (see Appendix B.1) of the regulation, aim to protect individuals from potential harm caused by certain AI practices [9].

Title II Article 5 of the AI Act lists AI systems whose use is considered unacceptable as they contravene Union values and Fundamental Rights. The prohibitions cover practices that utilize subliminal techniques beyond individual's consciousness, known as 'dark patterns', to manipulate people's behavior (Article 5.1.a). Additionally, AI systems that exploit vulnerabilities of specific vulnerable groups, such as children or person with disabilities, in ways likely to cause physical or psychological harm (Article 5.1.b) are also prohibited. Moreover, the prohibition applies to AI systems used for social scoring by public authorities to evaluate or classify natural persons based on their social behavior (Article 5.1.c), leading to detrimental or unfavourable treatment of certain individuals or groups in unrelated social contexts (Article 5.1.c(i)), or treatment that is unjustified or disproportionate to their social behavior or its gravity (Article 5.1.c(ii)).

The use of biometric categorisation systems to classify natural persons according to sensitive or protected attributes or characteristics is not allowed in the EU, except for therapeutic purposes, as outline in Article 5.1.ba. Additionally, the use of AI systems in 'real-time' biometric identification systems in publicly accessible spaces is a prohibited practice (Article 5.1.d). The use of AI systems for the analysis of recorded footage of publicly accessible spaces through 'post' remote biometric identification systems, unless subject to a pre-judicial authorisation in accordance with Union law, is also prohibited (Article 5.1.e). The technical inaccuracies of AI systems intended for the remote biometric identification of natural persons can lead to biased results and discriminatory effects, posing a significant intrusion into the rights and freedoms of individuals concerned.

The latest amendment in May 2023 [11] has also banned several practices. Firstly, the use of AI systems for making risk assessments of natural persons or groups regarding actual or potential criminal or administrative offenses based on profiling of individuals' personality traits and characteristics is prohibited (Article 5.1.da). Secondly, the use of AI systems that employ untargeted scraping of facial images from the internet or CCTV footage to create or expand facial recognition databases (Article 5.1.db). Lastly, the use of emotion recognition system in law enforcement, border management, workplaces, and educational institutions is prohibited (Article 5.1.dc).

4.1.2. High-Risk

In accordance to the pyramid of criticality, High-Risk AI systems are classified one level below the category of unacceptable risk and are subject to regulation within the framework of the AI Act. Addressing concerns and requirements associated with High-Risk AI Systems, Articles 6 to 51 of the Act provide dedicated provisions. High-Risk AI systems are identified based on their potential to cause significant harmful impact on the health, safety, and fundamental rights of individuals in the European Union. It is important to note that, unlike prohibited practices, High-Risk AI systems themselves do not directly violate Union values but rather pose a potential threat to these values [6]. The risks presented by High-Risk AI systems to the human dignity, privacy, data protection and personal data, freedom of expression and information, non-discrimination, and other values is weighed against the benefits of the systems.

There are two main categories of High-Risk AI systems. The first category includes AI systems designed to function as safety components of products that are subject to third party ex-ante conformity assessment. These AI systems pose a high risk due to the specific characteristics of the sectors they operate in and their particular use cases [12]. The second category encompasses the standalone AI systems with fundamental right implications, explicitly listed in Annex III. This acknowledges that, regardless of the specific sector involved, there may be exceptional instances where AI applications for certain purposes classified as high-risk due to the potential harm they can cause [12].

1. AI systems intended to be used as safety component of products that are subject to third party ex-ante conformity assessment

This category is covered by Sectorial Product Legislation, and an AI system is considered high-risk if it meets two conditions. **The first condition** is that the AI system is used as a product or a safety component of a product covered by one of the EU harmonisation instruments listed in Annex II of the AI Act (Article 6.1.a). The harmonisation instruments mentioned in Annex II include machinery (Annex II.A.1), toys safety (Annex II.A.2), recreational craft and personal watercraft (Annex II.A.3), lifts and safety component of lifts (Annex II.A.4), equipment and protective systems intended to use in potentially explosive atmospheres (Annex II.A.5), radio equipment (Annex II.A.6), pressure equipment (Annex II.A.7), cableway installations (Annex II.A.8), personal protective equipment (Annex II.A.9), appliances burning gaseous fuels (Annex II.A.10), medical devices (Annex II.A.11), and in-vitro diagnostic medical devices (Annex II.A.12). Additionally, other instruments of Other Union Harmonisation Legislation (Annex II.B) related to aviation security (Annex II.B.1), marine equipment (Annex II.B.4), interoperability of the rail system (Annex II.B.5) and unmanned aircrafts and their engines, propellers, parts and remote-control equipment (Annex II.B.7), approval and market surveillance of two- or three-wheel vehicles and quadricycles (Annex II.B.2), agricultural and forestry vehicles (Annex II.B.3), and motor vehicles and their trailers are included (Annex II.B.6).

The second condition for classifying an AI system as a high-risk is when the system, whether functioning as a safety component or as a standalone product, is required to undergo a third-party conformity assessment (Article 6.1.b). This requirement aligns with the existing sector-specific regulations and aims to update and harmonize the safety legislation within the horizontal framework introduced by the AI Act [12].

2. Standalone AI systems with mainly fundamental rights implications

Standalone AI systems not covered by sectoral legislations mentioned above can still be classified as High-Risk if they fall under one or more critical areas and use cases listed in Annex III of the AI Act and pose a risk to the health, safety, or fundamental rights of natural persons (Article 7.1).

Annex III provides a list of eight categories in which an AI system may be classified as high-risk. The first category includes biometric and biometrics-based systems, including emotion recognition systems (Annex III.1). However, emotion recognition system are only qualified as high-risk when used in certain contexts such as law enforcement, border management, workplace and education institutions (Article 5.1.dc). The second category is management and operation of critical infrastructure such as road, rail and air traffic, supply of water, gas, heating, eletricity, and critical digital infrastructure (Annex III.2). The third category covers AI systems in education and vocational, including those used for admission decisions (Annex III.3.a), assessing appropriate education levels (Annex III.3.b), and monitoring student behavior during tests (Annex III.3.c). The fourth category relates to worker recruitment, application screening and evaluating candidates (Annex III.4.a), as well as AI systems to influence decisions such as promotion and termination of an employment contract, task allocation and performance monitoring (Annex III.4.b).

The fifth category concerns access to essential private services and public services and benefits, including eligibility for public assistance benefits and services (Annex III.5.a), credit scoring (Annex III.5.b), health and life insurance (Annex III.5.ba), and benefits of emergency services (Annex III.5.c). The sixth category involves the use of AI in law enforcement to support law enforcement authorities as polygraphs and similar tools, to evaluate the reliability of evidence in the investigation or prosecution of criminal offences, to profile a natural persons, and to use as crime analytics (Annex III.6). The seventh category is in the area of migration, asylum and border control management, including the use as polygraphs and similar tools (Annex III.7.a), to asses a risk such as security risk, health risk of individual who intend to enter the EU territory (Annex III.7.b). The use of this system in this area also includes the verification of documents (Annex III.7.c), the assessment of the veracity of evidence of the asylum or residence permits application (Annex III.7.d), the purpose of monitoring, surveilling or processing data in the context of border management activities (Annex III.7.da), and migration movement and border crossing forecasting (Annex III.7.db). The final category encompasses the administration of justice and democratic processes, including AI systems intended to assist a judicial authority or administrative body to do research and interpret facts and the law (Annex III.8.a), to influence the outcome of an election or referendum or voting behavior of individual (Annex III.8.aa), and to be used by large online platforms in their recommender systems (Annex III.8.ab).

It should be noted that this list of critical areas and use cases is dynamic and subject to adjustment as technology advances [6].

Furthermore, all High-Risk AI systems, including those intended to be used as safety components of products, must comply with mandatory requirements outlined in the AI Act. These requirements include risk management measures (Article 9), utilization of high-quality data and appropriate data governance practices (Article 10), thorough documentation and record-keeping for traceability (Article 11 and 12), transparency in the system's (Article 13), human oversight (Article 14), and ensuring accuracy, robustness, and cybersecurity of AI systems (Article 15). By imposing these requirements, the AI Act aim to strike a balance between harnessing the benefits of High-Risk AI systems and safeguarding fundamental rights and Union values.

4.1.3. Limited-Risk

On the third level of the pyramid of criticality, certain AI systems with limited risk are regulated by Article 52 of the AI Act. This article outlines transparency obligations for systems that (i) interact with individual, (ii) use emotion recognition system or biometric categorisation system, or (iii) generate or manipulate image, audio or video content (such as 'deep fakes'). However, this obligation does not apply to AI systems authorised by law for detecting, preventing, investigating, and prosecuting criminal offences.

For AI systems intended to interact with natural persons, they must be designed and developed in a way that individuals are aware they are interacting with an AI system (Article 52.1). For AI systems that intend to use to detect emotions or determine association with categories based on biometric data, users shall inform the individuals exposed to such system (Article 52.2). Additionally, AI systems that generates or manipulates image, audio or video content must disclose that the content has been artificially generated or manipulated (Article 52.3).

These transparency obligations aim to ensure that individuals are properly informed when engaging with AI systems, particularly regarding the recognition of their emotions or characteristics through automated means, as well as the generation of AI-generated content. Providing individuals with this information empowers them to make informed choices. The underlying risk is the potential for individulas to be misled by mistakenly believing they are interacting with another person or perceiving manipulated content as authentic. These situations can erode trust in new technologies, which is undesirable. Therefore, the transparency obligations are crucial for building trust and upholding fundamental rights within the Union [6]. It is worth noting that Article 52 clarifies that these transparency obligations apply to AI systems with limited risk, regardless of whether they are labeled as "high-risk" or not [45].

4.1.4. No/Minimal Risk

AI systems with no or minimal risk are classified at the bottom of the criticality pyramid. This category falls outside the regulatory framework of the AI Act. The majority of AI systems currently used in the EU belong to this category and pose little to no risk, such as spam filters [47]. It is important to note that referring to this category as 'minimal/no risk' rather than 'non-existent risk' is justified, as risk can never be completely eliminated in various aspects of life, incuding AI [6].

Although 'minimal risk' AI systems are outside the scope of the AI Act regulation, Article 69 of the AI Act proposes a voluntary code of conduct to be developed for these systems. This provision aims to encourage companies to adhere to the same standards of transparency, human oversight, and robustness required for High-Risk AI systems, rather than imposing mandatory obligations [47]. Even though video games or spam filters may be categorized as minimal or no risk, there is no guarantee that the data generated by these systems will not be misused or pose harm to individuals [6]. Therefore, continuous monitoring and staying up-to-date with advancements are necessary to prevent any potential misuse or harm to individuals.

4.2. Potential Challenges

This section explores the potential challenges that arise from the current classification of AI systems proposed by the European Union (EU), as discussed in Chapter 2 by scholars. One significant challenge is the need for a clearer distinction between different risk levels under the AI Act. The lack of well-defined boundaries between risk categories leads to uncertainties in classifying AI systems, resulting in borderline cases where systems may fall into multiple categories. These borderline cases are referred to as Non-Obvious cases in this report. Furthermore, alongside these non-obvious cases, there are also obvious cases that are considered during the evaluation stage in Chapter 7.

To effectively manage AI systems, it is crucial to establish a clear division between risk classes. Scholars, such as Lim et al (2022) [26], argue that distinct regulatory legislation is necessary for different AI classifications to address risks appropriately. For instance, it would be reasonable to differentiate chatbot services from the risks associated with autonomous vehicles.

There are several factors contributing to the lack of clarity between risk classes:

(1) Too narrow scope of Unacceptable Risk: The proposed classification of biometric surveillance systems in the AI Act draft, specifically for remote biometric identification systems, has a narrow scope. This narrow scope fails to meaningfully address the risks these systems pose to fundamental rights, resulting in a wide range of surveillance practices. It is recommended to reconsider the classification of biometric surveillance systems and introduce clear prohibitions in the draft AI Act before its enactment [5].

(2) Exceptions and loopholes in Prohibited AI Practices: Some exceptions and loopholes exist regarding prohibited AI practices. For instance, the use of systems in prohibited practices for law enforcement purposes, although permitted in exceptional cases, can often be easily justified.

(3) Vague wording in certain cases: Some sections of the AI Act are vaguely worded, which partially reflects essential guarantees for the use of surveillance measures. However, this vagueness may fail to prevent the use of AI systems for mass surveillance. The draft AI Act should provide clearer and more explicit provisions to address this concern [5]. Additionally, critics have described the draft AI Act as a lengthy and sometimes opaque document lacking consideration for its interaction with other laws and future enforcement [34].

(4) Distinction between High-Risk and Limited Risk categories: The classification of emotion recognition systems and biometric categorization systems as generally limited-risk AI systems, with some cases falling into the high-risk category, fails to meaningfully address the significant risks these systems pose to fundamental rights. Simply relying on transparency rules to mitigate the risks while neglecting the need for additional limitations and requirements is insufficient [5]. Hupont et. al. (2022) [18] also mentioned several facial processing applications, that might falls under High-Risk or Limited Risk depends on the context.

(5) Reliance on interpretation to assess risk class of AI systems: The interpretation of the AI Act to assess the risk class of AI systems can potentially lead to misclassification without a proper understanding of the legislation. The AI Act contains explicit and implicit references to facial processing, but it may not specifically mention situations where emotion recognition or biometric categorization systems are exploited in other high-risk use cases listed in Annex III [18].

(6) Need for more interdisciplinary approach: There is a call for a more profound and interdisciplinary approach to regulate subliminal AI systems, considering the connections between law and neuroscience. Complex questions arise, such as the moral difference between nudges and subliminal messages and the impact on individuals' freedom of choice. These questions highlight the limitations of binary approaches and emphasize the importance of a broad and interdisciplinary debate based on sound scientific findings [34].

(7) Dilemmas and paradoxes caused by dualistic or dichotomous thinking: The dualistic or dichotomous

thinking applied in the classification of high-risk and non-high-risk AI systems leads to dilemmas and paradoxes. The dichotomy between high-risk and non-high risk AI systems bring argument that non-risk AI systems does not mean the risk is actually low. In addition, the definition of High-Risk AI systems in the AI Act potentially excluding AI systems that are commonly recognized as harmful [12] [34].

The lack of clear risk classifications hinders investment and innovation, particularly in critical infrastructure, employment, law enforcement, and product safety sectors (Annex 2). The uncertainties surrounding risk classification significantly impact these areas. For example, it remains unclear whether 40% of AI systems used in enterprise functions, such as marketing, production, and purchasing, fall into the high-risk class or not [4].

Furthermore, providers that misclassify their AI system as not subject to the requirements of Title III Chapter 2 of the AI Act and place it on the market before the deadline for objection by National Supervisory Authorities shall be responsible and be subject to fines pursuant to Article 71 (Article 6.2.b).

Therefore, it is crucial to provide comprehensive guidance for the correct risk classification of AI systems, including clear instructions and examples, especially for AI systems used in generic and industry-agnostic enterprise functions.

Although the regulation has not been passed yet, there is an urgent need for clear categorization as soon as possible. This will enable businesses to predict whether their systems will be heavily regulated or not regulated at all, allowing them to adapt their planning for the coming years. Therefore, it is essential to define risk in the regulation and differentiate the proposed risk levels [6].

Provide comprehensive guidance for the correct risk classification of AI systems, including clear instructions and examples, especially for AI in generic and industry-agnostic enterprise functions.

4.2.1. Non-Obvious and Obvious Use Cases

In order to address the need for clarity in AI system classification, this study identifies certain cases that fall into multiple risk classes, referred to as Non-Obvious Cases. These cases serve as examples to gain insights into the classification of AI systems across different risk levels. Additionally, Obvious Cases are also provided for each risk level.

4.2.1.1. Non-Obvious Case

The Non-Obvious Cases arise due to the lack of clear division between risk classifications, as discussed in the previous sub-section. Table C.1 presents several examples of non-obvious cases compiled from academic papers. From this compilation, four cases are selected for evaluation to address SQ4. The rational for selecting these use cases is twofold: (1) they are mentioned in academic papers or references, indicating that they could potentially fall under multiple risk classes depending on the context or application domain, or (2) different references classify these AI systems differently.

Based on those indications, four Non-Obvious Cases are selected and summarized in Table 4.1.

Table 4.1: Non-Obvious AI Systems Use-Case and Rationale

Case	Category	Source	Rationale
AI system to measure a truck driver's fatigue and playing a sound to push them to drive longer	Unacceptable Risk	"Manipulative systems that focus on distorting behaviour, such as measuring a truck driver's fatigue and playing a sound that pushes them to drive longer" [39]	It is explicitly mentioned in the URL that this AI system is prohibited due to its intention to manipulate systems to distort behavior. However, for the evaluation in this research, this system is assumed as part-of the Non-Obvious case due to the slight contravening value of 'measuring fatigue' and 'push to drive longer'.
AI systems designed for social robots for children with autism to capture their behavior to assist treatment	High-Risk/Limited Risk	"High-Risk/Limited Risk" as listed in Table 4, p. 7 [18]	From the paper, this system might fall under High-Risk or Limited Risk. However, for the evaluation in this research, with the assumption that this system is intended for good purpose, the system is set to the lowest possible level of this case, Limited Risk.
AI systems for automatic transcription or enhancement of speech	High-Risk/Minimal/ No Risk	"High-Risk/Minimal/No Risk" as listed in Table 4, p. 7 [18]	From the paper, this system might fall under High-Risk or No/Minimal Risk. However, for the evaluation in this research, assuming that this system is not causing significant harm to natural persons, the system is set in minimum risk level as No/Minimal Risk.
AI systems to assess recidivism risk by providing quantitative risk assessments	High-Risk	"One can debate under which of the categories quantitative recidivism risk assessment falls. For instance, it is debatable whether it can be qualified as profiling, which is defined in Article 3(4) of Directive 2016/680" [46]	It is mentioned in the paper that AI systems to assess recidivism risk are considered High-Risk. However, there is the possibility that this case falls under Prohibited. Therefore, this use case is included as a Non-Obvious/borderline case for the evaluation part of this research.

4.2.1.2. Obvious Case

In contrast to the Non-Obvious Cases, the Obvious Cases are those explicitly mentioned in the proposed AI Act. Four examples of Obvious Use Cases are listed in Table 4.2.

First, selected use case for Minimal Risk is the spam filter case. Although Minimal Risk falls outside the regulatory framework of the AI Act, it is mentioned on the European Commission's official website [1], which states that "*The proposal allows the free use of minimal-risk AI. This includes applications such as AI-enabled video games or* ***spam filters***." Additionally, several papers also refer to spam filters as an example of minimal/no risk applications [6] [47].

For High-Risk cases, the use of emotion recognition systems is mentioned in the latest amendment of the AI Act under Article 1(aa) [11], except when applied within areas of law enforcement, border management, workplace, and education institutions.

Regarding Unacceptable Risk, the selected use case is AI systems using remote biometric identification systems. This use case is listed under Prohibited AI Practices in the proposed AI Act, specifically in Article 5d [9].

Lastly, the use case of chatbots is mentioned in the AI Act for some specific AI systems that are obliged to meet minimum transparency obligations. The Explanatory Memorandum of the AI Act proposal explicitly refers to chatbots [9].

Considering that the EU has mentioned these four use cases in its regulations on AI, the categories of the use cases as presented in Table 4.2 are considered ground truth data in this research. This ground truth data will be further utilized to evaluate this research's proposed decision tree framework.

Table 4.2: Obvious AI Systems Use-Case and Rationale

Case	Category	Source	Rationale
AI system to filter unwanted mails and keep them separated from useful emails to reduce time and effort	Minimal/No Risk	"..., minimal-risk AI systems (such as **spam filters**, computer games,..." p.7 [6] "...the fourth category – that of systems that pose a minimal risk (e.g. **spam filters**)" p.104 [47]	Several papers mentioned Spam Filters under Minimal/No Risk category. To make it not too explicit for the evaluation, the 'spam filters' use case is rephrased as presented in the table based on explanation by Khandelwal and Bhargava [20] (paper spam filtering using AI)
AI system use emotion recognition system to identify/recognize patient's emotion	High-Risk	"AI systems intended to be used to make inferences about personal characteristics of natural persons on the basis of biometric or biometrics-based data, including emotion recognition systems, with the exception of those mentioned in Article 5" p. 122, Article 1(aa) - High-Risk AI Systems [11] "the placing on the market, putting into service or use of AI systems to infer emotions of a natural person in the areas of law enforcement, border management, in workplace and education institutions" p. 129, Article 5(dc) - Prohibited AI Practices [11]	The latest amendment of AI Act mentioned emotion recognition systems as High-Risk with the exception of those mentioned in Article 5 (applied within areas of law enforcement, border management, workplace and education institution)
AI systems using remote biometric identification of political protesters creates a significant chilling effect on the exercise of freedom of assembly and association	Unacceptable Risk	"the use of 'real-time' remote biometric identification systems in publicly accessible spaces" p.129, Article 5d - Prohibited AI Practices [9]	The remote biometric identification systems are mentioned as Prohibited/Unacceptable Risk in the proposed AI Act
AI system that automatically converse with people in place for a human being and can interact with them	Limited Risk	"For some specific AI systems, only minimum transparency obligations are proposed, in particular when chatbots or 'deep fakes' are used." p.3 - Transparency for Certain Systems Application [9]	Chatbot use case is mentioned in the Explanatory Memorandum of first draft of the AI Act. To make it not too explicit for the evaluation, the'chatbot' is rephrased as presented in the table based on explanation by Lalwani, et.al. [24]

4.3. Results Discussion

The objective of this chapter is to define the existing classification of AI systems, which will serve as features for the proposed framework (SQ2), and to identify potential challenges in the current classification that will be used as to formulate design principles. However, there are several concerns that arose during the analysis, as outlined below:

1. Concern on No/Minimal Risk identification

While the regulation does not encompass No/Minimal Risk, it is essential to acknowledge that even minimal risk implies some level of potential harm to individuals in the EU. Furthermore, a vast majority of AI systems used within the EU are classified as No/Minimal Risk. Therefore, this research still includes No/Minimal Risk to comprehensively analyze AI experts' understanding of AI system classification.

2. Concern on Limited Risk identification

Another consideration is that the analysis of existing AI systems is based on the AI Act, which is currently under discussion. For this research, the latest amendment from May 2023 was utilized. Several changes have been made in this draft and proposal, such as the inclusion of emotion recognition systems in Article 5 - Prohibited Practices, whereas previously they were only mentioned in High-Risk and Limited Risk practices. Furthermore, there is a growing debate on whether emotion recognition systems should be completely prohibited due to concerns about their scientific validity, potential inaccuracies, biases, and their contribution to discrimination and social inequalities [34]. Thus, before the AI Act is enacted, the existing classification of AI systems for the proposed framework in this research should be adjusted.

The latest amendment of the AI Act in May 2023 only mentions changes to specific articles. Some articles are mentioned in the first draft of the AI Act but not in the amendment, such as Article 52 - Transparency Obligations. Hence, in this research, transparency obligations still refer to the first draft of the AI Act proposal [9]. Consequently, emotion recognition systems in the amendment are mentioned in Prohibited and High-Risk practices and also included as AI systems that have to comply with transparency obligations (Limited Risk).

3. Concern on Ground Truth for further analysis

As non-obvious cases are derived from academic papers and not explicitly mentioned in the AI Act, there is no ground truth available for evaluation. On the other hand, for obvious cases that are explicitly mentioned in the proposed AI Act, the categories of these use cases as specified in the AI Act will be regarded as ground truth in this research.

Consequently, distinct approaches will be employed to evaluate the decision tree framework's performance for both obvious and non-obvious cases.

4. Concern on the Obvious and Non-Obvious Cases

To address concerns related to clarity, this study identifies Non-Obvious Cases that fall within the borderline between risk classes. These cases, along with the Obvious Cases, are summarized in Table 3.2 to assess the effectiveness of the AI Act framework in classifying AI system use cases into specific risk levels. Furthermore, these use cases contribute to answering SQ4.

The selection of use cases is based on academic papers or the AI Act itself However, this selection may present issues as the chosen use cases are formulated in concise sentences based on the AI Act or references, with minimal additional contextual information. This decision was made deliberately to evaluate the proposed framework. Altering the provided sentences or adding more context could potentially lead to different classifications, causing them to be incomparable to existing sources. Ultimately, with the 'minimal' context of the use case, we can also gain more insights into what context is needed to make AI systems can be classified better.

In some cases, additional explanations were provided to prevent them from being easily guessed. For instance, the use case of a chatbot/spam filter was reformulated based on its translation according to a specific academic paper. Table 4.2 includes some rationales for such cases.

The non-obvious cases were selected based on their potential to lead to misclassification due to ambiguous descriptions. For example, the case of 'AI systems designed for social robots for children with autism to capture their behavior to assist treatment' is intriguing, as it encompasses elements such as 'social robots,' 'children,' 'autism,' 'behavior capture,' and 'treatment assistance,' all of which could potentially lead to classification as Unacceptable Risk, High-Risk, or Limited Risk.

Lastly, certain decisions were made based on assumptions, as mentioned above, for both obvious and non-obvious cases. These assumptions may impact the evaluation of the proposed framework. To minimize the potential limitations associated with these assumptions, discussions were held with legal experts to gain a deeper understanding of the AI Act, its challenges, and the selection of obvious and non-obvious cases.

4.4. Summary of the Chapter

In summary, AI systems governed by the AI Act are categorized based on risk levels, with each risk classification accompanied by specific obligations. Unacceptable risk AI systems are strictly prohibited for use within the EU region, with certain exceptions. High-risk AI systems require compliance with a set of requirements before they can be introduced to the EU market. Certain AI systems necessitate transparency obligations if they intend to interact with or impersonate natural persons. AI systems that do not fall under these three categories are classified as No/Minimal risk where a voluntary code of conduct similar to the requirements for High-Risk systems are recommended.

While this risk classification framework provides benefits in protecting natural persons, challenges exist within the current classification of AI systems. One such challenge is the lack of clarity in classifying certain AI system use cases. Without clear boundaries between risk levels, organizations may encounter issues related to budget allocation and innovation, and could potentially face penalties for misclassification under the AI Act. Several issues contributing to the lack of clarity have been discussed in this chapter, including the narrow scope of Unacceptable Risk, exceptions and loopholes in prohibited AI practices, vague wording in certain cases, the distinction between high-risk and limited-risk categories, reliance on interpretation for risk classification, the need for a more interdisciplinary approach, and the dilemmas and paradoxes arising from dichotomous thinking. These discussions serve as considerations for addressing the identification of Design Principles in response to SQ2.

To address the clarity issues, this study identifies Non-Obvious Cases that fall in the borderline between risk classes. These cases, along with the Obvious Cases, are utilized to assess the effectiveness of the AI Act framework in classifying AI system use cases into specific risk levels. Furthermore, these use cases will also contribute to answering SQ4.

5

Differentiating Features of Each Risk Classification

This chapter addresses the second sub-question of the research: *What are criteria that differentiate each level of the AI Act classification?*. Referring to Design Science Methodology, this stage defines the solution's requirements. The research approach in this chapter includes desk research and interviews with legal experts. The desk research is done by understanding the abstraction of existing AI systems classification (SQ1) in the AI Act by conducting thematic analysis and analyzing which criteria distinguish one risk class from others. Insight from the discussion with legal experts is also taken into consideration. The process to extract the distinctive criteria from the existing AI systems classification is done in Miro. In addition, design principles are derived from the challenges identified in the research and discussions with legal experts to guide the development of the proposed framework.

Section 5.1 presents the selected distinctive criteria for each risk class using thematic analysis, while Section 5.2 proposes the design principles. A brief discussion of the Section 5 analysis is provided in Section 5.3.

5.1. Thematic Analysis of the AI Act

From the thematic analysis, all mentioning theme (features) of the AI Act content are visualized in Table D.1, Table D.2, Table D.3, and Table D.4 in the Appendix D. These tables summarize the key themes extracted from the AI Act content.

Based on the analysis of the AI Act, thirteen themes can be identified across all risk levels. These themes are as follows:

1. Protected Value

The AI Act explicitly aims to safeguard Fundamental Rights and Union values. Unacceptable Risks, High-Risk AI systems, and Limited Risks all consider the protection of specific values. For example, AI Systems with unacceptable risk use practices that contradict EU values and fundamental rights, while High-Risk AI systems also take into account the adverse impact on health, safety, fundamental rights, and the environment [5]. Limited Risk is regulated to ensure trustworthiness from an individual perspective. The understanding of the distinction between risk levels requires a comprehensive grasp of the 'Protected Values'.

2. Objective/Intention

The intended purpose of AI systems are mentioned in every risk level. The way how the AI system is classified is based on its intended purpose. For instance, the AI systems which intentionally exploit vulnerabilities of specific groups are classified as Unacceptable Risk. As set out in the Explanatory Memorandum of the AI Act, the classification 'high-risk' is based on the intended purpose of the AI system, in line with existing EU product safety legislation. The classification of AI systems depends not only on their function but also on their specific purpose and modalities of use [45]. The same thing goes to Limited Risk while in the Article 52, AI systems with intention to interact with human or generate/manipulate content are subject to transparency obligations.

3. Affect/Impact

The 'Impact' theme focuses on how AI systems affect humans. For instance, AI systems that materially distort

human behavior can have physical or psychological effects, or cause significant harm to specific groups or individuals. If an AI systems cause physical or psychological effect to individuals, then they are categorized as Unacceptable Risk.

4. Technology/Use-Case

While similar to the 'Objective/Intention' theme, the 'Technology/Use-Case' theme emphasizes specific technologies or use-cases mentioned in the AI Act. Biometric identification systems are mentioned across all risk levels. It is important to note that certain use-cases for biometric identification systems are classified as Unacceptable Risks, while others may be categorized as High-Risk, depending on the context and domain. For example, emotion recognition systems is considered as Unacceptable Risk if they are used within law enforcement, border management, workplace and education institutions domain. If the emotion recognition systems are applied within the domain mentioned in Annex III but outside those for prohibited cases, then these AI systems will be classified as High-Risk.

The biometric identifiction systems belong to prohibited risk if they are operated in 'remote', 'in real time', 'in publicly accessible places', and 'for the purpose of law enforcement' [34].

5. Relatedness to Prohibited Practices in the EU

This theme is specifically mentioned for Unacceptable Risk. For prohibited practices, the AI Act also mentioned how it is prohibited, such as social control practices (social scoring) that is prohibited since the use of social scoring itself is not allowed in the EU. In addition, if AI systems materially distort human behavior, it is also prohibited because of violating Fundamental Rights and/or Union values. Furthermore, AI systems that deploy subliminal components or AI systems that exploit vulnerabilities of individuals or specific groups are violating Fundamental Rights and/or Union values, such as human dignity, which then makes them prohibited to use in the EU.

6. Risk or Potential Risk

Considering the risk-based approach of the regulatory framework, this theme is mentioned across all risk levels. It provides a more detailed assessment of the risks associated with AI systems, particularly in terms of potential risks to human dignity, discrimination, and other factors.

7. Benefit

The AI Act weighs the risks posed by High-Risk AI systems against their benefits, taking into account factors such as human dignity, privacy, data protection, freedom of expression and information, and non-discrimination. If an AI system is not prohibited, its benefits are considered along with potential risks to determine its risk classification.

8. Intertwined Regulation

The AI Act harmonizes with existing legislation, and each risk level specifies the intertwined regulations that must be complied with. For example, the use of biometric identification systems is mentioned in the context of criminal offenses and sectorial legislation for product safety. Compliance with these regulations determines the risk classification of AI systems. If an AI system used as safety of a product or a product that already regulated prior the AI Act proposal, and included in one of the Harmonised Legislation in Annex II, then this AI systems is classified as High-Risk.

According to the explanatory memorandum of the AI Act draft, the high-risk classification is depend on the purpose of the sysstem in connection with existing product safety legislation [6].

9. Input Data

The type of data used in AI systems is mentioned for Unacceptable Risks, with a focus on biometric data or biometric-based data. Other types of data, such as social behavior and sensitive information like gender identity, race, ethnic origin, migration status, and political orientation, are also referenced in the AI Act.

10. Exemption

For some AI system use-cases, several exemptions are also mentioned. All AI systems for military purposes should be excluded from this regulation, as stated in the AI Act. Some Prohibited AI system are allowed to be

used for research purposes, such as AI systems with subliminal message. If AI systems is designed to be used solely for cybersecurity purposes, then they are also excluded from the High-Risk group.

11. Type (of AI system)

This theme is extracted from High-Risk AI systems as outlined in Article 5 AI Act. There are two main categories of High-Risk AI systems, those are: (1) AI systems intended to be used as a safety component of products that are subject to third party ex-ante conformity assessment and (2) Standalone AI systems with mainly fundamental rights implications. This includes AI systems that are product or safety components (Article 6(1)) or systems used in the areas listed in Annex III of the draft AI Act (Article 6.2), including such areas as biometric identification and categorisation, education, employment, law enforcement, migration, asylum and border control [5].

12. Domain

If an AI system belongs to the domains listed in Annex III, the AI Act categorizes them as High-Risk. The existing classification is limited to eight domains, such as biometric identification, education, employment, law enforcement, migration, asylum, and border control [32].

13. Requirement

The AI Act aims to regulate the use of AI systems within the EU to ensure compliance with Fundamental Rights and Union values. Each risk level has specific requirements that AI systems must comply with. Unacceptable Risks are strictly prohibited, High-Risk AI systems are subject to conformity assessment and mandatory requirements based on sectorial legislation or predefined domains, and Limited Risks have limited transparency rules. Minimal/no risk AI systems are encouraged to follow a voluntary code of conduct for assurance, despite being outside the scope of the regulation [5].

Based on these 13 themes obtained from the thematic analysis, four themes are selected as differentiating features for each risk class. They are **Protected Values, Objective/Intended Purpose, Domain,** and **Technology/Use Case**. The selection is based on understanding how central each theme is to distinguish the risk class, such as the Protected Values. Protected Values can be used to analyze whether AI systems are potentially causing significant harm to the AI systems, which makes them an Unacceptable Risk/High-Risk or has less/no significant harm to the individuals(Limited Risk/Minimal Risk). In the subsequent chapter, these features will be highly considered in generating the proposed framework. Although crucial for the classification process, the' Requirement' theme is not included as a feature since it is the output of the classification process.

5.2. Design Principles

This section presents a set of design principles that can guide the development of an effective AI system classification framework. These principles address the identified challenges related to clarity, distinction between high-risk and limited risk, reliance on interpretation for risk assessment, and dilemmas arising from dualistic thinking. As mentioned by [4], standardization and clear guidance are necessary for AI system classification.

The following six design principles are proposed as guidance for creating the classification framework:

1. Clarity

The principle of clarity emphasizes the need to make the AI system classification framework easily understandable to users and stakeholders. The clarity also refers to the readability of the content represented in the framework [2]. It involves presenting information in a clearly and unambiguously manner, minimizing ambiguity and confusion. By designing for clarity, the framework enhances transparency, reduces the risk of misinterpretation, and enables stakeholders to make informed decisions based on the system's categorization.

2. Simplicity

The principle of simplicity emphasizes the importance of keeping the AI system's classification framework straightforward and uncomplicated. It involves avoiding unnecessary complexity, intricate terminology, and convoluted structures. Simplifying the design promotes user-friendliness and ease of understanding. A simple classification framework allows users to grasp the categorization criteria easily, reducing cognitive load

and potential errors. By simplifying the design, the framework facilitates broader adoption, as users from various backgrounds can comprehend and apply the classification system more effectively. Failure to incorporate simplicity into the framework's design can reduce its utility [2].

3. Obligation-Required

Obligation-required design principle ensures that the AI system's classification framework aligns with the obligations and requirements specified by the AI Act or relevant regulations. It encompasses incorporating legal and regulatory obligations into the design. Adhering to obligation-required design principles ensures compliance with legal and regulatory frameworks. By explicitly considering and reflecting the obligations mandated by the AI Act, the classification framework becomes more robust, reliable, and legally sound.

4. Sequential

The sequential design principle emphasizes organizing the AI system's classification framework in a logical and step-by-step manner, enabling a structured decision-making process. It involves establishing a sequence of criteria or questions to guide the classification process. Sequential design promotes consistency and uniformity in classifying AI systems. It helps users navigate through the classification framework systematically, ensuring comprehensive coverage and minimizing subjective interpretations. By following a predefined sequence, it brings clarity and reduces ambiguity, facilitating more accurate and reliable categorization outcomes.

5. Representative

Representative design principle involves making the AI system's classification framework reflective of real-world scenarios, emerging technologies, and societal considerations. It necessitates periodic reviews and updates to ensure the framework remains relevant and adaptable. This principle acknowledges the dynamic nature of AI systems and the evolving landscape of technology and society. By regularly evaluating and updating the classification framework, it can accommodate new AI developments, address emerging risks, and align with evolving legal and ethical standards. Being representative enables the framework to remain robust, up-to-date, and responsive to changing circumstances.

6. Value-First

As the objective of the AI Act proposal is to protect Fundamental Rights and Union values, the value consideration should be prioritized. By adopting a value-first approach, the design ensures that the classification framework goes beyond mere technical considerations. It takes into account the value consideration of each AI practices.

By adhering to these design principles, an effective and well-rounded AI system classification framework can be developed. Such a framework enhances usability, compliance, accuracy, and alignment with the AI Act.

5.3. Results Discussion

This section presents the results obtained from the analysis, which aimed to identify the distinguishing features of each risk level in the AI Act classification through thematic analysis. Furthermore, design principles corresponding to these findings are formulated. These outcomes address the second sub-question of the research, which focuses on understanding the characteristics that differentiate AI system risk levels in the AI Act classification. The insights gained from this analysis will be used to develop the proposed framework in SQ3. There are three things to consider from the results of this analysis.

1. Iteration with Legal Experts

First, it is important to acknowledge that the selection of differentiating features and the formulation of design principles may be subject to author interpretation. This potential bias can lead to misclassification of AI systems. To mitigate this risk, complementary references were consulted, and an iterative process was employed to extract the thematic analysis. Additionally, discussions with experts were conducted to ensure a comprehensive understanding of the findings.

The iterative process was done twice to understand the AI Act classification better. The first iterative process was conducted individually to code and group based on the AI Act (as seen in Appendix D). To eliminate the author's bias and strengthen the understanding of the legal terms, the second iteration was conducted with

two legal experts by asking several questions in a semi-structured interview to understand the context and legal terms. Such question is *'What are the key features of each class?'*. This question was proposed to acknowledge what differentiates each risk class. After the semi-structured interview was done, the abstraction process was then continued.

2. On-going discussion of the AI Act

Second, these selections were made prior to the public release and implementation of the AI Act in the EU. Therefore, adjustments may be necessary once the AI Act is available and in effect.

3. Concerns of oversimplification or generalization

Finally, the design principle of simplicity may raise concerns about oversimplification or generalization of the legal context, which typically allows for open interpretation. However, considering the challenge of clarity in the AI Act, simplicity is incorporated in a manner that facilitates the usability of the proposed decision tree framework.

5.4. Summary of the Chapter

Through the analysis of the output from SQ1, the existing AI systems classification in the AI Act, the identified challenges, and the abstraction process along with discussions with legal experts, four distinctive features have been identified as variables for the decision tree used in classifying each risk level. These features include **Protected Values**, **Objective/Intended Purpose**, **Domain**, and **Technology/Use Case**.

Protected values are a primary concern across all risk levels. Unacceptable Risk considers the objective/intended purpose of AI systems, while High-Risk also considers the domain of application and its significant impact on safety, humans, and the environment as outlined in Annex II or Annex III. Limited Risk focuses on protected values and the technology/use-case of AI systems. AI systems that do not fall into these three risk levels are categorized as minimal/no risk. Design principles are derived to aid the process of generating the proposed framework based on the challenges identified and the AI Act itself.

Furthermore, in addition to the distinctive features, high-level design principles have been formulated based on the findings of SQ1. These design principles include **clarity**, **simplicity**, **obligation-required**, **sequential**, **representative**, and **value first**. These principles serve as the boundaries for generating the decision tree used in the classification of AI systems.

6

Proposed Framework

This chapter addresses the third sub-question, "What possible framework can be designed to improve the classification process of AI systems?". This sub-question is associated with the Design Science Methodology's 'design and develop artifact' stage. The possible framework as the artifact for the classification process is determined, and after careful consideration, the decision tree framework is selected. The decision tree is created through a creative idea-generation process in Miro, considering the features and design principles from SQ2. An iterative approach is employed to refine the decision tree, ensuring its compliance with design principles.

The generated decision tree framework is presented in Section 6.1. Furthermore, the analysis of the decision tree according to the designated design principles is elaborated in Section 6.2.

6.1. Decision Tree Framework

The decision tree framework is chosen as the proposed classification framework for AI systems under the AI Act. Decision trees offer a clear and structured approach to decision-making, characterized by if-then rules in a flowchart type [40][29]. This aligns with the design principle of sequential organization, as the decision tree guides users through step-by-step classification by presenting sets of questions.

Moreover, it is also simple to understand and easy to implement [29]. These advantages correspond to the clarity and simplicity of design principles formulated to develop the proposed framework.

The decision tree is constructed based on the features and design principles derived from the SQ2 output. Figure 6.1 provides an overview of the decision tree, illustrating the sequential assessment. A total of 20 questions are formulated, organized according to the pre-selected aspects: Protected Value, Objective/Intention, Domain, and Use-Case/Technology (refer to Table 6.1). Further elaboration on these aspects will be presented in the subsequent sections.

In order to facilitate a comprehensive understanding, the decision tree is complemented with additional definitions and terms. The results of the risk classification include the identified risk class, the corresponding risk value, necessary requirements, and any applicable exemptions.

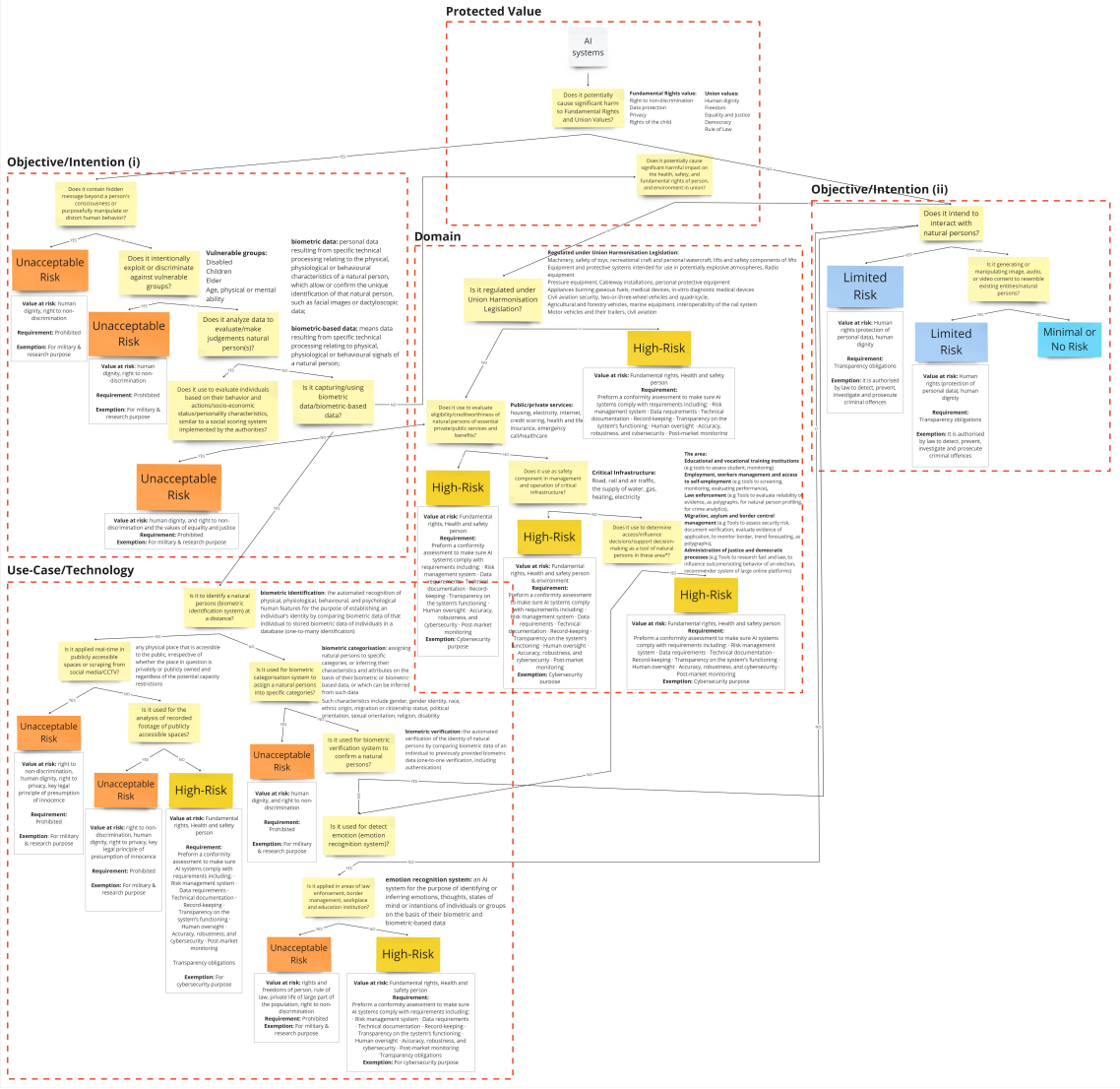

Figure 6.1: Decision Tree Overview

6.1.1. Protected Value - Decision Tree

The classification of AI systems under the AI Act aims to safeguard Fundamental Rights and Union Values. Consequently, the first question addresses the Protected Value aspect. The initial step involves excluding practices that are prohibited (Figure 6.2. Hence, the first question to ask is, "Is it potentially caused significant harm to Fundamental Rights and Union Values?". Depending on the answer, the assessment proceeds to either Unacceptable Risk (if the answer is 'Yes') or Limited Risk (if the answer is 'No'). Notably, High-Risk AI systems are considered to have the potential for significant harm to Fundamental Rights and Union Values, and thus, they are evaluated sequentially after the Unacceptable Risk assessment. This first question divide the AI systems into those posing significant risks and those with less risk to Fundamental Rights and Union Values. To add more information, several Fundamental Rights and Union Values are mentioned in the tree.

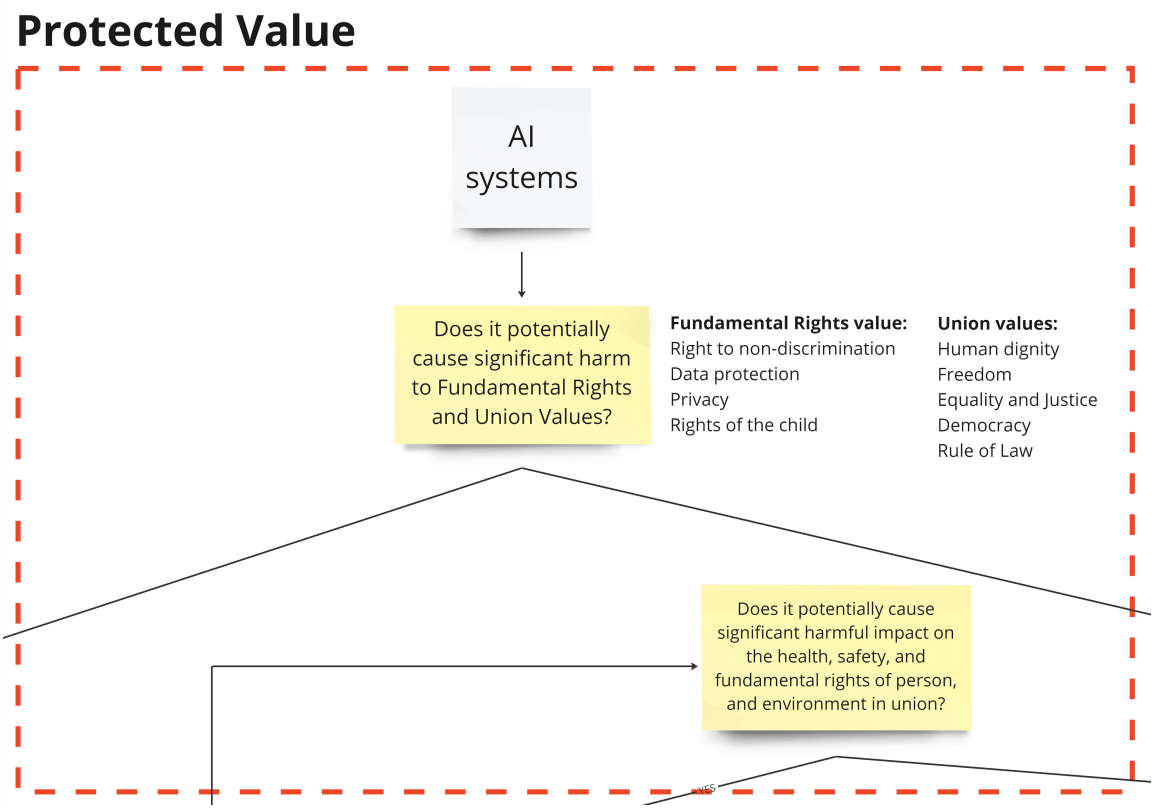

Figure 6.2: Protected Value - Decision Tree

Another relevant question pertaining to Protected Value is whether the AI system could potentially cause significant harm to the health, safety, fundamental rights of individuals, or the environment within the Union. This question categorizes AI systems into High-Risk or non-High-Risk, with the latter further categorized as Limited Risk or No/Minimal Risk.

6.1.2. Objective/Intention (i) - Decision Tree

The Objective/Intention features are divided into two types: (i) Objective/Intention to assess Unacceptable Risk and (ii) Objective/Intention to assess Limited Risk. Each type of feature corresponds to a specific classification output. This section presents the Objective/Intention (i) part (Figure 6.3), which consists of questions based on Article 3 of Prohibited Practices in AI systems. These questions are asked sequentially to assess Prohibited Practices.

The first question focuses on hidden messages or intentional manipulation of human behavior beyond conscious awareness, as outlined in Article 5.1.a of the AI Act. If the answer is 'Yes,' the AI system falls under Unacceptable Risk; otherwise, it proceeds to the next question.

The next question aims to identify whether the AI system is intentionally designed to exploit or discriminate against vulnerable groups such as the elderly or disabled community. This practice is prohibited under Article 5.1.b, leading to a classification of Unacceptable Risk. If the answer is 'No,' the assessment moves to the next question.

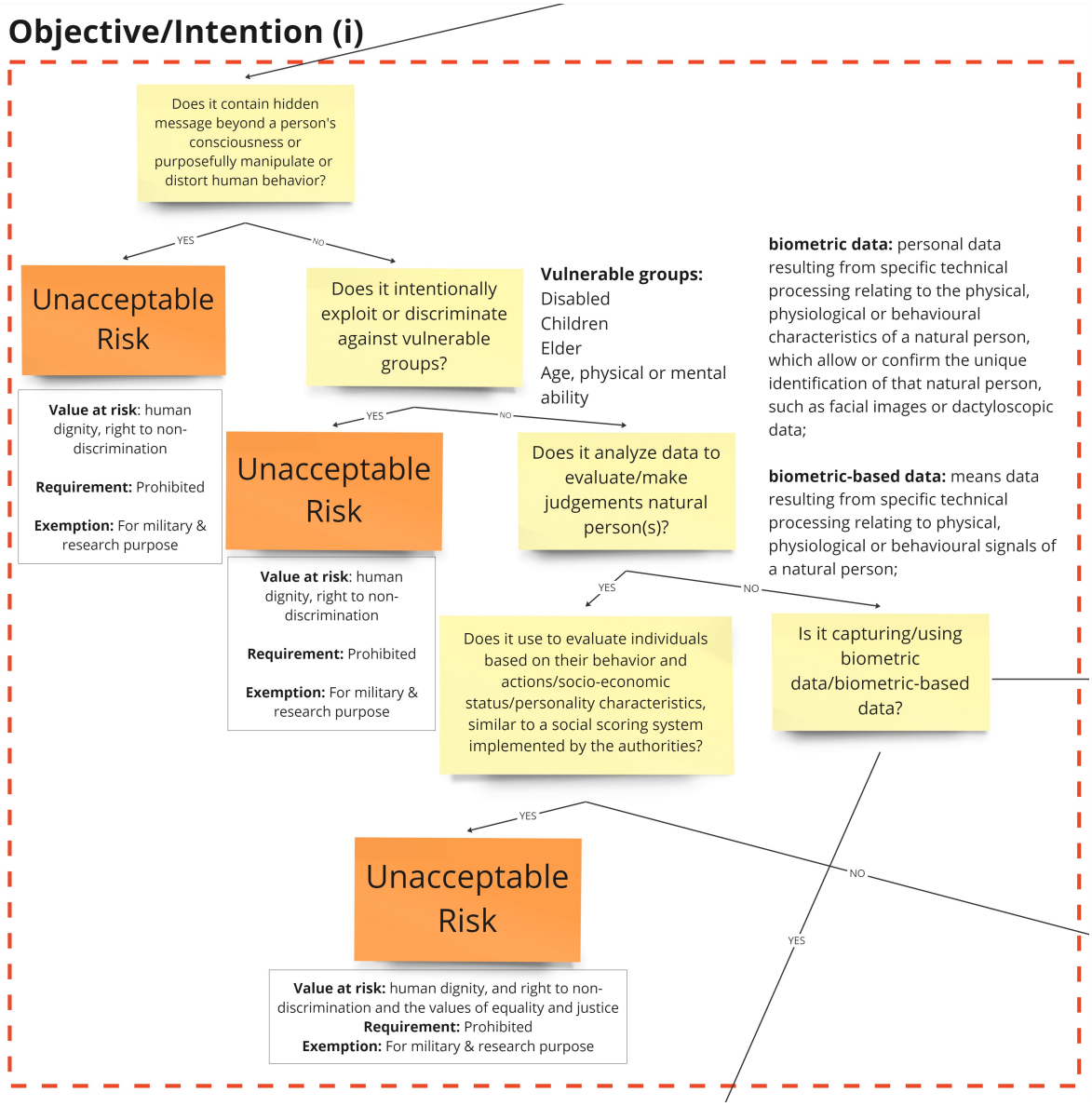

Figure 6.3: Objective/Intention (i) - Decision Tree

Subsequent questions assess specific AI systems under prohibited practices. For example, the question "Does it analyze data to evaluate or make judgments about natural persons?" examines the use of AI systems like Credit Scoring or Social Scoring. The answer to this question determines whether the system falls under Unacceptable Risk or requires further evaluation for social scoring purposes. The assessment continues with additional questions based on the specific context of the AI system.

Another question focuses on the use of AI systems to evaluate individuals based on behavior, actions, socio-economic status, or personality characteristics, similar to a social scoring system implemented by authorities. This use is classified as Unacceptable Risk under Article 5.1.c. However, if the answer is 'No,' the assessment proceeds to determine whether the evaluation is intended for assessing eligibility or creditworthiness of individuals for essential private or public services, which falls under the High-Risk category.

The last question derived from the Objective/Intention (i) feature is whether the AI system captures or uses biometric data. If the answer is 'No,' the assessment moves to evaluate the system as High-Risk. However, if the answer is 'Yes,' more detailed questions regarding the use of biometric identification systems are asked to assess the system's classification, considering exceptions and contexts. This question selection is based on the understanding that biometric systems are mentioned in the AI Act and can be classified as Unacceptable

Risk, High-Risk, or Limited Risk.

6.1.3. Use-Case/Technology - Decision Tree

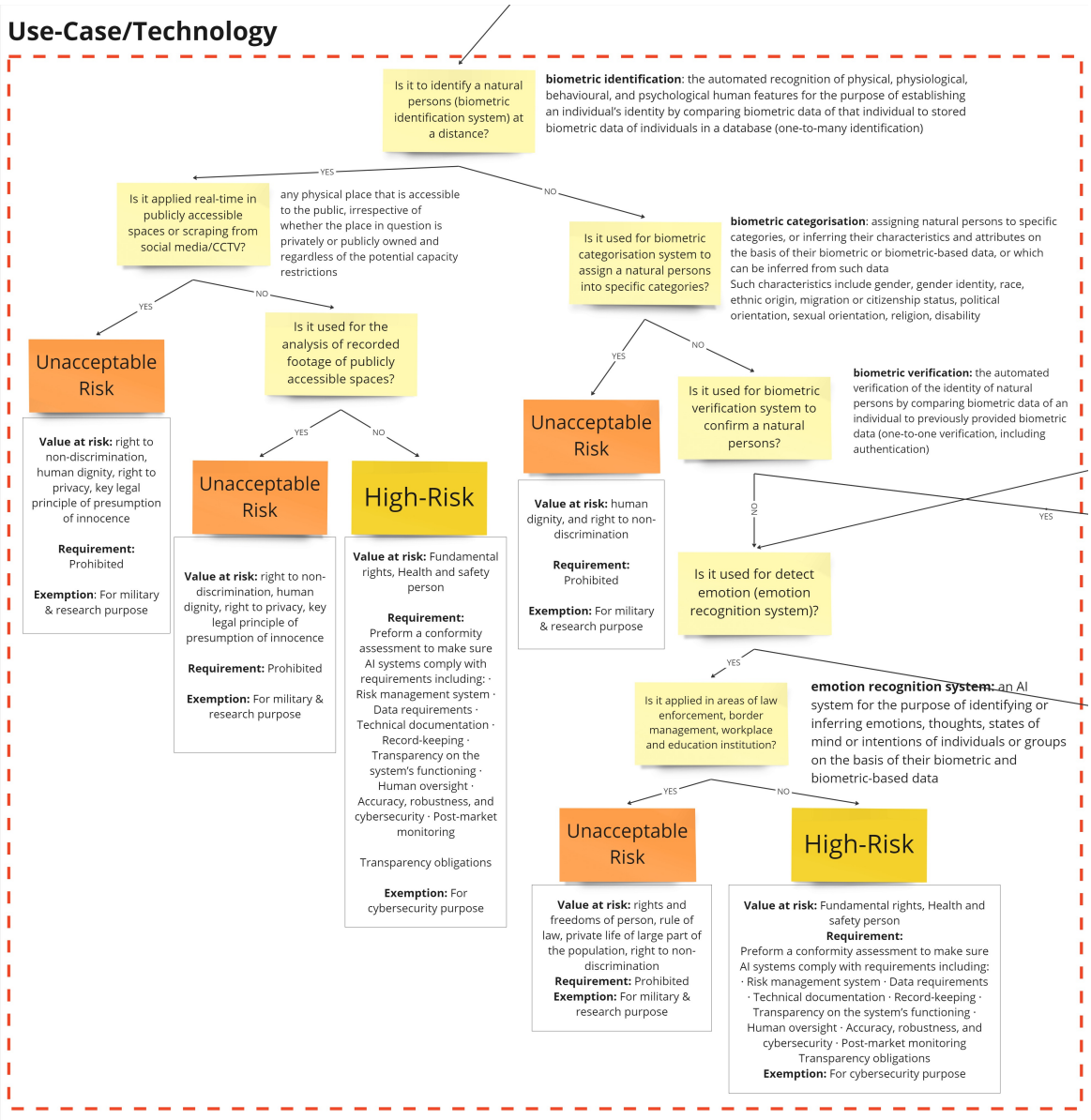

Figure 6.4: Use-Case/Technology - Decision Tree

The AI Act specifically mentions biometric identification systems in Article 5, Annex III, and Article 52. Therefore, use-cases involving biometric identification systems are grouped to assess whether the AI systems are classified as Unacceptable Risk, High-Risk, or Limited Risk. Prohibition keywords for these systems include operation at a distance, real-time usage (including short delays) in publicly accessible places, and law enforcement purposes [34]. The questions in 'Use-case/Technology' feature are depicted in Figure 6.4.

Another grouping focuses on the use of biometric identification systems at a distance. If the system is used at a distance and within the domain of prohibited practices, it is likely to be classified as Unacceptable Risk. To ensure the classification accuracy, more questions are asked, such as: 'Is it applied real-time in publicly accessible spaces or scraping from social media/CCTV?' from Article 5.1.db and 'Is it used for the analysis of recorded footage of publicly accessible spaces?' that mentioned in Article 5.1.e and Annex III.1.

Suppose the AI system is not used at a distance. In that case, the assessment moves to determine whether

it is used for biometric categorization (Article 5.1.ba), biometric verification (Annex III.1), or emotion recognition (Article 1.1.dc and Article 6.1.aa). The use of biometric categorization is prohibited, while biometric verification, as mentioned in Article 3.1 that such biometric verification whose sole purpose is to confirm that an individual is the person he or she claims to be. If the AI systems are intended to be used for biometric verification, as mentioned, it is then classified as Limited Risk.

For emotion recognition systems, which are mentioned in both Prohibited Practices (Article 5) and High-Risk AI systems (Annex III), more detailed questions are asked. If the system is applied in areas such as law enforcement, border management, workplace, or educational institutions, it falls under prohibited practices (Article 5). Otherwise, if it is not used within those domains, it is classified as High-Risk (Annex III, Article 6,1.aa). It's important to note that emotion recognition systems are also mentioned in Transparency Obligations (Article 52), requiring compliance with transparency rules.

6.1.4. Domain - Decision Tree

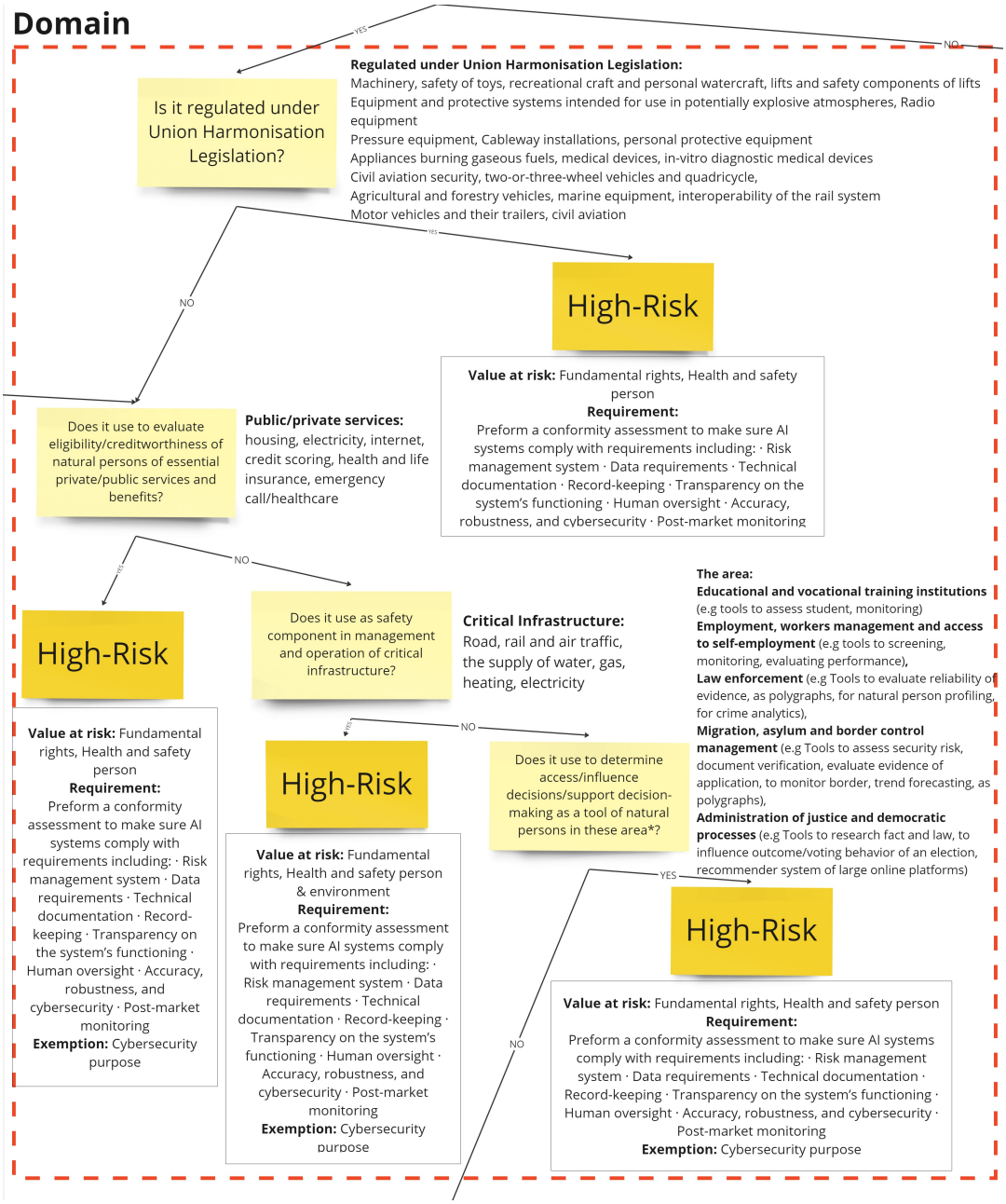

Figure 6.5: Domain - Decision Tree

The previous sub-sections mainly assess the Unacceptable Risk of AI systems. According to the AI Act, the High-Risk assessment begins by excluding prohibited practices. Then, High-Risk classification is determined by checking whether certain AI systems are regulated under sectoral legislation prior to the AI Act proposal or if they are standalone products used in critical areas or use-cases mentioned in Annex III. The questions within the Domain feature are derived from Article 6 and Annex III (Figure 6.5).

As explained in Chapter 4, there are two main categories of High-Risk AI systems: (1) AI systems intended to be used as safety components of products subject to third-party ex-ante conformity assessment and (2) standalone AI systems. For the first category, if the systems are regulated under other sectoral legislation, they are classified as High-Risk AI systems. If not, further assessment is required to determine if they are mentioned in critical domains listed in Annex III. Therefore, the first question in the 'Domain' feature is whether the system is regulated under Union Harmonization Legislation.

The next question aims to evaluate whether the AI system is used to evaluate the eligibility or creditworthiness of natural persons for essential private or public services and benefits. This system is classified as High-Risk under Annex III.5.

Another question is to evaluate wether the AI system is a safety component in management and operation of critical infrastructure such as road, transport, energy, and others. This particular systems are classified as High-Risk according to Annex III.2.

The last question derived from Annex III.3/4/6/7 evaluates whether the High-Risk AI system is used to make access/influence decisions or support decision-making in areas such as education, workplace institutions, law enforcement, migration and border control management, and administration of justice and democratic processes. If an AI system falls within these categories and domains, it is classified as High-Risk AI systems.

The use-case of biometric identification systems is already addressed in the Use-Case/Technology feature in the previous section, as it encompasses all technologies related to biometric/biometric-based systems.

6.1.5. Objective/Intention (ii) - Decision Tree

This section focuses on the Objective/Intention of an AI system, relating to the Objective/Intention (i) feature. The classification in this section corresponds to Limited Risk (Figure 6.3). The questions are based on Article 52, which covers Transparency Obligations.

The first question determines whether the AI system intends to interact with natural persons, as stated in Article 52.1. The next question is related to deepfake technology, asking whether the system generates or manipulates image, audio, or video content to resemble existing entities or natural persons, as mentioned in Article 52.3. If the AI system falls under either of these questions, it is classified as Limited Risk and must comply with Limited Risk requirements. However, if the AI system does not fall under these questions, it is classified as Minimal/No Risk and can voluntarily adhere to a code of conduct (Article 69).

There is one more use-case for Limited Risk, which is the use of emotion recognition systems mentioned in Article 52.2. However, this use-case is already grouped within the assessment of biometric identification systems in the previous section and is not mentioned again here. It should be noted that even though emotion recognition systems are classified as High-Risk, they still need to comply with transparency rules according to the output of the classification.

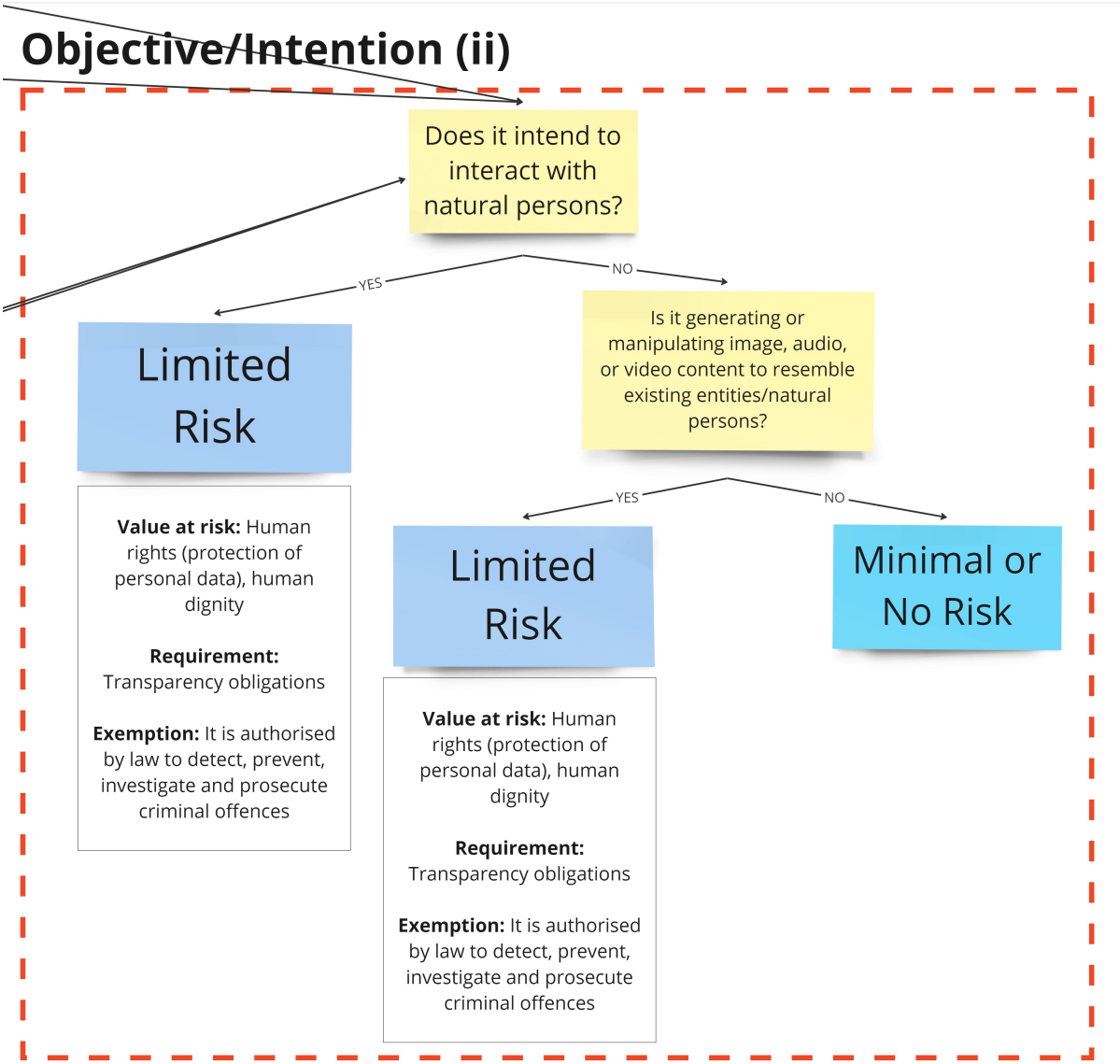

Figure 6.6: Objective/Intention (ii) - Decision Tree

6.2. Desicion Tree Analysis

This section aims to address the sub-question "What possible framework can be designed to enhance the classification process of AI systems?" Taking into account the identified challenges and design principles (SQ2), a decision tree is proposed as a framework to facilitate the classification process of AI systems under the AI Act.

1. Decision Tree Selection

The decision tree is selected based on several reasons. Firstly, it offers interpretability, making it easier to understand and explain the classification process. Additionally, decision trees are known for their simplicity and ease of implementation. Although decision trees have limitations, such as the potential for oversimplification or generalization, these disadvantages are carefully weighed against the advantages. In this research, the decision tree is chosen after undergoing an iterative process and considering input from legal experts.

2. The Role of Iterative Process in Decision Tree Development

The iterative process plays a crucial role in the development of the decision tree. The process of developing the decision tree was iterated three times. The initial iteration was conducted by the author on an individual

Table 6.1: List of Decision Tree Questions

Code	Question	Option	Result	Features
Q1	Does it potentially cause significant harm to Fundamental Rights and/or Union values?	Yes	Q2	Protected Values
		No	Q19	
Q2	Does it contain hidden message beyond a person's consciousness or purposefully manipulate or distort human behavior?	Yes	UR	Objective/ Intention (i)
		No	Q3	
Q3	Does it intentionally exploit or discriminate against vulnerable groups?	Yes	UR	
		No	Q4	
Q4	Does it analyze data to evaluate/make judgements natural person(s)?	Yes	Q5	
		No	Q6	
Q5	Does it use to evaluate individuals based on their behavior and actions/socio-economic status/personality characteristics, similar to a social scoring system implemented by the authorities?	Yes	UR	
		No	Q16	
Q6	Is it capturing/using biometric data/biometric-based data?	Yes	Q7	Use-case/ Technology
		No	Q14	
Q7	Is it to identify a natural persons (biometric identification system) at a distance?	Yes	Q8	
		No	Q10	
Q8	Is it applied real-time in publicly accessible spaces or scraping from social media/CCTV?	Yes	UR	
		No	Q9	
Q9	Is it used for the analysis of recorded footage of publicly accessible spaces?	Yes	UR	
		No	HR	
Q10	Is it used for biometric categorisation system to assign a natural persons into specific categories?	Yes	UR	
		No	Q11	
Q11	Is it used for biometric verification system to confirm a natural persons?	Yes	Q19	
		No	Q12	
Q12	Is it used for detect emotion (emotion recognition system)?	Yes	Q13	
		No	Q19	
Q13	Is it applied in areas of law enforcement, border management, workplace and education institution?	Yes	UR	
		No	HR	
Q14	Does it potentially cause significant harmful impact on the health, safety, and fundamental rights of person, and environment in union?	Yes	Q15	Protected Values
		No	Q19	
Q15	Is it regulated under Union Harmonisation Legislation?	Yes	HR	Domain
		No	Q16	
Q16	Does it use to evaluate eligibility/creditworthiness of natural persons of essential private/public services and benefits?	Yes	HR	
		No	Q17	
Q17	Does it use as safety component in management and operation of critical infrastructure?	Yes	HR	
		No	Q18	
Q18	Does it use to determine access/influence decisions/ support decision-making as a tool of natural persons in these area*?	Yes	HR	
		No	Q12	
Q19	Does it intend to interact with natural persons?	Yes	LR	Objective/ Intention (ii)
		No	Q20	
Q20	Is it generating or manipulating image, audio, or video content to resemble existing entities/natural persons?	Yes	LR	
		No	NR	

basis. The second iteration of the decision tree was conducted subsequent to conducting interviews with two legal experts in order to solicit their feedback on the initial iteration. Finally, following the incorporation of comments provided by legal experts, the decision tree underwent simulation by another two experts in order to obtain their feedback prior to the experiment.

Initially, the decision tree consisted of more than 20 questions, including inquiries about the exemptions associated with each use case. However, through iterative cycles and pre-tests, it became evident that having over 20 questions, especially for the longest path, was excessively long and discouraged respondents from engaging with the tree. Consequently, after the third iteration, the number of questions was condensed to a maximum of 20 for the longest possible path. As a trade-off, the "Exemptions" component for each risk level was excluded. To ensure that respondents remained aware of the potential exemptions applicable to their AI systems, the exemption information was provided as additional information once the decision tree result was obtained, along with the "Value at Risk" and "Requirements" details.

Another thing that came up during the second round of improvements was that some terms in the decision tree needed to be explained. Terms like "What are the Fundamental Rights/Union values?" or the definition of Vulnerable Groups, or explanations of biometric data and biometric-based data, needed more context. Interestingly, these terms were left out in the first version. However, the legal experts pointed out that explaining them well is important. If not, especially for respondents who are not legal experts, it would be tough to grasp the whole picture and classify the AI systems correctly.

Moreover, iteration was important in ensuring that each question posed to respondents was comprehensible. The questions were carefully refined to maintain consistency with the language and context of the AI Act. Although it is still possible that respondents may find certain questions vague or unclear, the iterative process aimed to enhance the readability and understandability of the decision tree.

For further improvement, the list of questions for the decision tree can be comprehensively analyzed one by one by AI experts, either from legal experts or from a more technical view, to ensure the questions' comprehensiveness.

3. Decision Tree Analysis correspond to the Design Principle

In Section 6.1, design of decision tree is elaborated. In alignment with the iterative process, the decision tree underwent analysis following the Design Principles during each iteration. This analysis was conducted individually, drawing insights from expert feedback obtained during the iteration. For instance, when legal experts expressed concerns about the decision tree's excessive length impeding their comprehension of the questions, the author inferred that the 'simplicity' principle was not adequately addressed. Subsequently, improvements were implemented to align with the 'simplicity' principle. Similarly, instances arose where legal experts noted the lack of clarity in certain questions. In response, the author deduced that the decision tree failed to achieve the 'clarity' principle. Consequently, additional information was incorporated into the decision tree framework to enhance the fulfillment of the 'clarity' principle.

To ensure the eligibility of this decision tree for use in the next stage of the research (Evaluation stage), it is evaluated based on the following design principles: clarity, simplicity, obligation-required, sequential, representative, and value-first.

Clarity. The proposed decision tree maintains clarity by formulating questions that align with the terminology used in the corresponding articles of the AI Act. Furthermore, additional information is provided within the decision tree to aid understanding of certain definitions or contexts.

Simplicity. The framework is designed to be simple and easy to implement by utilizing a decision tree structure. The questions are designed to have binary Yes/No options, and the longest path in the classification process is limited to a maximum of 20 questions. While this choice may be subject to debate, an iterative process has ensured that 20 questions are the minimum necessary to cover the relevant criteria.

Obligation-required. The outcome of the classification process is presented not only in terms of the risk class of a given AI system but also in terms of the type of obligations required for that system.

Sequential. Leveraging the sequential nature of decision trees, the classification process follows a sequential flow. The assessment begins with filtering prohibited risks, then high-risk systems, followed by limited risk systems. If an AI system does not fall under the limited risk category, it is categorized as having minimal or no risk.

Representative.Each question in the decision tree is designed to be representative of the latest amendment of the AI Act [11]. The questions use consistent terminology and definitions. However, for ambiguous terms such as "subliminal message," the question includes a representation using the definition provided in the AI Act. The majority of the questions represent specific articles of the AI Act.

Value-first. The proposed decision tree begins with a question related to Fundamental Rights and Union values, aiming to separate AI systems with potentially significant risks to these principles from those with lower risks. Furthermore, similar to the "Obligation-required" principle, the output of the decision tree indicates the values at risk for each AI system category, based on the AI Act draft.

In summary, the decision tree has been designed in accordance with the specified design principles. Therefore, this decision tree will be evaluated in the next stage of the Design Science Methodology to address SQ4. It is important to note that this decision tree can be considered a "modified" version, as it does not strictly adhere to the structure of a traditional tree, where each node only has one parent [33]. In this decision tree, certain questions can originate from multiple paths, such as Q19 - "Does it intend to interact with natural persons?" which can arise from Q1, Q14, Q11, and Q12 (see Table 6.1). Considering the advantages of the decision tree, the sequential nature of AI system classification, and to enhance communication within this report, this framework is referred to as the Decision Tree.

Finally, this decision tree will be used to answer SQ4 in the subsequent chapter.

6.3. Summary of the Chapter

In summary, the generated decision tree is constructed based on the selected themes extracted from SQ2: Protected Value, Objective/Intention, Use-Case/Technology, and Domain. The longest path for the decision tree is 20 questions sequentially made by assessing the possibility of Unacceptable Risk to No/Minimal Risk. The decision result will include the risk level and additional information such as value at risk, requirements of the associated risks, and exemptions for specific use cases.

In the end, the decision tree is analyzed, including its analysis corresponding to the formulated design principle. This analysis ensures that the decision tree is made accordingly to make it still within the boundary of the design principle.

7

Decision Tree Evaluation (Obvious Cases)

Chapter 7 addresses the sub-question: "How to evaluate the proposed framework and what improvements can be drawn from the evaluation?". This sub-question corresponds to the Evaluation phase within the Design Science Methodology. To achieve this sub-question objective, a series of interview sessions were conducted with AI experts from both legal and non-legal (technical) backgrounds, following the guidelines outlined in the Interview Protocol (refer to Appendix A.2). These interviews were designed to assess the performance of the decision tree and gain a deeper understanding of the classification process. The feedback and comments provided by the experts were analyzed and interpreted to formulate comprehensive recommendations for refining the decision tree framework.

The assessment of the decision tree's performance is divided across two chapters, with Chapter 7 focusing on the evaluation of the decision tree for Obvious Cases and Chapter 8 dedicated to evaluating the decision tree for Non-Obvious Cases, along with the qualitative analysis derived from the interviews.

Section 7.1 presents the results of this analysis, incorporating measures of similarity agreement such as Krippendorff's Alpha, accuracy, precision, recall, and the F1 score to evaluate the decision tree's performance specifically for Obvious Cases. Section 7.2 delves into a comprehensive discussion of the findings from each analysis. Finally, Section 7.3 summarizes the key points in this chapter.

7.1. Decision Tree Performance

This section presents the evaluation of the decision tree performance for Obvious cases. Data from 16 participants which composed from 7 legal experts and 9 non-legal experts who completed the experiment to evaluate the decision tree as mentioned in Chapter 3 were collected and used for the analysis. Based on the composition of the respondents, there were seven respondents with a legal background and nine with a non-legal background, which was more technical in nature, such as data scientists or researchers. Thus, analyzing the previous findings for the decision tree performance concerning the different backgrounds of the respondents can provide a broader perspective.

The analysis focuses on assessing the effectiveness and efficiency of the decision tree's performance. The effectiveness evaluation encompasses various criteria, such as Inter-Rater Reliability agreement quantified by Krippendorff's Alpha, accuracy, precision, recall, and F-1 score. A comparison is made between the two approaches: (1) Without DT (Decision Tree) approach, basically the approach without the decision tree (classification based on the AI Act article only), and (2) With DT approach, a decision tree-based classification. The quantitative performance results are further analyzed in subsequent subsections, considering the overall performance of both approaches, as well as the performance of the decision tree based on the respondents' backgrounds.

Simultaneously, in assessing decision tree efficiency, the time taken by each respondent to classify the use cases with and without the decision tree is measured. This time duration is separately recorded for each respondent and each use case.

However, it is important to note that due to the limited number of respondents, the quantitative analysis is presented to gain insights rather than provide definitive recommendations. The analysis will be complemented by qualitative analysis in subsequent sections.

7.1.1. Inter-Rater Reliability

The level of agreement was assessed based on widely accepted statistical reliability measures in conjunction with the percent agreement: Kα < 0.01, indicating poor agreement, Kα = 0.01–0.20 indicating slight agreement, Kα = 0.21–0.40 indicating fair agreement, Kα = 0.41–0.60 indicating moderate agreement, Kα = 0.61–0.80 indicating substantial agreement, and Kα = 0.81–1.00 indicating almost perfect agreement [13].

The overall inter-rater reliability for both approaches, Without DT and With DT, were computed and are presented in Table 7.1.

Table 7.1: Inter-Rater Reliability - Obvious Case

Approach	Krippendorff's Alpha		
	Legal Respondents	Non-Legal Respondents	All Respondents
Without DT	0.512	0.268	0.320
With DT	0.426	0.450	0.440

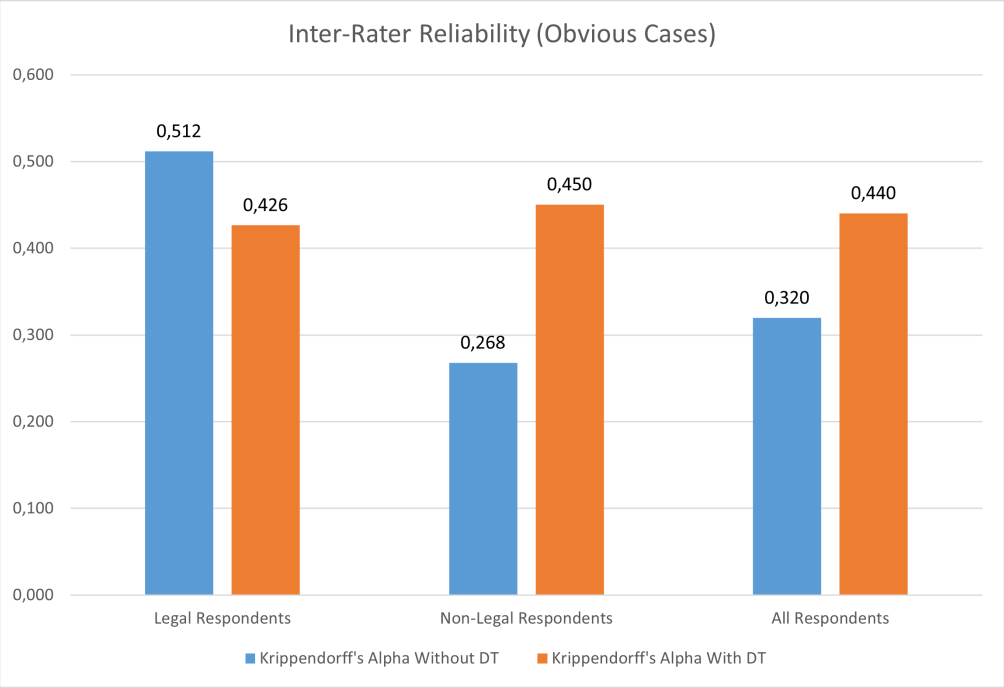

Figure 7.1: Inter-Rater Reliability of Obvious Cases (All, Legal, and Non-Legal Respondents)

When considering all respondents (legal and non-legal experts) who classified obvious use cases without the decision tree, a Krippendorff's Alpha value of 0.33 was obtained. In contrast, when using the decision tree approach, Krippendorff's Alpha value increased (Kα = 0.44), indicating that the experts tended to exhibit higher agreement when utilizing the decision tree than classification without it.

On closer examination of different types of respondents, Krippendorff's Alpha for legal respondents without the decision tree was 0.512, which showed a slight decrease compared to the approach with the decision tree. This decrement indicates that legal experts tend to exhibit less agreement when using the proposed decision tree and are more likely to have a higher agreement when they classify AI systems without it. However, both approaches indicated moderate agreement with and without the decision tree framework.

Conversely, for non-legal respondents, Krippendorff's Alpha was 0.268 for classification without the decision tree and 0.450 for classification with the decision tree. There was an increment in Krippendorff's Alpha when non-legal respondents classified obvious use cases, indicating that non-legal respondents were more likely to agree on the classification result when using the decision tree.

Comparing Krippendorff's Alpha for legal and non-legal respondents, it is evident that legal experts had less similarity agreement when classifying the Obvious Use Cases using decision tree. Meanwhile, the agreement of the non-legal respondents was slightly higher than that of the legal respondents, with Krippendorff's values showing Kα = 0.426 and Kα = 0.44, respectively.

7.1.2. Accuracy

In this study, accuracy refers to the correct classification of AI systems in comparison to their corresponding ground truth. Specifically, accuracy signifies the percentage of respondents who accurately classified the risk level of AI systems based on the determined ground truth.

Table 7.2: Accuracy (%) - Obvious Case

Approach	**Accuracy (%) - Obvious Case**		
	Legal	**Non-Legal**	**All Respondents**
Without DT	71.43%	61.11%	65.63%
With DT	78.57%	66.67%	71.88%
Differences	7.14%	5.56%	6.25%

The accuracy comparison, encompassing all respondents, including those with legal and non-legal backgrounds, is presented in Table 7.2 and visually depicted in Figure 7.2. Evidently, the decision tree enhances accuracy for all respondents, irrespective of their legal or non-legal expertise, when classifying AI systems with DT compared to the without DT approach.

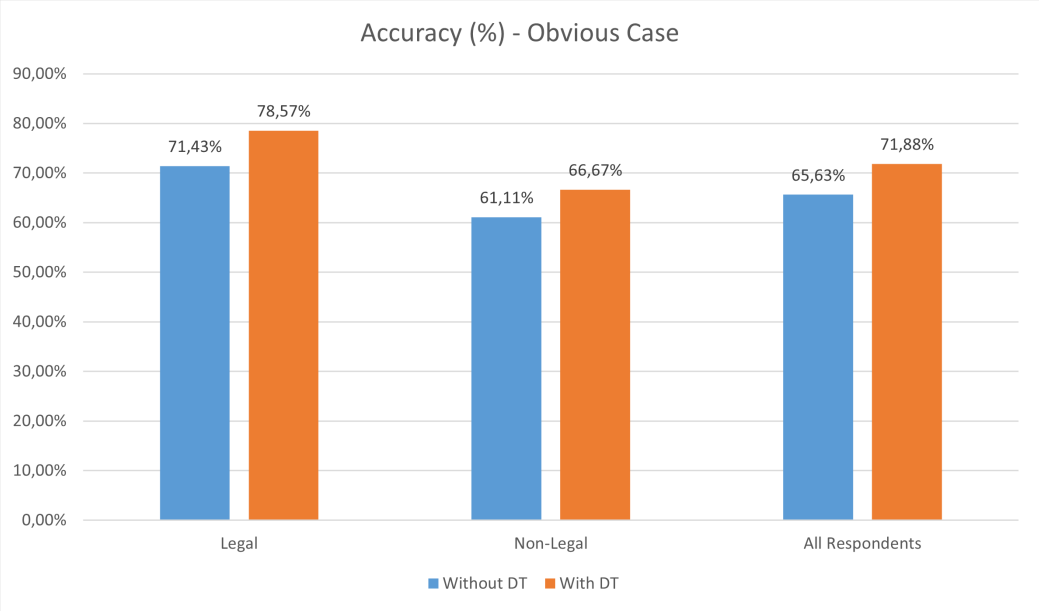

Figure 7.2: DT Performance: Accuracy (%) - Obvious Case

Furthermore, the improvement in performance with the decision tree is more significant for legal respondents during the classification of Obvious Cases, showing an approximate difference of 7.14% compared to classification without the decision tree. This suggests that the decision tree aids legal respondents in more accurate AI system classification.

From Figure 7.2, it is evident that the decision tree also contributes to improved performance for non-legal respondents when classifying AI systems. The enhancement is approximately 5.56%, albeit slightly lower than the performance improvement observed among legal respondents.

These findings collectively indicate that using the decision tree for AI system classification improves accuracy,

regardless of whether the respondents have legal or non-legal backgrounds. Notably the accuracy enhancement is particularly notable among legal respondents.

Additional metrics, such as precision, recall, and F-1 score, were evaluated to comprehend the reliability of these results, as detailed in the subsequent section.

7.1.3. Precision

Precision is another metric used to evaluate the performance of a decision tree. It measures the ratio of correctly predicted positive occurrences (true positives) to all instances projected as positive, including both true positives and false positives (Ouni, 2021) [36]. In the context of this study, precision denotes the ratio of accurately predicted AI systems use cases (true positives) relative to the total use cases predicted as positive.

Table 7.3: Precision (%) - Obvious Case

Approach	Precision (%)											
	UR			HR			LR			NR		
	Legal	Non-Legal	All	Legal	Non-Legal	All	Legal	Non-Legal	All	Legal	Non-Legal	All
Without DT	80.00%	57.14%	66.67%	50.00%	50.00%	50.00%	50.00%	66.67%	57.14%	100.00%	100.00%	100.00%
With DT	75.00%	66.67%	70.00%	100.00%	50.00%	75.00%	66.67%	50.00%	58.33%	100.00%	100.00%	100.00%
Differences	-5.00%	9.52%	3.33%	50.00%	0.00%	25.00%	16.67%	-16.67%	1.19%	0.00%	0.00%	0.00%

To delve deeper into the decision tree's performance, the precision calculation is presented in Table 7.3 and visually depicted in Figure 7.3. This visualization offers a comparative view of precision results, showcasing the precision differences between classification with and without the decision tree for all respondents, as well as for legal and non-legal respondents.

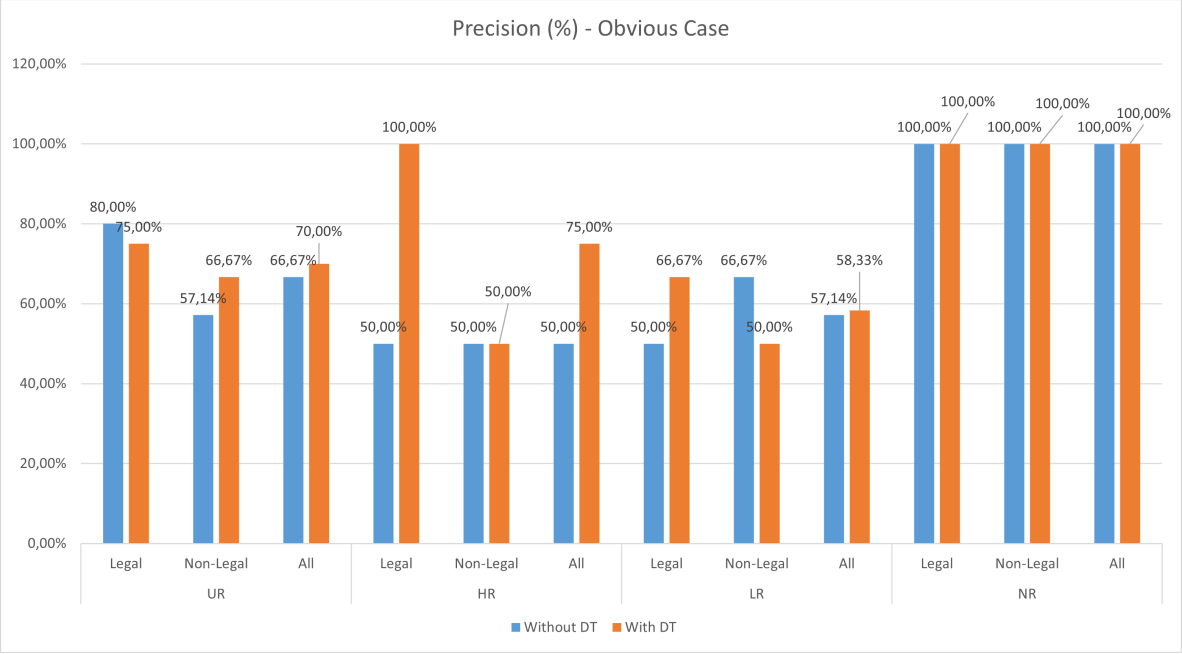

Figure 7.3: DT Performance: Precision (%) - Obvious Case

Generally, the precision of all respondents' classifications across all cases outperforms the classification without the decision tree, as demonstrated in Figure 7.3. This indicates that the decision tree slightly enhances the classification's precision. This improvement is particularly noticeable for the High-Risk use case, with a precision difference of approximately 25%. Additionally, for the No/Minimal Risk case, all respondents correctly classify the use case using both approaches— with and without the decision tree—suggesting no difficulties in classifying No/Minimal Risk cases.

Upon closer examination, for the Unacceptable Risk use case, all respondents exhibit a higher rate of classifying the case when using the decision tree, with a precision of around 70%, compared to 66.67% without the decision tree. However, differentiating between legal and non-legal respondents reveals that legal respondents display a decrement of approximately 5%, dropping from 80% for classification without the decision tree to 75% for classification with the decision tree. In contrast, non-legal respondents show an increment from 57.14% to 66.67%.

Regarding the High-Risk Use Case, all respondents demonstrate a significant improvement in precision, with differences of around 25%. This substantial enhancement arises primarily from legal respondents' precision values, which remain constant with both approaches (without and with the decision tree). In contrast, the Limited Risk use case exhibits a slight increase in precision, about 1.19%. The decision tree enhances classification for legal respondents, with a precision difference of 16.67%. However, non-legal respondents' precision seems to decline (-16.67%), as they could better classify the Limited Risk use case without the decision tree (66.67%) compared to using the decision tree (50%).

Lastly, for the No/Minimal Risk use case, all respondents, both legal and non-legal, correctly classify the use case, resulting in a 100% precision value, regardless of whether the decision tree is employed or not. This value underscores that all respondents can easily classify the No/Minimal Risk use case.

The higher precision values indicate that the decision tree yields a greater accuracy in classifying Unacceptable Risk, High-Risk, Limited Risk, and No/Minimal Risk cases, thereby leading to more precise identification of use cases than the approach without a decision tree. There is an exception, notably for Unacceptable Risk classification among Legal Respondents and Limited Risk classification among Non-Legal Respondents.

7.1.4. Recall

Recall means how many positive cases the decision tree correctly predicted over all the positive cases [36]. The recall calculation is presented in Table 7.4 and shown in Figure 7.4.

Table 7.4: Recall (%) - Obvious Case

Approach	Recall (%)											
	UR			HR			LR			NR		
	Legal	Non-Legal	All	Legal	Non-Legal	All	Legal	Non-Legal	All	Legal	Non-Legal	All
Without DT	100.00%	100.00%	100.00%	25.00%	75.00%	50.00%	100.00%	33.33%	50.00%	75.00%	50.00%	62.50%
With DT	100.00%	80.00%	87.50%	66.67%	20.00%	37.50%	80.00%	100.00%	87.50%	66.67%	80.00%	75.00%
Differences	0.00%	-20.00%	-12.50%	41.67%	-55.00%	-12.50%	-20.00%	66.67%	37.50%	-8.33%	30.00%	12.50%

Upon analyzing Figure 7.4, it is evident that, for all respondents, the decision tree excels in identifying Limited Risk and No/Minimal Risk cases, with both approaches (without the decision tree and with the decision tree) showing differences in recall value of 37.5% and 12.5%, respectively. Interestingly, the decision tree demonstrates comparatively lower performance in correctly identifying High-Risk and Unacceptable Risk cases, resulting in differences in recall value of -12.5% for both cases.

Turning to legal respondents, the decision tree showcases better performance in capturing High-Risk cases while showing no improvement for Unacceptable Risk cases (Figure 7.4). However, the recall score is lower for Limited Risk and No/Minimal Risk cases compared to the approach without the decision tree.

For non-legal respondents, the decision tree's performance is notably higher in classifying the Limited Risk and No/Minimal Risk use cases. Meanwhile, the recall performance for Unacceptable Risk and High-Risk cases has decreased compared to the classification performance without the decision tree.

Comparing the recall scores between legal and non-legal respondents, it becomes evident that the decision tree performs better for legal respondents in correctly classifying Unacceptable and High-Risk cases, whereas it performs better for non-legal respondents in correctly classifying Limited and No/Minimal Risk cases.

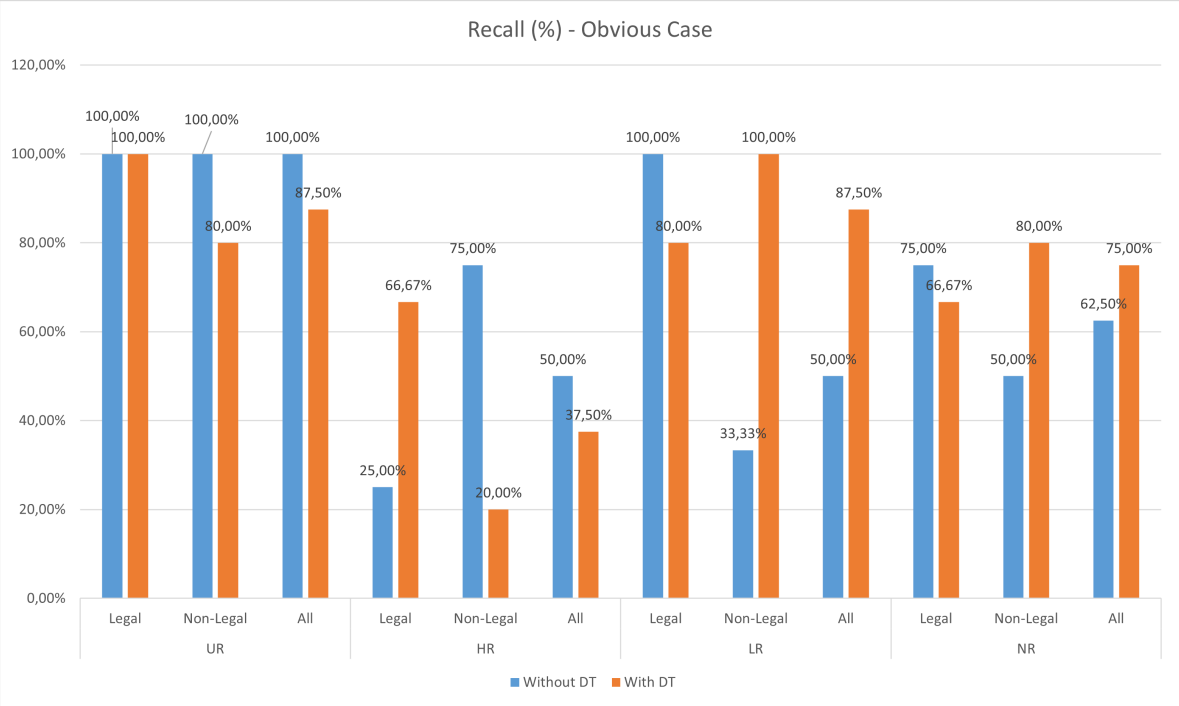

Figure 7.4: DT Performance: Recall (%) - Obvious Case

7.1.5. F1-score

The F1 score is a comprehensive metric that harmonizes precision and recall measurements [36]. This score represents the harmonic mean between precision and recall, and the F1 score values are presented in Table 7.5 and illustrated in Figure 7.5.

Table 7.5: F1 Score (%) - Obvious Case

Approach	F1-score (%)											
	UR			HR			LR			NR		
	Legal	Non-Legal	All	Legal	Non-Legal	All	Legal	Non-Legal	All	Legal	Non-Legal	All
Without DT	88.89%	72.73%	80.00%	33.33%	60.00%	50.00%	66.67%	44.44%	53.33%	85.71%	66.67%	76.92%
With DT	85.71%	72.73%	77.78%	80.00%	28.57%	50.00%	72.73%	66.67%	70.00%	80.00%	88.89%	85.71%
Differences	-3.17%	0.00%	-2.22%	46.67%	-31.43%	0.00%	6.06%	22.22%	16.67%	-5.71%	22.22%	8.79%

Reviewing Figure 7.5, it is apparent that the F1 scores for All Respondents, for both approaches (without and with the decision tree), exhibit improvement for Limited Risk and No/Minimal Risk cases, with F1 score enhancements of 16.67% and 8.79%, respectively. For High-Risk cases, there is no difference in F1 scores between the approach without the decision tree and the approach with the decision tree, with values of 50%. However, there is a slight decrement in the F1 score for Unacceptable Risk cases, declining from 80% without the framework to 77.78% with the framework.

The decision tree approach generally enhances the F1 score compared to the approach without the decision tree, except for Unacceptable Risk cases. This result indicates that the decision tree provides a balanced classification of AI systems, demonstrating good precision and recall for High-Risk, Limited Risk, and No/Minimal Risk categories.

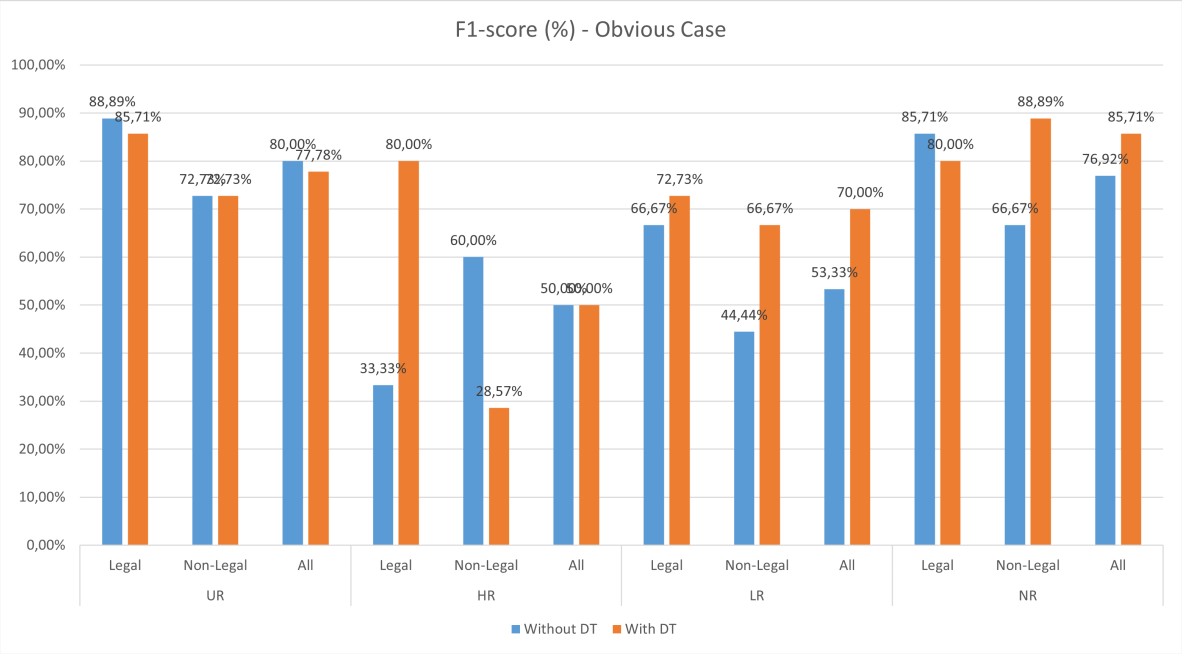

Figure 7.5: DT Performance: F1-score (%) - Obvious Case

Further delving into the F1 score for legal respondents, as depicted in Figure 7.5, the decision tree performs better regarding the F1 score for High-Risk and Limited Risk cases. However, for Unacceptable Risk and No/Minimal Risk cases, the F1 score is slightly lower compared to the classification without the decision tree.

On the other hand, for non-legal respondents, the F1 score is better for Limited Risk and No/Minimal Risk cases. For Unacceptable Risk, the performance of the classification with the decision tree is similar to the classification without the decision tree. However, the F1 score for non-legal respondents drops significantly from 60% for the classification without the decision tree to 28.57% for the classification with the decision tree.

In summary, for all types of respondents, the decision tree performance shows improvement in all cases except for the Unacceptable Risk case. However, the F1 score for Legal Respondents tends to exhibit higher improvement for High-Risk and Limited Risk classifications. Meanwhile, for non-legal respondents, the decision tree classification performs better in classifying all cases except for the High-Risk case.

7.1.6. Confusion Matrix

To gain deeper insights, we visualize the confusion matrix for All Respondents, both with and without the decision tree (Figure 7.6). This analysis aims to understand the patterns of AI system classification across all cases.

This section presents an in-depth examination of the findings, focusing on Figure 7.6, which illustrates the classification outcomes achieved with and without the decision tree framework.

It is evident from the analysis of this figure that respondents encountered challenges in identifying High-Risk, Limited Risk, and No/Minimal Risk. Specifically, with regards to the High-Risk classification, an intriguing trend emerged where approximately half of the respondents misclassified it as either Unacceptable Risk or Limited Risk. In the context of the Limited Risk category, respondents exhibited a tendency to classify cases as either Limited Risk or High-Risk, with one respondent even labeling it as Unacceptable Risk.

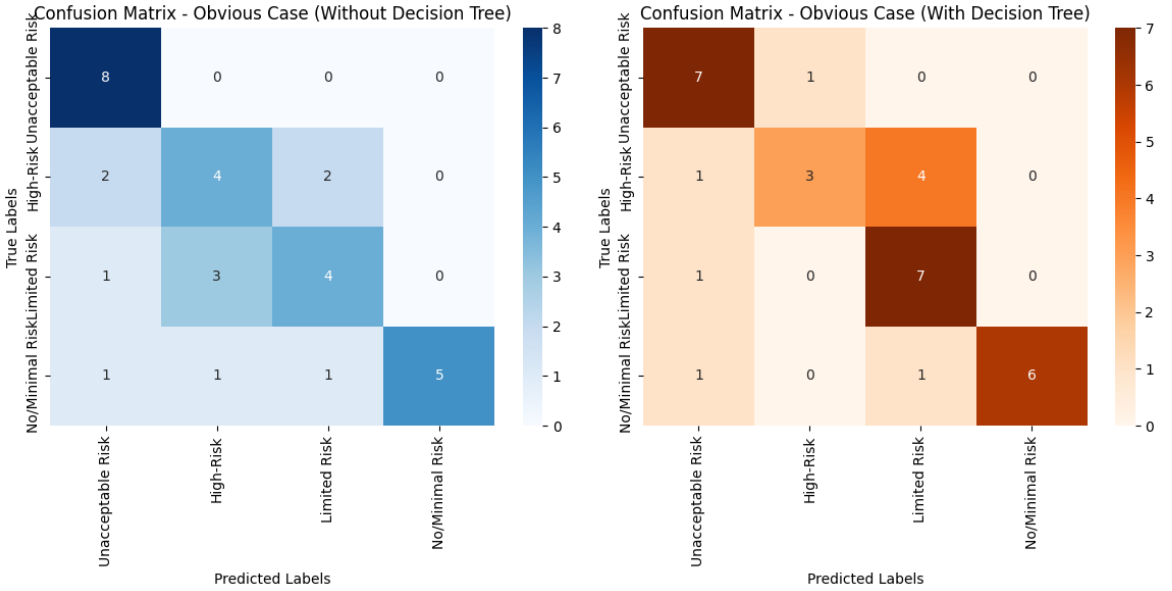

Figure 7.6: Confusion Matrix - All Respondents

On the other hand, when we compare these results with the classification based on the decision tree, a clear improvement in precision is observed for the categorization of Limited Risk and No/Minimal Risk. Specifically, in cases where there is Limited Risk, such as AI system that automatically converse with people in place for a human being and can interact with them (Case 8), an impressive seven out of eight respondents accurately labeled it as Limited Risk. This improvement is particularly significant considering the consensus among most respondents that this case poses minimal threat to Fundamental Rights and Union values, leading them to confidently assign it to the Limited Risk category. Interestingly, there was only one respondent who chose an Unacceptable Risk classification, motivated by a meticulous evaluation of whether the case distorts human behavior.

Additionally, the decision tree-based classification method highlights a trend where some respondents tend to misclassify High-Risk cases as Limited Risk. For example, the use of AI system use emotion recognition system to identify/recognize patient's emotion (Case 2) invoked diverse viewpoints. Nearly half of the respondents believed that this system, designed to assist patients, would not significantly breach Fundamental Rights or Union values. This perception led them to categorize it as Limited Risk. Conversely, the other half acknowledged substantial contravention of these principles but stopped short of deeming it outright prohibited as Unacceptable Risk, resulting in its classification as High-Risk.

To summarize, the implementation of the decision tree demonstrates an improved ability to differentiate between Limited Risk and No/Minimal Risk classifications in Obvious Cases. However, a significant concern arises regarding the potential misclassification of High-Risk cases as Limited Risk.

This analysis further delves into the performance disparities among respondents based on their backgrounds:

a. Legal Respondents

After carefully analyzing the interview data, it becomes evident that legal experts showcase a high level of proficiency in categorizing the provided use cases. Their skill shines particularly when distinguishing between Unacceptable Risk and Limited Risk, underscoring their expertise in legal terms. However, a noticeable variation arises in their perspectives regarding No/Minimal Risk and High-Risk case, with some experts categorizing them as Limited Risk.

Regarding Case 1 - AI system to filter unwanted mails and keep them separated from useful emails to reduce time and effort which determined as No/Minimal Risk, the majority of legal experts accurately label this case as Minimal/No Risk. However, Respondent O classified it as Limited Risk. This divergence can be attributed to the ambiguity introduced by the term 'interact' in Q19 (Does it indent to interact with natural persons?) and its contextual interpretation, leading to varied viewpoints.

Examining the High-Risk case concerning Case 2 - AI system use emotion recognition system to identify/recognize patient's emotion, most legal experts appropriately classify it as High-Risk. However, Respondent O's decision tree-based classification diverges, placing it under the Limited Risk category. Upon closer examination, Respondent O acknowledges this case is not significantly harm to the Fundamental Rights and/or Union values. Most respondents who classified this case as Limited Risk highlight that their classification was misguided due to not fully considering the potential harm to Fundamental Rights or Union values posed by the system.

In essence, legal experts, despite minor disparities, exhibit an improvement alignment between both approaches, with and without decision tree framework.

b. Non-Legal Respondents

Within a broader context, for non legal respondents, it's noticeable that the decision tree's classifications closely mirror the ground truth across different risk categories.

A compelling example pertains to the Limited Risk category, which involves Case 8 - AI system that automatically converse with people in place for a human being and can interact with them. Without decision tree, non-legal respondents exhibited diverse classifications, ranging from Unacceptable Risk and High-Risk to Limited Risk. This divergence emerged due to varied interpretations of terms such as "interaction" and "system design purpose." However, with the guidance of the decision tree, a consensus emerged, leading to a convergence of classifications among most non-legal respondents, now closely aligned with the ground truth. The decision tree's guidance effectively addressed the ambiguities that previously gave rise to differing viewpoints.

In summary, although some discrepancies persist, the majority of non-legal respondents' classifications align quite well with the ground truth, highlighting the decision tree's notable accuracy and agreement with the true classifications.

7.1.7. Time Performance

This study aims to assess the time efficiency of the decision tree classification by examining the time experts took to categorize AI system use cases. The initial expectation is that using the decision tree framework would lead to shorter classification duration than classifications conducted without the decision tree.

Table 7.6 displays the time taken by respondents in two distinct approaches to categorizing AI systems. As described in Table 3.3, each respondent categorized four use cases using both approaches. The boxplot diagram in Figure 7.7 provides a visual comparison of the two approaches.

Figure 7.7 reveals that the median duration for the "With DT" approach is significantly shorter than that of the "Without DT" approach, indicating a notable improvement in time efficiency.

Moreover, when examining the distribution of time durations, the "Without DT" classification displays a symmetric distribution, while the "With DT" classification has a right-skewed distribution. The "Without DT" approach shows a more uniform data distribution than the "With DT" classification, demonstrating increased variability.

From a broader perspective, the range of durations for AI system classification using the decision tree method spans from around 180 to 800 seconds, equivalent to 3 to 13.33 minutes. However, an outlier is noticeable within the "With DT" approach, attributed to respondent C, who took approximately 1586 seconds or 26.43 minutes. This outlier behavior can be attributed to respondent C's meticulous approach, characterized by comprehensive case-by-case analysis and thorough scrutiny of details. Given respondent C's legal background, it is clear that the decision tree framework aligns with his approach to AI system classification. Hence, they delved deep into the decision tree's questions and cases.

Table 7.6: Classification Duration per Respondent

Respondent	Duration (second)		Background	Familiarity with the AI Act
	Without DT	With DT		
A	57	113	Legal	Yes
B	388	251	Non-Legal	No
C	567	1586	Legal	Yes
D	742	600	Non-Legal	No
E	342	219	Non-Legal	No
F	729	765	Legal	Yes
G	302	244	Non-Legal	No
H	560	285	Non-Legal	No
I	186	395	Legal	Yes
J	737	368	Non-Legal	No
K	89	180	Legal	Yes
L	485	209	Non-Legal	No
M	189	209	Non-Legal	No
N	421	506	Legal	Yes
O	89	190	Legal	Yes
P	228	480	Non-Legal	Yes
Average	381,9	412.5		
Median	365	268		

Given the presence of an outlier, a median is calculated to determine the central tendency for both approaches, "Without DT" and "With T." The calculations confirm that AI system classification using the decision tree method is faster than without the decision tree. However, it is important to note that this outcome lacks statistical significance.

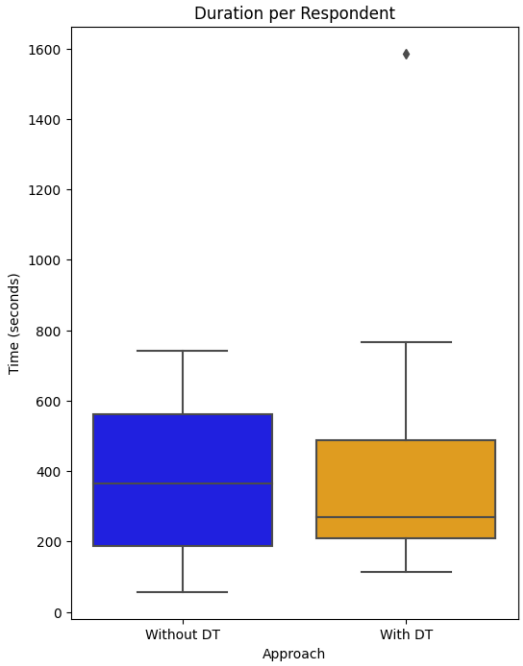

Figure 7.7: DT Performance: Classification Duration per Respondent

Upon analyzing Table 7.6, it becomes apparent that under the "Without DT" method, some respondents, namely A, K, and O, managed to classify AI systems in about one minute. Their swift classification of the presented AI system use cases can be attributed to their extensive legal experience and familiarity with the AI

Act, allowing them to categorize cases based on their intuitive interpretation of the legislation. This contrasts with the "With DT" method, in which the systematic examination of decision tree questions elongates the classification process.

In this section, we present the results of the classification duration specifically for Obvious Cases, as detailed in Table 7.7 and visually depicted in Figure 7.8.

Table 7.7: Classification Duration - Obvious Case (Average and Median)

Obvious Case	**Duration - Average (second)**		**Duration - Median (second)**	
	Without DT	**With DT**	**Without DT**	**With DT**
Case 1	83.53125	102.90625	78.5	66.625
Case 2	79.25	105.125	91	57.875
Case 7	121.125	77.59375	111.125	67
Case 8	106.625	75.125	95	66.125

Upon reviewing Table 7.7, it becomes evident that, in the context of Obvious Cases (Case 1, 2, 7, 8), the mean time required for classifying each case is generally shorter when respondents undertake AI system classification without utilizing the decision tree framework. This trend is observed, except for Case 2. However, considering the outlier mentioned above and focusing on the median values, the time required for AI system classification is reduced when employing the decision tree framework, in contrast to classification without the decision tree. This observation offers a preliminary insight into the potential time-saving attributes of the decision tree, albeit with a marginal difference of less than 50 seconds.

Figure 7.8 further illustrates the patterns observed. Under the "without T" approach, notable outliers arise in Case 1, Case 7, and Case 8. A closer examination of individual respondents' performance reveals that Respondent C took approximately 309 seconds to classify Case 1. This particular instance underscores the variability within the data and the potential impact of individual approaches on classification durations.

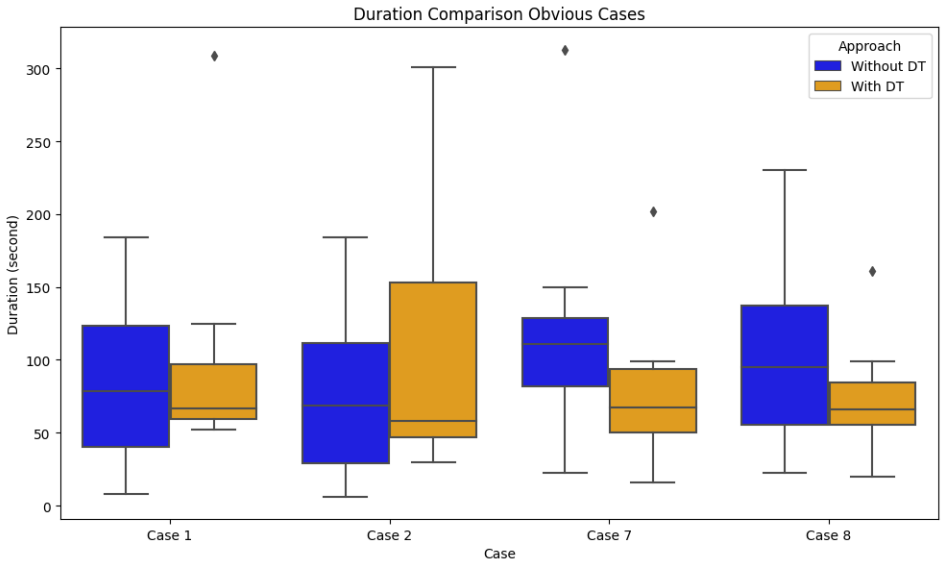

Figure 7.8: DT Performance: Classification Duration - Obvious Case

For Case 1, the outlier can be attributed to the response from Respondent F, who possesses a legal background. An analysis of the response indicates that when using the DT approach, Respondent F scrutinized the decision tree word by word and carefully considered the context of Case 1. This led Respondent F to classify Case 1 as No/Minimal Risk, corresponding to the decision tree's last assessment. However, starting from the initial question, which inquired whether the AI system has the potential to significantly harm Fundamental Rights or Union values, Respondent F answered with a "Yes," arguing that this AI system could pose a data

protection risk. Furthermore, Respondent F found it challenging to predict the potential consequences of the AI system in this case, potentially leading to significant harm. Consequently, Respondent F initially selected the Unacceptable Risk assessment and then proceeded to answer all subsequent questions until the latest assessment, No/Minimal Risk assessment. As a result, the longer time Respondent F requires to classify Case 1 can be attributed to this detailed process.

Turning to the outlier in Case 7 for both approaches, this particular case involves AI systems employing remote biometric identification of political protesters, potentially impacting the exercise of freedom of assembly and association. Respondent D's need for more context and technical insight into Case 7 contributed to the extended classification duration.

Another outlier in Case 7 pertains to Respondent P's classification using the DT framework. The extended duration can be attributed to the same reasons as the approach without the DT framework. Respondent P highlighted the need for additional context and technical information about the case, particularly regarding the systems underlying the AI system in this scenario. Understanding these underlying systems could potentially reveal other uses of the AI system and its implications.

Finally, the last outlier occurs in Case 8 for the Decision Tree framework approach, involving Respondent P. Case 8 revolves around an AI system that automatically converse with people in place for a human being and can interact with them. Respondent P's query regarding whether the system is fully automatic or pretends to be human introduces a degree of ambiguity, resulting in an unclear description of Case 8.

7.2. Results Discussion

This section provides a comprehensive discussion of the results the quantitative analysis.

7.2.1. Classification Performance for Obvious Cases without Decision Tree

For the classification of Obvious Cases without the decision tree, the classification of AI systems reveals that respondents exhibit moderate agreement among themselves in categorizing all obvious cases. However, there is room for improvement in this performance.

In terms of accuracy, the achievement of only 65.63% indicates that the classification process without the decision tree lacks standardization or tools to facilitate the classification of AI systems.

The analysis of interviews indicates that the classification carried out by all legal respondents without the decision tree showcases low performance in terms of reliability (reproducibility) and accuracy. This lower score suggests that respondents encountered difficulties and had differing interpretations when classifying AI systems.

In terms of the time taken for the classification process, the duration for each case is already relatively short. However, there is potential for improvement to expedite the classification process using the decision tree framework. The longer time required to classify cases might indicate issues in understanding the cases themselves or the AI Act articles.

7.2.2. Decision Tree Performance for Obvious Cases in General

In terms of similarity agreement between respondents, the decision tree enhances the level of agreement, especially among legal respondents, although the increase is slight. Nevertheless, there is potential for the decision tree to further enhance similarity agreement between respondents. This result implies that the proposed decision tree framework can potentially enhance the reproducibility of AI system classification.

In terms of accuracy, the decision tree generally brings about improvement in correctly classifying AI systems, making their categorization more consistent with the determined ground truth. When considering the F1-score, which harmonizes precision and recall, the decision tree enhances the performance of AI system classification.

Lastly, the time performance of using the decision tree to classify AI systems shows potential for speeding up the classification process compared to the approach without the decision tree.

Overall, using the decision tree to classify Obvious Cases enhances the classification of AI systems in terms of reproducibility, accuracy, and efficiency compared to the classification approach without the decision tree.

However, it is important to note that this improvement is not highly significant and could benefit from further enhancement.

7.2.3. Decision Tree Performance Considering Legal and Non-Legal Respondents

Examining the inter-rater reliability between legal and non-legal respondents in classifying Obvious Cases reveals that the similarity agreement among non-legal respondents is higher than among legal respondents. Nonetheless, the similarity agreement for both types of respondents when using the decision tree indicates moderate agreement (Kα = 0.41-0.60).

Upon observation, it is evident that the decision tree exhibits higher accuracy when used by respondents with a legal background than those without legal expertise. The decision tree performance among legal respondents surpasses that of non-legal respondents.

Comparing the recall scores between legal and non-legal respondents, it becomes clear that the decision tree performs better for legal respondents in correctly classifying Unacceptable and High-Risk cases, while it excels for non-legal respondents in correctly categorizing Limited and No/Minimal Risk cases.

In summary, for all types of respondents, the performance of the decision tree shows improvement in all cases except for the Unacceptable Risk case. However, the F1 score for Legal Respondents tends to exhibit a higher improvement for High-Risk and Limited Risk classifications. Meanwhile, for non-legal respondents, the decision tree classification fares better in classifying all cases except for the High-Risk case.

In conclusion, the decision tree's performance for classifying Obvious Cases is generally better among legal and non-legal respondents. This difference can likely be attributed to legal experts' familiarity with the AI Act and legal terminology, while non-legal respondents might encounter challenges in comprehending certain contextual aspects. This suggests that the proposed decision tree needs to effectively translate legal terms into more accessible language for individuals without a legal background. Terms that were particularly difficult to understand include 'interaction between human and machine' and comprehending the AI system's design purpose.

7.3. Summary of the Chapter

In conclusion, this study engaged in a series of interview sessions involving 16 participants encompassing both legal and non-legal backgrounds. The outcomes of these interviews were subjected to both quantitative and qualitative analysis. Chapter 7 will delve into the analysis focusing on the Obvious Case scenarios. The quantitative examination of Obvious Cases revolved around aspects such as similarity agreement, accuracy, precision, recall, F1-score, and the time performance of the decision tree. The insights from the discussion with respondents are also provided to further enrich the quantitative findings.

The study's findings revealed that the decision tree played a valuable role in categorizing and enhancing the comprehension of AI system classification for Obvious Cases. On the whole, the level of agreement among respondents was positively impacted, reflecting an improvement. Additionally, the decision tree showcased enhanced accuracy in the classification process compared to without decision tree classification. Furthermore, there is potential for time-saving when respondents employ the decision tree framework to classify AI systems, in contrast to when it is not utilized.

8

Decision Tree Evaluation (Non-Obvious Cases) & Qualitative Analysis

Chapter 8 represents a continuation of addressing the sub-question: "How to evaluate the proposed framework and what improvements can be drawn from the evaluation?". This sub-question falls within the Evaluation stage of the Design Science Methodology, with a specific focus on assessing the performance of the decision tree for Non-Obvious Cases.

For the Non-Obvious cases, the evaluation employs the measure of Inter-Rater Reliability, which seeks to measure the extent of agreement among respondents, both those with legal backgrounds and those without. Additionally, this section incorporates a qualitative analysis of the interview outcomes with the respondents.

The findings regarding the decision tree's performance for Non-Obvious Cases are presented in Section 8.1. This evaluation primarily centers on the metric of Inter-Rater Reliability. Section 8.2 provides an in-depth discussion of the qualitative analysis arising from the interviews. The outcome of each analysis is elaborated in Section 8.3. Lastly, Section 8.4 presents a summarizing overview of the chapter.

8.1. Decision Tree Performance

This section is dedicated to the evaluation of the decision tree's performance in addressing Non-Obvious Cases. As delineated in Chapter 3, assessing the decision tree's performance in classifying non-obvious cases entails quantifying similarity agreement using Krippendorff's Alpha as a measure for Inter-Rater Reliability.

An agreement table has been formulated to provide a comprehensive grasp of the subject. This table furnishes an intricate analysis of the decision tree's performance for each case.

8.1.1. Inter-Rater Reliability

The overall inter-rater reliability for both approaches, specifically classification without the decision tree and classification with the decision tree, has been computed and is presented in Table 8.1.

Table 8.1: Inter-Rater Reliability - Non-Obvious Case

Approach	**Krippendorff's Alpha**		
	Legal Respondents	**Non-Legal Respondents**	**All Respondents**
Without DT	0.200	0.157	0.231
With DT	0.121	-0.032	0.086

The outcomes of this calculation indicate that, concerning Non-Obvious Cases, there exists a higher level of similarity agreement among respondents when they perform classifications without employing the decision tree (Kα = 0.231), as opposed to when they utilize the decision tree (Kα = 0.086). Applying the decision tree for classifying Non-Obvious cases yields lower agreement levels among respondents. In contrast, the agreement tends to be more reasonable when the classification is carried out without incorporating the decision tree.

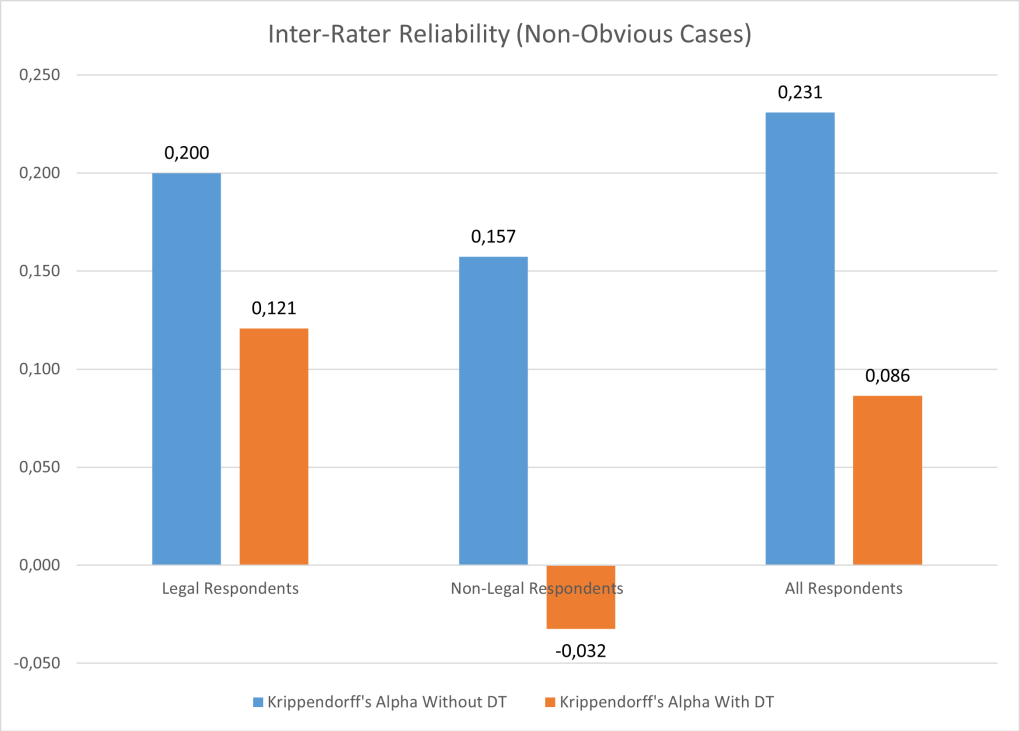

Figure 8.1: Inter-Rater Reliability of Non-Obvious Cases (All, Legal, and Non-Legal Respondents)

Figure 8.1 visually demonstrates this trend, showcasing a decline in Krippendorff's Alpha from classifications conducted without the decision tree to those conducted with the decision tree. This pattern remains consistent regardless of whether the respondents are legal experts. This observation implies that, in this particular context, the decision tree does not enhance the similarity agreement between respondents in the classification of AI systems. Instead, it seems to lead to an elevation in disagreement among respondents.

The subsequent sub-section presents an agreement table for a more comprehensive analysis of the factors contributing to this diminished agreement among respondents for specific Non-Obvious cases. This table offers a detailed breakdown to facilitate a more thorough exploration.

8.1.2. Agreement Table

The agreement table, depicted as a seaborn matrix, illustrates the degree of classification agreement among all respondents for specific cases. Figure 8.2 presents the agreement table. As explained in Chapter 3, the Non-Obvious cases are denoted as Case 3, Case 4, Case 5, and Case 6. In the table, these case numbers are allocated as rows, while the columns correspond to the risk levels (Unacceptable Risk, High-Risk, Limited Risk, and No/Minimal Risk).

From the insights provided by Figure 8.2, it becomes apparent that, without the decision tree, respondents tend to exhibit higher levels of agreement when classifying Case 4 as a High-Risk instance. For Case 3, respondents are divided, half categorizing it as an Unacceptable Risk case, while the remaining respondents assign it to High-Risk, Limited Risk, or No/Minimal Risk categories. In the case of Case 5, the classifications tend to sway towards No/Minimal Risk or Limited Risk. As for Case 6, a majority of respondents classify it as an Unacceptable Risk, while others categorize it as High-Risk.

However, when the decision tree is employed for classification, the similarity agreement among respondents does not display an improvement. Specifically, for Case 3 and Case 6, the level of agreement among respondents remains comparable to that observed in the classification without the decision tree. On the contrary, for Case 4, the level of agreement diminishes compared to the classification without the decision tree. Notably, employing the decision tree predominantly classifies Case 4 as an Unacceptable Risk instance. For Case 5, the classification pattern remains consistent with the decision tree, where respondents tend to categorize it as either No/Minimal Risk or Limited Risk.

8.1.2.1. Agreement Table for All Respondents

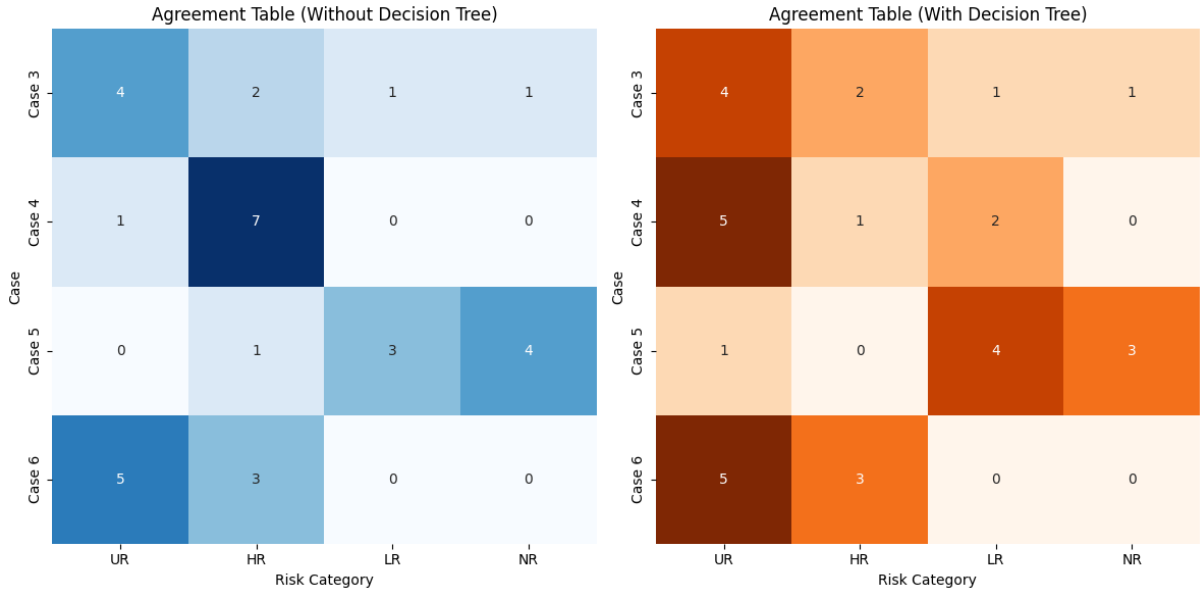

Figure 8.2: Agreement Table - All Respondents

The analysis of Case 4 is particularly intriguing, as it reveals divergent classification outcomes between the two approaches—classification with and without the decision tree. Most respondents categorized Case 4 (AI systems designed for social robots for children with autism to capture their behavior to assist treatment) as High-Risk. This classification resulted by the fact that the object in this system is 'children with autism' which considered as vulnerable groups and the consideration that this treatment is purposefully design to bring benefit to those children with autism. Consequently, Unacceptable Risk is deemed an unlikely classification due to the system's beneficial intent for the vulnerable group. Similarly, Limited Risk classification is also avoided, as the system's involvement with vulnerable individuals requires more than just 'transparency requirements.' Consequently, the majority of respondents deemed this case to fall under the High-Risk category.

With the decision tree, the outcomes varied, with classifications mostly divided between Unacceptable Risk and Limited Risk. The Unacceptable Risk classification stemmed from the initial question, where respondents perceived the AI system to pose significant harm to Fundamental Rights and/or Union values. Conversely, the Limited Risk classification was derived from respondents' determination that the social robots and their associated benefits would not significantly harm Fundamental Rights and/or Union values. This conclusion led to the Limited Risk classification after considering the AI system's interaction with natural persons (humans). Consequently, the classification of this AI system using the decision tree framework exhibited a division into two distinct groups.

To provide a deeper understanding, each case will be qualitatively expounded upon, along with specific considerations for both legal and non-legal respondents.

a. Legal Respondents

In the context of Non-Obvious cases, legal respondents without the decision tree encountered differing opinions when classifying cases 3, 4, and 6. However, for Case 5, most legal respondents concurred that this AI system should be classified as Limited Risk. In the absence of the decision tree, Case 5 was often classified as No/Minimal Risk.

In the case of classifying an AI system designed to measure a truck driver's fatigue and play a sound to encourage longer driving (Case 3), the outcomes for both approaches were similar, with classifications spanning from Unacceptable Risk to High Risk. The argument for classifying this case as Unacceptable Risk is grounded in the notion that it distorts human behavior by encouraging overwork. On the contrary, some respondents

saw it as a High-Risk scenario, viewing the system as beneficial support within the work environment—a category covered by the AI Act. With the decision tree, legal respondents agreed that the AI system would cause significant harm to Fundamental Rights and/or Union values. However, there was less consensus on whether the system would lead to behavior distortion or benefit the driver and employer.

Similarly, legal respondents held diverse opinions when classifying Case 6 (AI system to assess recidivism risk through quantitative risk assessments). With or without the decision tree framework, classifications were dispersed across Unacceptable Risk and High-Risk categories. The argument for classifying this case as Unacceptable Risk stems from the use of this AI system in law enforcement, a context mentioned in Article 5 of the AI Act. Legal respondents believed that recidivists are part of the vulnerable groups outlined in Article 5, which prohibits exploiting such groups. Conversely, others contended that the case should be classified as High-Risk due to its applicability in supporting law enforcement authorities. One respondent emphasized the significance of technological details when quantitatively assessing recidivism risk. If biometric data is involved, it should be categorized as Unacceptable Risk; if not, it should be classified as High-Risk.

In the case of Case 4—an AI system designed for social robots to assist in treating children with autism by capturing their behavior—legal respondents' opinions also diverged. Without the decision tree, a majority categorized this case as High-Risk, considering its potential to both benefit and potentially harm children with autism's Fundamental Rights and/or Union values. However, using the decision tree, the classifications shifted between Unacceptable Risk and Limited Risk. Unacceptable Risk classification was rooted in the assumption that children with autism constitute a vulnerable group, making the AI system unsuitable for them. Conversely, Limited Risk classification was based on the belief that the AI system would not cause significant harm to humans, leading the decision tree to a direct Limited Risk assessment. In Case 4, the decision tree generated more diversity and less agreement among legal respondents.

Interestingly, unlike the prior cases, the decision tree yielded greater agreement among legal respondents in classifying Case 5 (AI system for automatic speech transcription or enhancement). Legal respondents concurred that this AI system posed Limited Risk. They believed it would not significantly harm Fundamental Rights and/or Union values, but it should still comply with transparency requirements. Without the decision tree, legal respondents debated whether this case should fall under Limited Risk or No/Minimal Risk. Although all legal respondents agreed that this AI system would not pose significant harm to Fundamental Rights and/or Union values, consensus was lacking regarding the necessity of transparency requirements.

In summary, the decision tree resulted in higher agreement among legal respondents for Case 5. However, for Cases 3, 4, and 6, the decision tree did not seem to enhance agreement among legal respondents.

b. Non-Legal Respondents

For non-obvious cases, the classification patterns of AI systems became more diverse when considering respondents without a legal background.

In the case of Case 3—an AI system measuring a truck driver's fatigue and playing sounds to extend driving time—both with and without the decision tree, classifications spanned all risk levels. Arguments supporting Unacceptable Risk centered on the belief that "pushing drivers to drive longer" violates human dignity and distorts human behavior. In contrast, other respondents contended that the AI system would benefit drivers and employers, making it suitable for work environments (Article 6 - High-Risk). Some even argued that the AI system would not harm Fundamental Rights and/or Union values, resulting in Limited Risk or No/Minimal Risk classifications. No significant agreement emerged among respondents.

Case 4—an AI system designed to capture behavior in children with autism to aid treatment using social robots—generated consensus without the decision tree, with non-legal respondents largely classifying it as High-Risk due to its potential to aid children with autism. However, with the decision tree, most non-legal respondents classified it as Unacceptable Risk. This was based on the assumption that the AI system would significantly harm Fundamental Rights and/or Union values, potentially exploiting vulnerable groups like children with autism. This was compounded by the involvement of biometric or emotion recognition systems to capture children's behavior.

The classification of Case 5 (AI system for automatic speech transcription or enhancement) also displayed variation across all risk levels, with no significant agreement among non-legal respondents. Notably, Respondent J classified this case as Unacceptable Risk, holding the belief that all AI systems inherently pose significant risks to humans. Excluding Respondent J, the classifications hovered around Limited Risk and

No/Minimal Risk. This indicated that non-legal respondents, except for Respondent J, agreed that the case would not significantly harm human behavior. However, opinions differed on whether this case should adhere to transparency regulations, with some contending that the AI system's interaction with humans warranted transparency.

Lastly, the classification of Case 6—an AI system assessing recidivism risk through quantitative assessments—varied between two risk categories: Unacceptable Risk and High-Risk. The pattern persisted regardless of the presence of the decision tree, indicating that the decision tree's performance was not enhanced. All respondents agreed that Case 6 would cause significant harm to Fundamental Rights and/or Union values. While one respondent argued for High-Risk classification due to the AI system's application in law enforcement (Article Annexes III – High Risk), the majority viewed it as Unacceptable Risk. They believed the quantitative risk assessment, possibly involving biometric systems, related to a vulnerable group, causing potential misjudgment. The different interpretations of the AI Act's terms contributed to this disparity.

Overall, the decision tree did not amplify agreement levels among non-legal respondents across the presented cases. Nevertheless, it aided non-legal respondents in assessing whether the AI systems significantly harmed Fundamental Rights and/or Union values. It also highlighted difficulties in comprehending terms like 'significant harm to Fundamental Rights and/or Union values,' 'vulnerable groups,' and 'interaction between human and machine,' contributing to lower agreement levels among respondents.

8.1.3. Time Performance

The time measured when non-legal respondents classified each non-obvious cases is presented in Table 8.2 and visualized in Figure 8.3. The non-obvious case in this study is plotted in Case 3, 4, 5 and 6.

Table 8.2: Classification Duration - Non-Obvious Case (Average and Median)

Non-Obvious Case	**Duration - Average (second)**		**Duration - Median (second)**	
	Without DT	**With DT**	**Without DT**	**With DT**
Case 3	75.6875	114	63.75	80
Case 4	102.5625	131.09375	90.25	58.75
Case 5	87.40625	104.3125	69.25	93.5
Case 6	107.6875	114.84375	89.5	67

From Table 8.2, it is evident that non-legal respondents required more time to classify non-obvious cases when utilizing the decision tree than those without the decision tree. Analyzing the average duration reveals that all cases are processed more quickly without the decision tree.

Further examination of the data is presented in Figure 8.3. Notably, all cases display outlier data. To assess the central tendency of the data, the median is calculated. However, even after excluding the outliers, the median itself demonstrates that using the decision tree does not yield superior performance in classifying non-obvious cases, except for Case 4 and Case 6.

A closer investigation of the outlier data aligns with the explanation for total data outliers in Sub-section 7.1.7. Specifically, it is observed that all outliers in Case 3, 4, 5, and 6 for the "With DT" approach are attributed to Respondent C. During the interview, Respondent C took longer to classify each case due to meticulous consideration of the AI systems and the decision tree questions. This detailed approach may stem from his legal expertise, indicating a careful evaluation process.

Conversely, the outlier in Case 4 for the "Without DT" approach is attributed to Respondent F. While classifying Case 4—an AI system designed for social robots to assist in treating children with autism—he meticulously reviewed the AI Act article by article. He deliberated each possibility, ultimately determining that Case 4 does not violate the prohibitions outlined in Article 5. His careful step-by-step evaluation consumed more time compared to other cases.

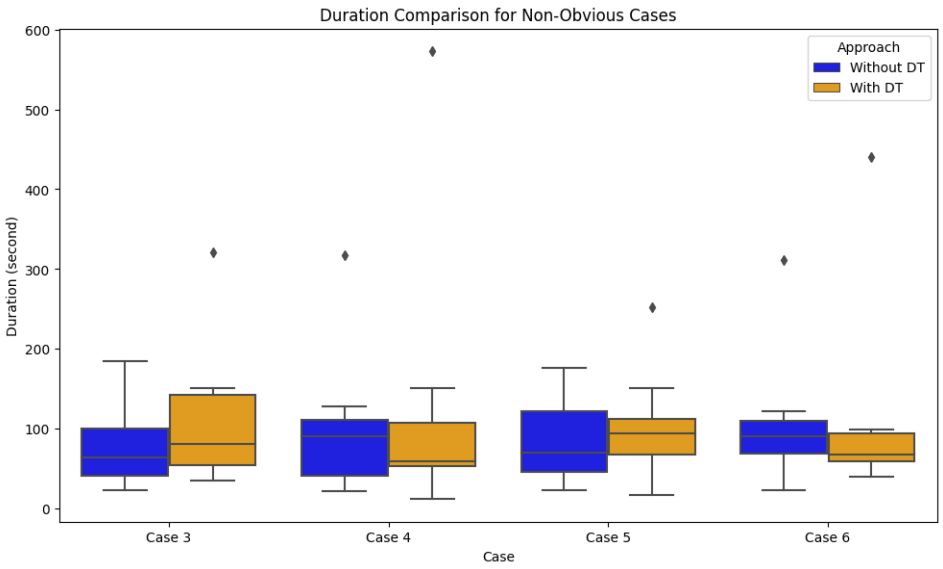

Figure 8.3: DT Performance: Classification Duration - Non-Obvious Case

Another outlier arises from Case 6, involving AI systems using remote biometric identification of political protesters to suppress freedom of assembly and association. Here, Respondent D, lacking legal expertise, encountered confusion while interpreting the term "significant chilling effect" in Case 6. Once he grasped its meaning, he rapidly classified Case 6 as an Unacceptable Risk.

In summary, the challenges in understanding the contextual nuances of Cases 3, 4, 5, and 6 influenced the time required for their classification. A median analysis reveals that using the decision tree decreased classification duration for Case 4 and Case 6 while prolonging it for Case 3 and Case 5.

8.2. Qualitative Analysis

To enhance the insights of this research, a semi-structured interview was conducted to gather feedback on the performance of the decision tree and the process of classifying AI systems. Qualitative analysis of the interview data revealed several high-level insights:

8.2.1. Benefits of Proposed Decision Tree

All respondents (100%) expressed that the decision tree was helpful in categorizing AI systems, citing various benefits associated with its use.

1. Helpful to categorize

Firstly, the decision tree was found to be helpful in terms of categorization. Respondent I highlighted that it increased awareness of the risks associated with AI systems, which might be overlooked when reading articles. Respondent M echoed this sentiment, emphasizing the benefits of considering different circumstances instead of lengthy paragraphs.

"So that's also very useful because yeah, again, you see high risk for example. And if you don't really look very closely to this articles, you will not see that the notion of risk also pertains to fundamental rights, but also to the environment and safety. So it's really nice" - Repondent I

"Decision tree is helpful, instead of more paragraph. More beneficial to read several circumstances" - Respondent M

According to Respondent L, the decision tree facilitated a quick and comprehensive understanding of what is allowed and not allowed, enabling a rapid overview of the AI Act. Respondent N added that the decision tree aided in categorizing AI systems into two main groups: concerning categories (Prohibited Risk and High-Risk) and non-concerning categories (Limited Risk and No/Minimal Risk).

"Like I said before, uh, I like the fact that the things that are not allowed on one side. So you can really easily see them and I think it also allows you to skim the entire thing quite quickly" - Respondent L

"It was very good because it separate the components from the AI act into yes, no questions. Or if then questions whatever you wanna call it and it helps to understand the differences, especially between this." - Respondent N

2. Structured and Visualized

Secondly, respondents appreciated the structured and visual nature of the decision tree. Respondent B noted that it allowed for a clear and structured perspective on the regulations, making it easier to classify AI systems. Respondent C highlighted that the decision tree enhanced understanding of the AI Act and the associated risk classes.

"As I explained before, this classification three really helps me to have my way of thinking in a more structured way big in comparison with just reading the acts. Because as we know that law language is always a bit boring and then this one help us to to see the thing in more clear perspective and in a more structured way. So the keyword is in structured way, so you you make everything more structure and more visualize and you confront like a words branches of words into semi figure. It helps a lot" - Respondent B

"You basically modeled out for making the decision tree that is a lot and that is not close at hand like you have to know the structure of the law in order to find even your risk classes. So yeah, that's helping. That's making knowledge structured" - Respondent C

3. Necessity of the Assessment Model like the Proposed Decision Tree

Thirdly, respondents emphasized the necessity of a risk assessment model like the proposed decision tree. Respondent A stated that with the enforcement of the AI Act, such a model becomes increasingly important as it simplifies the classification process and logically assigns risk levels to AI systems. Respondent J further supported this viewpoint, emphasizing the need for such tools in the operationalization of the AI Act.

"I think it was definitely a very nice research on interpreting the risk assessment model because I feel like we do need something like that. Because if we don't really have that kind of pathway to get to that final answer or what kind of a risk is an AI system pose and this is that pathway. So it really simplifies things a lot. It makes it easier. It helps you logically place your answer." - Respondent A

"With the operationalizing of the AI Act, so I think that the development of this kind of tools will be more and more needed" - Respondent J

4. Mental road map in classifying AI systems

Fourthly, respondents regarded the decision tree as a mental roadmap for classifying AI systems. Respondent A likened it to a visual representation of their own thought process in attributing risk to different AI systems. It provided a step-by-step approach to determine the appropriate risk class.

"I started seeing the kind of mental road map that I use in my mind in front of me in front of my eyes. So yeah, so very simple like you know, it was just it. I just saw a visual representation of the way that my mind works when I think of attributing risk to different AI systems. So that was really cool to look at." - Respondent A

5. Concise framework

Lastly, respondents commended the decision tree for providing a concise framework. Respondent F noted that the decision tree offered a more concise representation of the AI Act compared to the actual text, eliminating unnecessary prompts and including only key points relevant to classifying AI systems. This concise framework was particularly beneficial for individuals who were not accustomed to reading legal texts.

"It is helpful because it just gives a much more concise framework than the actual text of the AI act, which can be useful for people who are not used to reading legal text. It also provides in this sort of necessary prompts and it excludes all the ones that are not relevant. Based on what you've chosen earlier, so that can be useful for all, especially if you're not used to dealing with legal text, you don't have to look at the stuff that's not relevant" - Respondent F

8.2.2. Growing Concerns of Proposed Decision Tree

In addition to the mentioned benefits, respondents expressed growing concerns regarding the decision tree. These concerns revolved around the need for legal consultation, adaptability of the decision tree, level of understanding of the AI Act, assumptions, and personal biases.

1. Double check to Legal Experts

Despite the decision tree's usefulness in categorizing the AI Act, Respondent A emphasized the importance of consulting legal experts to ensure that the output and explanations align with the AI Act. Respondent G echoed this sentiment, indicating that while they were satisfied with the decision tree, they would still consult the law book to verify the classification of AI systems.

"Just just double check with the legal experts if this applies as well, because there was this one thing I feel like I noticed which you then changed the term of the bot interaction with people". - Respondent A

"I would rather consult with the legal team. Yeah, and not relying completely to the decision tree itself because it fails for some. Yeah, for some scope for some edge cases" - Respondent G

2. Adaptability of the decision tree

Respondent A highlighted that the AI Act draft is subject to change before its establishment, and therefore, the decision tree needs to be adjusted accordingly once the AI Act is enforced in the EU.

"Like you just said it's like obviously your first model, you might end up making changes to it in the future as things new things come up. It's a very good first step. I remember I told you, like I don't like the how I think of this classification on a scale. But then again, that's going to make it as obscure as it is already. So I think for now it's it's pretty good, it's pretty good" - Respondent A

3. Level of understanding of the AI Act

Respondents A and F expressed concerns regarding the level of background knowledge required to comprehend the AI Act. The decision tree's use of terms consistent with the AI Act's definitions could lead to incorrect answers if users lack a sufficient understanding of the law. This issue is particularly relevant for individuals with technical backgrounds who may not fully grasp the underlying context and reasons behind the regulations.

"I think the weakness would be if somebody, I mean you need to have a certain level of background of understanding that you AI acts. I would say like if somebody's following the tree, they can get to the wrong answer. Not because of something being wrong with the tree, but just because of not having sufficient knowledge. You know, they might be like, OK, because you need to have a certain understanding of how the law works and how things are argued in courts and how evidence works. Someone with no knowledge, They might trace it wrongly. You know, they might be like, I think a lot of, like, personal opinions come into play, right. Like, I think that this is this effects fundamental rights. When it actually does not, it could be a bit problematic on that end, but I don't think it's a creators problem. I think it's more so a user's problem." - Respondent A

"The drawback might be that especially if you haven't actually read the legal text and you're using this sort of say you're an AI developer and you don't really want to look into the law and you're just using this tool, then yeah, then you don't really have the context behind why these rules exist in what they're for, which might make it more difficult to understand these." - Respondent F

4. Assumptions

Respondent N pointed out that despite the decision tree's clear criteria and interactive nature, it relies heavily on assumptions, such as the definitions of human rights and significant risk. Different interpretations of harm and risk may arise due to insufficient knowledge, leading to divergent perspectives among users.

"The only thing is that it relies a lot of a lot of the assumptions right of the of the interviewer. What's human rights? What's the judgment? Is it taking the place of a natural person or not? But in the end, I think that you've done the most you can do, which is to provide a clear criteria in a dynamic and interactive way." - Respondent N

5. Personal bias

Similar to assumptions, classifying AI systems using the decision tree may be influenced by the personal biases of respondents. The individual judgments and perspectives of legal and ethics experts can differ in their perception of AI systems based on their assessment of risks.

"So my decision could be biased. It could be influenced by my personal judgment. It is very possible, especially several design, that in several system that I involve in the design is also to the micro level. So my judgment could be influenced by my personal experience." - Respondent B

8.2.3. Additional Insights

In addition to the benefits and concerns discussed earlier, several additional insights emerged from the interviews, which can contribute to refining the decision tree and enhancing the classification of AI systems.

1. Interdisciplinary approach

Respondents A and J emphasized the importance of adopting an interdisciplinary approach and involving experts from various fields, not just legal professionals, to develop clear guidelines for classifying risks and ensuring the effectiveness of the classification system. They suggested combining scientific and management perspectives to gain a broader understanding.

"It's a very fresh take on the risk assessment model and I think that we need more clear guidelines on how to classify risk and this is a step in that direction. So I'm I'm really like your research. See, these are the kind of fields we need right now, which kind of just and you know, they combined the scientific aspect of things and also the management aspect of things" - Respondent A

"Meaning that we talk of, say, intentional manipulation. I mean, there are many words that have to be filled with practical and legal meanings in order to, you know, make the the classification system work." - Respondent J

2. Intriguing cases

During the interviews, it became evident to several respondents that there are still many ambiguous AI systems when it comes to classification. By exploring the classification of AI systems in this research, intriguing cases were identified, such as autistic robots falling into different risk categories. However, to effectively regulate such systems, it is crucial to determine the appropriate risk classification.

"I just also didn't expect how complicated it is because if you just see this. Examples. That's the AI act provides that you think that's OK. This is straightforward. Yeah, this is simple classification, but then it made me realize that there are so many cases that are not clear cuts and I think they propose the ACT is also not very helpful because there's lots of confusing definitions and terms that we can also require some additional kind of." - Respondent I

3. Standardizing formulation of AI systems

Respondent P said during the interview that particular use cases need more information in order to classify this AI systems, otherwise the respondent will use many assumptions.

"What's your understanding of this specific use cases and also what the you as the respondents want to what information that you need to know in order to classify this AI systems". - Respondent P

In order to improve the classification of AI systems, several respondents suggested including information about the type of data consumed, data use and handling, social context, output, technical details, distribution of responsibility, purpose and intention, information storage, and the system's users. However, further research is needed to determine the precise information required for effective classification..

"What type of data it consumes and what type of yeah. And what and how people are using them? And output. What type of data it consumes? What type of data they are you handling and processing?" - Respondent D

"what is social context in which it's being used?" - Respondent F

"And then what AI model they implemented? What is the output? What technicality specifically that you maybe that you need? Where for the data is really what kind of data, how they collect the data, it make an cause an ethical issue" - Respondent H

"I mean like the system should be able to trace back all of its action to one human in, in in the control in the design of this AI system - Distribution of Responsibility" - Respondent J

"Of course it's important, but what you need to know is the purpose of this system. The intention of the system and after that what is the output of this air system and also what what is the information that will be stored? Who will have the information access the subject? Who will have the information access and also from the impact assessment?" - Respondent O

4. Changing perceptions

Another insight gathered from the interviews was the changing perceptions of the respondents. Prior to using the decision tree, respondents B and G acknowledged that they did not give much consideration to personal privacy or associated risks when designing AI systems. However, through this research, they started thinking about the risk classifications for each system. Several respondents expressed the need for increased awareness regarding the risks associated with AI systems.

"My perception or paradigm before I take this interviews, I never really think about this kind of risk. Differentiation between, for example, an unacceptable risk or high risk. Artificial intelligence in such a way could lead to breaching personal privacy or personal information or being misused by several group of person. But I don't really think about this kind of perspective. To see the things in classification perspective. So it will also impact a after this decision of interview. Every time I heard about AI system, I will start to think about where which classification risk is the system addressing about?"- Respondent B

"Learned something today like that's called like fundamental value of human being or whatnot. So yeah, it's a good thing and that's make us also learn in terms of if I work on the company that create that kind of things I and and know I knew that this is the the basic" - Respondent G

5. Beneficial for developer to understand the boundaries

Respondents E and G, who possess technical expertise, emphasized the significance of understanding the AI Act through the decision tree to comprehend the boundaries of risk. Respondent P added that team leaders, AI creators, and developers should be well-informed about the AI Act.

"So it helpful in a case for you as developer to understand on the boundaries of the risk." - Respondent E

"But yeah, it's really throw me a line. Which one is like the prohibited one? Which one that really endangered the the freedom of of voice in the user, not user. I mean people that I also learned something about this fundamental failure"- Respondent G

"The idea is that any creators or developers or specific AI systems or data sets, they should actually be reading this in order to write proper system" - Respondent P

6. Vagueneess of the terms in regulation

Legal and non-legal experts among the respondents expressed difficulties in interpreting certain terms in the regulations, such as "significant harm" to fundamental rights. Even for legal professionals, the definition of significant harm may vary. They suggested the need for additional guidance or examples to clarify ambiguous terms in the regulations.

"Is probably just the vagueness of the terms in the regulation. Things like significant harm to fundamental rights? What does that mean? When does it become significant harm? And that makes it difficult when you encounter concrete case. My guess is that there might be a bit more guidance on this. Also, for the fundamental rights impact assessments, so they might already be solving that to some extent." - Respondent F

"But I think the problem is it. I still think it remains a little bit vague. Because, for example, it says, uh. For example, the first question already is about the fundamental rights or the Union values. And for me it was a little bit difficult to determine. I'm supposed to look at for example" - Respondent L

7. Ethic experts tend to analyze highest potential risk

During the interviews, several respondents with expertise in ethics tended to classify almost all AI systems as Prohibited Risk. They placed significant emphasis on privacy protection and considered the potential foreseeable risks associated with AI systems. According to them, all AI systems could potentially be prohibited depending on their context and the purpose for which they are used.

"Potential Is that it can be used by like battlefield or if we are talking about the automated vehicles, there's a potential that the AI system could kill the person inside. Or maybe you you you made the drone that can kill someone. Yet that's that's the potential risk that you need to assess first." - Respondent J

"Privacy concern, obviously, that's actually perfect because they require as they acquire the data, so whether it is problem there, emotion or access to the email that's a bit based on faith" - Respondent M

These additional insights provide valuable perspectives on adopting an interdisciplinary approach, identifying intriguing cases, standardizing AI system formulation, evolving perceptions, understanding boundaries for developers, interpreting vague terms in regulations, and the focus on potential risks by ethical experts. Incorporating these insights can enhance the decision tree's effectiveness and the overall classification of AI systems.

8.2.4. Recommendation for Decision Tree Improvement

The following recommendations emerged from the interviews regarding the enhancement of the decision tree tool:

1. Add more case studies/examples and explanation

To address the difficulty in understanding legal terms and definitions, it is advisable to incorporate more concrete definitions and examples that illustrate the questions at each level of the decision tree. Respondent P expressed challenges in identifying the specific vulnerable groups mentioned in the AI Act, leading to assumptions when answering questions in the decision tree. Additionally, respondent P suggested providing summaries of laws and listing all possible use cases related to each specific question in the decision tree.

"Perhaps or more more concrete definitions? Or perhaps examples that that represent those questions you know". - Respondent N

"And I would suggest if you if you just in in the case of listing the vulnerable groups you with the list them all, or you say for example, so that people understand that this is these are not the only three groups of vulnerable people or something." - Respondent P

"You just need to add more text and maybe have some pointers to summary of the laws, but if it's not listing all possible cases use cases then it should be saying this clearly." - Respondent P

2. More direction/guideline

The current decision tree requires respondents to proceed through the questions step by step without clear indications of their progress or the specific assessment they are conducting. Respondents C and L suggested the inclusion of more high-level information to inform users whether they are assessing an Unacceptable Risk or High-Risk system.

"This could also be super helpful to have, like a high level information about what is in the directive, like for each directive and then see also the name of the directive if that helps you like if you." - Respondent C

"Like I will give more explanation about in this session in the decision tree. I mean in this session you need to assess the intention of this AI system." - Respondent L

8.3. Results Discussion

This section delves into a comprehensive discussion of the results obtained from both quantitative and qualitative analyses.

8.3.1. Classification Performance for Non-Obvious Cases without Decision Tree

For non-obvious cases, evaluating the decision tree's performance involves assessing inter-rater reliability to determine similarity agreement and measuring the time taken to classify AI systems.

To analyze the decision tree's effectiveness, we begin by evaluating the performance of classifying AI systems without the use of the decision tree. The results reveal that for non-obvious cases, the agreement between respondents is poor ($K\alpha = 0.231$). This score indicates that respondents have varying preferences when classifying the provided non-obvious cases. Consequently, a tool that can enhance classification agreement among respondents might be beneficial.

Additionally, we assess the time required to classify AI systems. The findings demonstrate that the time taken for classifying each case ranges from one to two minutes, representing a reasonably efficient process. However, interviews revealed that when classifying without the decision tree, respondents often relied on their assumptions and understanding of risk under the AI Act instead of closely engaging with the provided AI Act articles.

8.3.2. Decision Tree Performance for Non-Obvious Cases in General

When employing the decision tree, the inter-rater reliability performance decreases compared to classification without the decision tree. All respondents' agreement level drops to 0.086, indicating a low level of agreement. This trend holds even for non-legal respondents, where the score becomes negative. The use of the decision tree does not enhance the overall classification performance of AI systems.

Further analysis reveals that these divergent results are due to respondents' differing arguments regarding two distinct risk levels. For instance, in Case 5, respondents leaned towards classifying it as either No/Minimal Risk or Limited Risk, while in Case 6, they oscillated between Unacceptable Risk and High-Risk classifications.

Considering the time performance, upon evaluating the central tendency of the data while excluding outliers, we find that the time saved in classifying non-obvious cases with the decision tree amounts to 10 to 20 seconds—only a slight improvement. Notably, for Case 4 and Case 5, the classification duration is notably longer compared to the classification without the decision tree. This discrepancy may stem from respondents' struggles to grasp the context of certain cases, as observed in Case 6, where the concept of "significant chilling effects" proved challenging to comprehend.

8.3.3. Decision Tree Performance Considering Legal and Non-Legal Respondents

As depicted in Table 8.1, both legal and non-legal respondents exhibit slight agreement (Kα = 0.01 - 0.2) in the first approach (Without DT). However, this score diminishes even further when the decision tree is employed for classification.

Comparing legal and non-legal respondents, the similarity agreement between legal experts is higher than that among non-legal respondents. The use of the decision tree leads to a decline in agreement among non-legal respondents (Kα < 0.01). Conversely, although using the decision tree results in a decrease in Krippendorff's Alpha, it remains within the range of slight agreement (Kα = 0.01 - 0.2).

In summary, for legal respondents, the decision tree contributes to higher agreement in Case 5; however, for Case 3, 4, and 6, it does not improve agreement among legal respondents.

In contrast, for non-legal respondents, the decision tree does not significantly enhance agreement in the four cases (Cases 3, 4, 5, and 6). However, it does aid non-legal respondents in assessing whether AI systems cause significant harm to Fundamental Rights and/or Union values. It is notable that respondents encounter difficulties understanding terms such as 'significant risk,' 'vulnerable groups,' and 'interaction between human and machine,' leading to a lower level of agreement.

Most non-legal respondents encountered challenges in comprehending the legal terms employed in the decision tree questions, which resulted in diverse assumptions during the experiment. Conversely, legal respondents prioritized distinguishing whether an AI system posed a 'significant risk,' highlighting that the term itself is unclear within the AI Act.

The findings underscore that both legal and non-legal experts face challenges when classifying non-obvious cases, regardless of the decision tree's application. This finding emphasizes the need for enhanced frameworks and tools like the decision tree to clarify AI system classification and streamline the classification process. Furthermore, it highlights the importance of clarifying risks and harm within the AI Act itself, especially for specific AI system use cases.

During the evaluation, certain extreme cases revealed that individuals with more ethical understanding often perceived higher risks or potential harm from AI systems, leading to their classification predominantly as Unacceptable Risk. Addressing this requires refining the decision tree by offering explicit guidelines and boundaries for potential risks or harm associated with similar AI systems. This could include examples or precise definitions. Additionally, policymakers involved in AI Act formulation should prioritize clarifying risks and harm identified as Unacceptable Risk or High-Risk.

Ultimately, while the decision tree proves more effective among legal experts, insights from non-legal backgrounds underscore the need for tools to comprehend AI system risks, facilitating boundary establishment during the design phase. An interdisciplinary approach is vital, translating legal terms into practical language for non-legal backgrounds. This aligns with scholarly discussions advocating interdisciplinary approaches to regulating specific AI systems, such as emotion recognition systems [5].

Finally, as the AI Act promotes AI literacy among AI system providers and deployers, developing simplified tools for understanding the AI Act becomes imperative, not only for legal use but also for practical application. Such tools would enhance comprehension of AI systems and their associated risks.

8.3.4. Decision Tree Performance of Obvious and Non-Obvious Cases

The evaluation of decision tree performance encompassed both quantitative metrics and qualitative insights. Overall, when comparing decision tree performance for Obvious Cases (Chapter 7) and Non-Obvious Cases (Chapter 8), it becomes apparent that the similarity agreement and time performance are distinguishing factors, as accuracy could not be measured for Non-Obvious Cases.

Agreement in similarity between respondents' decision tree-based classifications for Obvious Cases reflects moderate agreement (Kα = 0.44), while for Non-Obvious Cases, it shows slight agreement (Kα = 0.086). This discrepancy suggests that the decision tree performs better in classifying Obvious Cases than Non-Obvious Cases.

In terms of time performance, excluding outliers, it is evident that classification for Obvious Cases is more efficient than for Non-Obvious Cases. The latter's classification durations for certain cases (Case 3 and Case 5) even exceed those without the decision tree. However, it is important to note that the time saved is less than a minute.

The moderate to poor agreement in Inter-rater Reliability with the use of the decision tree, along with limited improvement in time performance, raises several potential concerns regarding the decision tree.

Firstly, users may require additional context and understanding of certain terms, definitions, and examples to effectively use the decision tree. Although the decision tree incorporates certain AI Act definitions, they may not be sufficient to guide respondents comprehensively.

Secondly, it is noteworthy that even without the decision tree, the classification of AI systems alone yielded low performance.. This suggests a need for further research. Qualitative analysis indicated that legal experts often expressed concerns about vague definitions and terms, leading to assumptions during risk identification. For instance, the term "significant harm" generated divergent interpretations among respondents. Some perceived AI systems as causing significant societal harm, while others held the opposite view. As the decision tree adheres to the AI Act's terms and definitions to ensure its "representative" design, respondents might have mistakenly placed Limited-Risk systems into Unacceptable Risk or High-Risk categories. This is evident in Figure 7.6, where Limited-Risk systems were frequently classified as Unacceptable Risk or High-Risk, instead of the lower risk (No/Minimal Risk). Respondents' confusion over certain terms, including "significant harm," influenced their perception of potential risks and led to higher risk classifications. Consequently, when this classification system is applied among the public, misclassifications could result in fines. Therefore, clear guidelines or divisions are essential to help individuals determine the risk level associated with their AI systems.

Lastly, the formulation of AI system use cases requires additional context. Respondents indicated the need for more detailed information about AI systems' data usage, handling and processing, social context, outputs, technical aspects, responsibility distribution, purpose and intention, information storage, and user identity. Currently, provided use cases consist of brief one-sentence descriptions. However, incorporating more information into AI system descriptions could render decision tree evaluations unverifiable, as there would be no ground truth for validation, especially for Obvious Cases. It is understood that different contexts can yield distinct risk level outcomes.

From a qualitative perspective, the decision tree effectively embodies its design principles, such as clarity and representativeness. Respondents indicated that using the decision tree improved their understanding of the AI Act and provided access to relevant information. The decision tree's design aligned with respondents' mental models for classifying AI systems, meeting their expectations. Furthermore, respondents found it

easy to navigate through the decision tree, step by step, to determine AI system classifications. Overall, the decision tree received positive qualitative feedback.

Lastly, qualitative assessments highlighted that respondents found it easier to classify use cases falling under Obvious Cases compared to those categorized as Non-Obvious Cases, which proved more challenging to assess. Factors contributing to this difficulty included the employed terms and definitions and limited contextual information about the AI systems themselves.

8.4. Summary of the Chapter

In conclusion, this chapter provides an in-depth analysis of the results obtained from the classification of Non-Obvious Case AI systems. The findings from these analyses are presented both quantitatively and qualitatively. The quantitative analysis focuses on similarity agreement and time performance, while accuracy measurement was unfeasible due to the lack of ground truth. The qualitative analysis offers insights into the decision tree's benefits, associated concerns, additional insights, and recommendations for improvement.

In summary, the decision tree did not exhibit a significant improvement in categorizing and facilitating a clearer understanding of non-obvious cases of AI systems. However, qualitative analysis suggests that the decision tree framework aids respondents in classifying AI systems by providing a more visualized and structured approach to classification. It serves as a mental guide for AI system classification. Nonetheless, further research is needed to enhance reproducibility (similarity agreement) for non-obvious cases, as Krippendorff's Alpha indicated low agreement among respondents.

9

Conclusions

In this chapter, the research sub-questions are reflected in order to address the main research question (Section 9.1). Based on those sub-questions, the limitations, further research, and recommendation are presented each in Section 9.2, Section 9.4, and Section 9.3. A discussion of relevance of this research to CoSEM is also explained in Section 9.5. Lastly, Section 9.6 presents the academic contribution of this research.

9.1. Revisiting the Research Questions

9.1.1. Answering Sub-Question 1

What is the existing AI systems classification, and what are possible challenges to the current classification?.

AI systems governed by the AI Act are categorized based on risk levels, adopting a multi-layer risk-based approach. Each risk levels associated with specific obligations. The proposed AI Act introduced this categorization of AI systems according to four different risk levels: 1) Unacceptable Risk (Title II), 2) High-Risk (Title III), 3) Limited Risk (TItle IV), and 4) Minimal Risk (Title IX).

AI systems categorized as Unacceptable Risk are strictly prohibited for use within the EU, with a few exceptions. These prohibited AI systems pose unacceptable risks that contradict the Fundamental Rights and Union values. The EU's prohibition practices influence the use cases falling under the category of Unacceptable Risk. For instance, AI systems used for credit scoring, which is not permitted in the EU, would be prohibited.

High-Risk AI systems require compliance with a set of requirements before they can be introduced to the EU market. These AI systems have the potential to cause significant harm to the health, safety, and fundamental rights of individuals within the EU. There are two main categories of High-Risk AI systems: those intended as safety components of products subject to third-party ex-ante conformity assessment, and standalone AI systems with significant implications for fundamental rights. AI systems falling under these categories must adhere to mandatory requirements, such as risk management measures, high-quality data and governance practices, traceability, transparency, human oversight, and robustness and cybersecurity measures.

Limited Risk is the third level of risk in the AI Act, which entails transparency obligations. For AI systems classified as Limited Risk, such as chatbots, users should be informed that they are interacting with a machine.

No/Minimal Risk falls outside the scope of regulation, but its significance should not be overlooked, as the majority of AI systems in the EU are categorized as No/Minimal Risk. While no mandatory obligations apply to AI systems falling under this category, the adoption of voluntary codes of conduct is encouraged.

While the risk classification framework provides benefits in safeguarding individuals, challenges persist within the current classification of AI systems. One such challenge is the lack of clarity in classifying certain AI system use cases. The absence of clear boundaries between risk levels can lead to issues related to budget allocation, innovation, and potential penalties for misclassification under the AI Act. Several factors contribute to this lack of clarity, including the narrow scope of Unacceptable Risk, exceptions and loopholes in prohibited AI practices, vague language in specific cases, the distinction between high-risk and limited-risk categories, reliance on interpretation for risk classification, the need for a more interdisciplinary approach, and the challenges arising from binary thinking.

To address these clarity issues, this study identifies Non-Obvious Cases that fall within the borderline between risk classes. These cases, along with Obvious Cases, are examined to evaluate the effectiveness of the AI Act framework in accurately classifying AI system use cases into specific risk levels.

9.1.2. Answering Sub-Question 2

What are features that differentiate each level of the AI Act classification?

Through the analysis of the output from SQ1, the existing AI systems classification in the AI Act, the identified challenges, and the abstraction process along with discussions with legal experts, four key features have been identified as variables for the decision tree used in classifying each risk level. These features are P**rotected Values, Objective/Intended Purpose, Domain**, and **Technology/Use Case**.

The first feature, **Protected Values**, plays a significant role in distinguishing AI systems based on their potential harm to individuals. It divides the AI systems into two main groups: Unacceptable Risk and High-Risk, and Limited Risk and No/Minimal Risk. The AI Act highlights the potential for significant harm in the first group, while the second group poses less or no significant harm to individuals.

The second feature, **Objective/Intended Purpose**, allows for the assessment of an AI system's purpose and helps determine its risk level. For example, AI systems that exploit vulnerabilities in vulnerable groups are classified as Unacceptable Risk, while those generating artificial content are classified as Limited Risk. The current approach assesses the intention based on the foreseeable purpose of the AI system.

The third feature is **Domain**, which helps differentiate between High-Risk and non-High-Risk AI systems. Understanding the domain in which an AI system operates aids in identifying its appropriate risk level.

The fourth feature, **Technology/Use Case,** considers the specific technologies mentioned separately in Unacceptable Risk, High-Risk, and Limited Risk. Identifying the intersection of technology and use case helps in understanding and classifying AI systems into their respective risk levels.

These four features serve as variables for the proposed framework. Additionally, high-level design principles have been formulated to ensure the comprehensive representation of these features. These design principles, including c**larity, simplicity, obligation-required, sequential, representative,** and **value first**, provide guidelines for generating the decision tree used in the classification of AI systems.

9.1.3. Answering Sub-Question 3

What possible framework can be designed to improve the classification process of AI systems?

The proposed solution to address the challenges in AI systems classification and improve the classification process is a decision tree framework. This framework is designed based on the identified features: Protected Value, Objective/Intention, Use-Case/Technology, and Domain. The decision tree is constructed according to these features, as illustrated in Figure 9.1.

In the development of the decision tree framework, the application of design principles and an iterative process are essential. Design principles provide guidance on the functional and visual aspects of the decision tree. For example, the decision tree is designed to provide output that includes the associated risk level, value at risk, obligations, and exemptions for certain AI systems. Design principles also set boundaries for the framework's development. To ensure simplicity, the number of questions in the decision tree is kept manageable, preventing users from being overwhelmed and enabling them to complete the assessment effectively.

The iterative process plays a critical role in designing the decision tree. Multiple iterations are conducted, involving experts' input and insights, to refine and improve the framework. Through these iterations, it was determined that a maximum of 20 questions is appropriate to maintain simplicity while still capturing the essence of the AI Act.

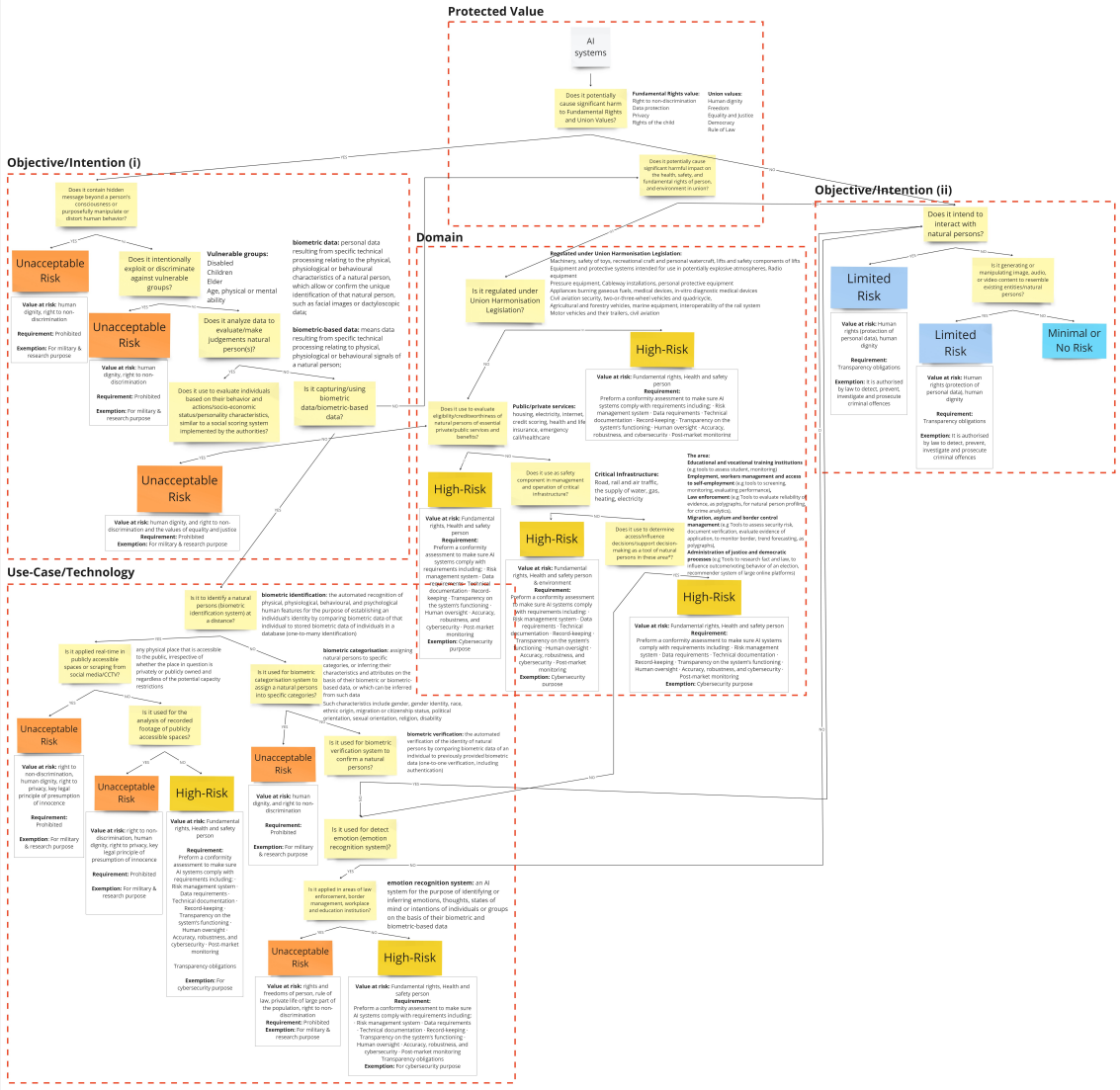

Figure 9.1: Decision Tree Overview

Overall, the decision tree framework is the result of a thoughtful and iterative design process, incorporating design principles and expert input to enhance the classification process of AI systems.

9.1.4. Answering Sub-Question 4

How to evaluate the proposed framework and what improvements can be drawn from the evaluation?

To assess the effectiveness of the proposed framework, a comprehensive evaluation was conducted involving 16 participants from both legal and non-legal backgrounds. The evaluation consisted of three sections: two experimental sessions and a semi-structured interview.

In the first two sessions of the experiment, participants were tasked with classifying various AI system use cases into the four categories based on the AI Act. In the first session, they classified the AI systems without the aid of the decision tree, relying on their interpretation of the AI Act. In the second session, they utilized the proposed decision tree to classify the AI systems. The third section consisted of a semi-structured interview, which aimed to explore the participants' opinions on the decision tree and their understanding of the classification process.

Both quantitative and qualitative data were collected and analyzed. The quantitative data focused on comparing the classification results between the two experimental sessions, measuring inter-rater reliability (sim-

ilarity agreement) and efficiency (time performance). Specifically for Obvious Cases, the accuracy, precision, recall, and F1-score can be measured since there are ground truth from the AI Act. On the other hand, the qualitative data were analyzed using thematic coding in software such as Atlas.ti to gain insights into the decision tree and the classification process.

When comparing the effectiveness of the decision tree in classifying Obvious Cases (discussed in Chapter 7) with its performance in handling Non-Obvious Cases (discussed in Chapter 8), two distinct factors emerge as key differentiators: similarity agreement and time efficiency. It is important to note that due to the absence of ground truth, measuring accuracy was not feasible for Non-Obvious Cases.

In terms of similarity agreement, comparing the decision tree-based classifications for Obvious Cases and Non-Obvious Cases reveals noteworthy differences. Specifically, the agreement among respondents' classifications using the decision tree for Obvious Cases demonstrates a moderate level of agreement (Kα = 0.44), while for Non-Obvious Cases, the level of agreement is only slight (Kα = 0.086). This notable variation strongly suggests that the decision tree is more effective in classifying Obvious Cases than Non-Obvious Cases.

Regarding the time performance, the classification process for Obvious Cases is notably more streamlined than that for Non-Obvious Cases. Interestingly, for specific instances within Non-Obvious Cases (such as Case 3 and Case 5), the time required for classification exceeds that of scenarios where the decision tree is not employed. However, it is important to highlight that the time saved through decision tree utilization remains relatively modest, amounting to less than a minute.

Based on the evaluation, several improvements were identified. It is suggested to incorporate more interdisciplinary approaches to translate legal terminology into more understandable language for non-legal participants. Additionally, providing more context to the AI system classification and including clear definitions, instructions, and real-life examples would enable participants to better assess the risk level of AI systems.

9.1.5. Answering Main Research Question

To what extent can the process of AI systems classification under the AI Act be improved?

Generally, the use of the decision tree shows slight improvement in classifying AI systems, in terms of accuracy, reproducibility, and time-efficiency. However, based on the evaluation, many insights can be gathered to improve the classification of the AI systems within the AI Act.

Regarding the decision tree's performance for both obvious and non-obvious use cases, it became evident that the decision tree faced more significant challenges when classifying non-obvious cases compared to obvious ones, particularly in terms of reliability and time efficiency. The lack of clarity within terms and definitions and limited contextual information posed notable difficulties in classifying non-obvious cases.

The performance of the decision tree also exhibited variations between legal and non-legal respondents. Legal experts demonstrated higher similarity agreement (reproducibility) and efficiency than their non-legal counterparts. This outcome suggested their familiarity with legal terminology and the nuances of the AI Act. However, it is noteworthy that both legal and non-legal respondents encountered challenges in classifying non-obvious cases, underscoring the need for enhancing decision tree frameworks or even reconsidering the creation of more effective tools to enhance clarity and streamline the classification process.

Furthermore, an essential observation emerged during the study – the lack of clarity within the decision tree could be attributed to the lack of clarity within the AI Act's articles, as highlighted by several respondents. This underscores the imperative for greater clarity within the terms mentioned in the AI Act, prompting a recommendation for an interdisciplinary approach to tackle these challenges and facilitate a comprehensive understanding of AI system risks.

The process of classifying AI systems under the AI Act can undergo significant improvement by implementing clear and comprehensive guidelines, structured decision-making frameworks, and fostering enhanced collaboration between legal and technical experts. To illustrate, developing a detailed decision tree that integrates specific criteria, definitions, and illustrative examples for various risk levels could provide a systematic method for classifying AI systems.

Such a decision tree would support both legal and non-legal professionals in accurately evaluating the risk class of AI systems and assigning them to appropriate regulatory requirements. Additionally, incorporating

practical case studies and hypothetical scenarios could effectively demonstrate the practical application of the AI Act's provisions in real-world scenarios, thus facilitating a more comprehensive understanding and fostering consistent interpretation among stakeholders.

Moreover, creating a platform or environment that encourages continuous feedback and active discussion among legal experts, AI practitioners, policymakers, and other relevant stakeholders would foster a dynamic and iterative classification process. This collaborative approach would aid in identifying ambiguities, addressing emerging challenges, and refining classification criteria over time. Ultimately, it would pave the way for a more refined and effective process of AI system classification under the AI Act, particularly when enforced to the broader public.

9.2. Limitations

While this study yields valuable insights into classifying AI systems based on risk using a decision tree approach, it is important to acknowledge certain limitations.

Firstly, the sample size of participants in this study might impact the generalizability of the findings. The participants, especially those with legal backgrounds, may only partially represent part of the population, potentially influencing the validity and reliability of the results. Future studies should aim for more extensive and diverse participant samples to bolster the robustness of the findings.

To enhance participant representability, an open survey approach could be implemented. This approach involves distributing the survey/questionnaires to legal and non-legal experts to gather a large number of participants. The methodology could resemble that used in this study, with participants classifying AI system use cases under two scenarios: one using only the AI Act and the other incorporating the decision tree framework. The responses from all participants would be gathered, measured, and analyzed. Several respondents could also be invited for more in-depth insights for semi-structured interviews. By employing this method, the participant sample size could be expanded.

Moreover, the number of use cases examined in this study may limit the comprehensiveness of the classification framework. The chosen use cases may not fully encompass the breadth and complexity of AI applications, possibly leading to biases or incomplete risk assessments. Future research should encompass a broader range of use cases, spanning various domains and application scenarios, to ensure a comprehensive and accurate classification framework.

To curate more relevant and compelling use cases, engaging in discussions with legal and non-legal experts and even involving business perspectives could be beneficial. This approach could yield more comprehensive use cases that truly capture the multifaceted nature of AI applications.

Additionally, through these discussions with diverse experts, the provided use cases for classification could be better described than the simple one-sentence approach utilized in this study. A concise sentence to define the AI systems may mislead participants and lead to differing assumptions. Offering greater context, technical detail, and legal nuances of the AI systems could significantly enhance the classification.

Furthermore, the evolving nature of regulations and legal frameworks related to AI systems could challenge the applicability of the decision tree approach. Given that the AI Act is currently undergoing discussions by the EU Commission, the decision tree might need updates to align with the latest developments. Acknowledging the dynamic regulatory landscape and ensuring that the decision tree model remains adaptable and relevant over time is essential.

Lastly, the potential for personal bias in developing the decision tree is worth considering. To mitigate this, engaging in more extensive discussions with legal and technical experts could help create a more comprehensive and impartial decision tree framework.

9.3. Further Research

This study opens up several avenues for further research in the field of AI system classification. The recommendation for further research can be divided into two big categories: (1) Research related to the standardization framework to classify AI systems based on the AI Act, (2) Research to improve the clarity of the risk classification of the AI Act.

1. Research related to the standardization framework to classify AI systems based on the AI Act

Firstly, the continuity of the decision tree's performance and its evaluation over time should be explored. As AI technologies and applications evolve, it is crucial to assess the effectiveness and relevance of the decision tree framework.

Additionally, future research should consider conducting more targeted studies in specific domains or industries. Researchers can delve deeper into industry-specific risk factors and challenges by focusing on particular sectors, allowing for more tailored and effective classification frameworks. Furthermore, exploring the perspectives and classification practices of legal professionals, organizations, and relevant stakeholders within these specific domains can provide valuable insights for refining the decision tree model and accommodating industry-specific considerations.

Quantitative research can also be conducted to evaluate the decision tree's performance using larger datasets with more industrial backgrounds. This can improve the performance of the classification framework. Additionally, incorporating qualitative research methods, such as interviews or focus groups, can provide rich insights into the decision-making processes and considerations of stakeholders involved in AI system classification.

Based on the analysis, several areas for improvement in AI systems classification under the AI Act have been identified. The current classification process faces challenges related to ambiguities in definitions, lack of contextual information, and difficulties in distinguishing between different risk levels. Addressing these challenges and introducing clearer guidelines to improve the process is recommended. One approach is to refine the decision tree used for classification, incorporating additional criteria and features that provide more clarity and context.

2. Research to improve the clarity of the risk classification of the AI Act

It is also identified that for non-obvious cases, even legal experts have difficulties in classifying them. Therefore, further research to analyze the reason for this difficulty might be crucial to improve the clarity of the AI systems classification under the AI Act. For example, specific research to find mutual understanding in an interdisciplinary approach to defining 'significant harm', 'vulnerable groups', and type of 'interaction' between human and machine, as those terms contributed to the low performance of the classification framework in this study.

Furthermore, the research can also focus on the different perspectives of legal experts towards certain High-Risk and Unacceptable Risk, and High-Risk and Limited Risk; since from the research, it is known that legal experts themselves have different understandings to determine which risk level is associated with certain AI systems. This research can help further standardize the classification of AI systems.

Another thing to consider is the formulated context of AI systems that can be used to analyze the AI systems. This could involve including more detailed information about the AI system's data consumption, data use, data handling, and data processing, social context, the output of the system, technical aspects, responsibility distribution, purpose and intention, information storage, and type of user who will use the AI systems. By providing a standard of this formulated context of AI systems, one can prepare what they need to know to classify their AI systems better.

Furthermore, an interdisciplinary approach is necessary to translate legal terms and concepts into more accessible language for non-legal backgrounds. This would ensure that individuals from various domains can effectively participate in the classification process and contribute their expertise.

9.4. Recommendation for Policy-making

For policymakers, gaining a comprehensive understanding of people's perceptions and achieving consensus in classifying AI systems, particularly in more complex cases, are important. The potential consequences of misclassification can lead to divergent interpretations and inconsistent categorizations by different stakeholders, ultimately undermining the effectiveness of regulatory efforts. To counter this, policymakers must prioritize the establishment of clear definitions and guidelines for the terms and concepts stipulated in the AI Act. By providing unambiguous explanations for terms like "significant harm," "vulnerable groups," and "interaction between human and machine," policymakers can effectively equip legal experts and AI practicioners with the necessary tools to make informed and consistent classifications.

Furthermore, policymakers need to recognize the significance of contextual considerations when evaluating AI systems for classification. The context within which an AI system operates plays a pivotal role in determining the variables that require assessment. Policymakers can guide users and organizations in classifying AI systems by offering a comprehensive framework that accounts for various factors such as data usage, potential societal impact, technical specifications, and intended purposes. This context-rich approach enables stakeholders to make accurate judgments about which risk level best aligns with their AI systems' characteristics and functionalities. Consequently, a well-structured and contextually informed framework facilitates more precise classification outcomes, reducing the likelihood of misinterpretations and inconsistent applications of AI regulations.

For instance, the case of AI systems designed for social robots for children with autism to capture their behavior to assist treatment. Without clear guidelines and definitions, different stakeholders might interpret the terms "significant harm" and "vulnerable groups" differently. Some might argue that misdiagnosis by the AI system could lead to significant harm to patients, while others contend that the system's intent to assist patients mitigates such harm. Additionally, the question of whether patients with pre-existing conditions should be considered vulnerable groups could lead to varying classifications.

However, if policymakers provide detailed definitions and guidelines, such as specifying that "significant harm" refers to life-threatening consequences or irreversible damage and that "vulnerable groups" encompass individuals with compromised health conditions, stakeholders can make more consistent and accurate classifications. Moreover, contextual considerations are crucial. In this case, the context involves the medical field, where the importance of minimizing misdiagnoses and protecting patients' health is essential. Therefore, with clear definitions and contextual understanding, stakeholders can confidently classify this AI system as falling within a specific risk level, aligning with the intended regulatory goals.

These recommendations could be incorporated into a regulatory sandbox, regulated within the AI Act. The recent amendment to the proposed AI Act introduces a regulatory sandbox as a controlled environment for safe development, testing, and validation of innovative AI systems [11]. The sandbox provides a space for businesses and regulators to collaboratively develop and regulate technologies.

Hence, based on this study, it is highly recommended that the regulatory sandbox not only facilitates AI technology development but also allows for continuous improvement and evaluation of AI systems classification within the AI Act. Regular evaluations within the sandbox involving stakeholders from business, legal, and technical backgrounds could ensure consistent and accurate classification. An interdisciplinary approach in the sandbox can enhance the clarity and accuracy of the AI Act, benefiting organizations and AI engineers alike.

9.5. Relevance to CoSEM

CoSEM master thesis projects aim to design solutions for large and complex contemporary sociotechnical problems which requires the consideration of technical, institutional, economic, and social knowledge. Therefore, this thesis research is highly relevant to the CoSEM study since this research involves understanding from legal, technical, and institutional approaches.

Moreover, this study also develops a decision tree framework to improve the accuracy of the AI systems classification. This decision tree refers to an artifact developed systematically and creatively, as presented during the stages of research, from understanding the context and challenges, defining features, developing the framework, and evaluating the framework. A systematic process is conducted, such as within the literature review, even in gathering variables to generate a decision tree. However, creative ways are also involved, such as generating the decision tree and stimulating the evaluation process to engage respondents during the interview session.

Tools and techniques used in this research are introduced during the lecture, such as Design Science Methodology. A step-by-step approach is conducted to answer the main research question by formulating the subquestions based on each stage according to Design Science Methodology.

Moreover, as mentioned before, this research also requires insights from legal and non-legal respondents, which presents the need for an interdisciplinary approach to solving the clarity issue of AI systems classification identified in the research.

The data collected, and the framework developed in this study can be utilized by organizations implementing AI systems to evaluate and understand the risks associated with their systems. This contributes to the overall management of complex systems within the CoSEM context. Furthermore, the findings of this study indirectly benefit the public by increasing transparency and understanding of the classification of AI systems. Organizations can ensure responsible and ethical implementation by evaluating and categorizing AI systems based on risk, thereby addressing societal concerns and fostering trust in AI technologies.

Moreover, this study provides policymakers with valuable insights and recommendations for a more accommodating and effective classification of AI systems. The interdisciplinary nature of CoSEM aligns with the complex and multifaceted challenges surrounding AI system classification. Integrating technological considerations, human perspectives, and legal and regulatory frameworks is vital in shaping policies that balance innovation, societal benefits, and risk mitigation. Therefore, this study contributes to the CoSEM field by bridging the gap between technology, humans, and institutions in the context of AI system classification.

9.6. Academic Contribution

The research gap related to the AI Act mentioned in Chapter 2 is the need for clear distinctions between the classification criteria of the AI systems, including the emerging use cases not yet addressed by the AI Act. Therefore, in this study, the decision tree framework is proposed with the assumption that this decision tree framework would lead to the improvement of AI systems classification, according to the AI Act. In doing so, this study makes several key contributions to the existing knowledge gaps:

First, the study introduces a decision tree framework that systematically outlines the decision-making process for classifying AI systems based on the AI Act. This framework considers the explicit use case within the AI Act and extends its scope to encompass more complex use cases that the current legislation may not explicitly cover.

Second, by delineating clear and specific criteria for differentiating between AI system risk levels, the proposed framework directly addresses the identified need for more distinct classification criteria. This reduces ambiguity and promotes a consistent approach to AI system classification.

Third, this study evaluates the performance of a decision tree in classifying AI systems based on the AI Act. By examining the accuracy, reliability (reproducibility), and efficiency of the decision tree in distinguishing between different risk classes, this study provides evidence of the utility of the decision tree for this purpose. The findings shed light on the strengths and weaknesses of the decision tree approach and offer insights into its practical application in real-world scenarios.

Fourth, this study also focuses on the context of non-obvious cases (borderline cases) that emerge from the AI systems classification based on the AI Act. Along with obvious cases, the distinction between non-obvious cases and obvious cases is crucial, as non-obvious cases often involve more complex risk assessments. Therefore, studies in particular non-obvious cases emerging from the AI systems classification will contribute to academic research.

Fifth, the proposed decision tree framework encourages understanding between legal and non-legal experts, promoting an interdisciplinary approach. This collaboration is crucial for developing comprehensive guidelines that consider legal, technical, and ethical aspects of AI system classification. Moreover, the qualitative analysis from this research offers insight that might be beneficial to developing a more standardized framework to help the classification process even more insights to enhance the AI Act.

Lastly, organizations that develop and deploy AI systems can benefit from the decision tree framework by understanding the risks associated with their systems and taking appropriate mitigation measures, especially for small-medium enterprises. The decision tree can serve as a valuable tool for software developers and engineers to assess and manage the risks settled within their AI systems.

Overall, the academic contribution of this study lies in the design and development of the proposed decision tree framework that fills the existing gaps within the AI systems classification framework in the AI Act. The framework addresses the need for clear distinctions and evaluates the decision tree performance in classifying obvious and non-obvious cases. This contribution enriches the academic discourse and informs policymakers, researchers, and practitioners about a potential solution to enhance the AI system classification process in a dynamic and rapidly evolving AI technology.

References

[1] Regulatory framework proposal on artificial intelligence. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Accessed: 2023-08-08.

[2] Hosam Al-Samarraie, Hassan Selim, and Fahed Zaqout. The effect of content representation design principles on users' intuitive beliefs and use of e-learning systems. *Interactive Learning Environments*, 24:1758–1777, 11 2016. ISSN 17445191. doi: 10.1080/10494820.2015.1057739.

[3] Hamza Alshenqeeti. Interviewing as a data collection method: A critical review. *English Linguistics Research*, 3(1):39–45, 2014. ISSN 1927-6028. doi: 10.5430/elr.v3n1p39.

[4] AppliedAI. Ai act: Risk classification of ai systems from a practical perspective, 3 2023.

[5] I. Barkane. Questioning the eu proposal for an artificial intelligence act: The need for prohibitions and a stricter approach to biometric surveillance. *Information Polity*, 27(2):147–162, 2022. doi: 10.3233/IP-211524.

[6] Johanna Chamberlain. The Risk-Based Approach of the European Union's Proposed Artificial Intelligence Regulation: Some Comments from a Tort Law Perspective. *European Journal of Risk Regulation*, 14(1):1–13, 2023. ISSN 21908249. doi: 10.1017/err.2022.38.

[7] Leona Chandra, Stefan Seidel, and Shirley Gregor. Prescriptive knowledge in is research: Conceptualizing design principles in terms of materiality, action, and boundary conditions. volume 2015-March, pages 4039–4048. IEEE Computer Society, 3 2015. ISBN 9781479973675. doi: 10.1109/HICSS.2015.485.

[8] European Commission. White paper - on artificial intelligence - a european approach to excellence and trust. 2020. URL https://www.cambridge.org/core/product/identifier/CBO9781107415324A009/type/book_part.

[9] European Commission. Proposal for a regulation laying down harmonised rules on artificial intelligence. 2021.

[10] European Commission. Annex - proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2021.

[11] European Commission. Draft compromise amendments on the draft report: Proposal for a regulation of the european parliament and of the council and harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2023.

[12] Jérôme De Cooman. Humpty Dumpty and High-Risk AI Systems: The Ratione Materiae Dimension of the Proposal for an EU Artificial Intelligence Act. *Market and Competition Law Review*, 6(1):49–88, 2022. ISSN 21840008. doi: 10.34632/mclawreview.2022.11304.

[13] Natasha R. Donnolley, Georgina M. Chambers, Kerryn A. Butler-Henderson, Michael G. Chapman, and Elizabeth Sullivan. A validation study of the australian maternity care classification system. *Women and Birth*, 32:204–212, 6 2019. ISSN 18781799. doi: 10.1016/j.wombi.2018.08.161.

[14] Lilian Edwards. The EU AI Act: a summary of its significance and scope. (April 2022): 1–26, 2022. URL https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf.

[15] Volpicelli Gian. Chatgpt broke the eu plan to regulate ai. https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/, 2023.

[16] L. M. Given. *The Sage Encyclopedia of Qualitative Research Methods*. Sage Publications, 2008.

[17] Philipp Hacker. A legal framework for ai training data—from first principles to the artificial intelligence act. *Law, Innovation and Technology*, 13(2):257–301, 2021. ISSN 1757997X. doi: 10.1080/17579961.2021.1977219. URL https://doi.org/10.1080/17579961.2021.1977219.

[18] I. Hupont, S. Tolan, H. Gunes, and E. Gómez. The landscape of facial processing applications in the context of the european ai act and the development of trustworthy systems. *Scientific Reports*, 12(1), 2022. doi: 10.1038/s41598-022-14981-6.

[19] Paul Johannesson and Erik Perjons. *An Introduction to Design Science*. Springer, second edi edition, 2021. ISBN 9783030781316. doi: 10.1142/9789811245473_0010.

[20] Yojna Khandelwal and Ritu Bhargava. Spam filtering using ai, 2021.

[21] Peter Kieseberg, Christina Buttinger, Laura Kaltenbrunner, Marlies Temper, and Simon Tjoa. Security considerations for the procurement and acquisition of artificial intelligence (ai) systems. *IEEE International Conference on Fuzzy Systems*, 2022-July, 2022. ISSN 10987584. doi: 10.1109/FUZZ-IEEE55066.2022.9882675.

[22] Mauritz Kop. Eu artificial intelligence act: The european approach to ai. *Transatlantic Antitrust and IPR Developments*, 2021. URL https://law.stanford.edu/publications/eu-.

[23] C.R. Kothari. *Research Methodology: Methods and Techniques*. New Age International Publishers, 2nd edition edition, 2004.

[24] Tarun Lalwani, Shashank Bhalotia, Ashish Pal, Shreya Bisen, and Vasundhara Rathod. Implementation of a chat bot system using ai and nlp. *International Journal of Innovative Research in Computer Science Technology*, 6:26–30, 5 2018. doi: 10.21276/ijircst.2018.6.3.2.

[25] Andreas Liebl and Till Klein. AI Act: Risk Classification of AI Systems from a Practical Perspective. Technical report, Initiative for Applied Artificial Intelligence, March 2023.

[26] E Lim, H Park, and B Kim. Review of the validity and rationality of artificial intelligence regulation: Application of the eu's ai regulation bill to accidents caused by artificial intelligence. 2022. doi: 10.32473/FLAIRS.v35i.130713.

[27] Britt Marie Lindgren, Berit Lundman, and Ulla H. Graneheim. Abstraction and interpretation during the qualitative content analysis process. *International Journal of Nursing Studies*, 108, 8 2020. ISSN 00207489. doi: 10.1016/j.ijnurstu.2020.103632.

[28] P. Marano and S. Li. Regulating robo-advisors in insurance distribution: Lessons from the insurance distribution directive and the ai act. *Risks*, 11(1), 2023. doi: 10.3390/risks11010012.

[29] Kapil Mittal, Dinesh Khanduja, and Puran Chandra Tewari. An insight into "decision tree analysis". *World Wide Journal of Multidisciplinary Research and Development*, 3:111–115, 2017. ISSN 2454-6615. URL www.wwjmrd.com.

[30] J. Mökander, M. Axente, F. Casolari, and L. Floridi. Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed european ai regulation. *Minds and Machines*, 32(2):241–268, 2022. doi: 10.1007/s11023-021-09577-4.

[31] Frederik Moller, Tobias Moritz Guggenberger, and Boris Otto. Towards a method for design principle development in information system. In Sara Hofmann, Oliver Muller, and Matti Rossi, editors, *Designing for Digital Transformation: Co-Creating Services with Citizens Industry*, pages 208–220. Springer, 12 2020.

[32] Benjamin Mueller. Artificial Intelligence Act Cost Europe ? (July), 2021.

[33] Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. An introduction to decision tree modeling, 6 2004. ISSN 08869383.

[34] Rostam J. Neuwirth. Prohibited artificial intelligence practices in the proposed eu artificial intelligence act (aia). *Computer Law and Security Review*, 48:105798, 2023. ISSN 02673649. doi: 10.1016/j.clsr.2023.105798. URL https://doi.org/10.1016/j.clsr.2023.105798.

[35] Alberto Orlando. Ai for sport in the eu legal framework. *2022 IEEE International Workshop on Sport, Technology and Research, STAR 2022 - Proceedings*, pages 100–105, 2022. doi: 10.1109/STAR53492.2022.9860029.

[36] Sarra Ouni, Fethi Fkih, and Mohamed Nazih Omri. Toward a new approach to author profiling based on the extraction of statistical features. *Social Network Analysis and Mining*, 11, 12 2021. ISSN 18695469. doi: 10.1007/s13278-021-00768-6.

[37] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24:45–77, 12 2007. ISSN 07421222. doi: 10.2753/MIS0742-1222240302.

[38] Raminta Pranckutė. Web of science (wos) and scopus: The titans of bibliographic information in today's academic world, 3 2021. ISSN 23046775.

[39] PwC. European Artificial Intelligence Act: many procedural and substantive requirements, year = 2022, url = https://www.pwc.nl/en/insights-and-publications/themes/digitalization/european-artificial-intelligence-act-many-procedural-and-substantive-requirements.html, urldate = 2023-06-22.

[40] Quinlan. Induction of decision trees, 1986.

[41] Bizhan Shabankhani, Jamshid Yazdani Charati, Keihan Shabankhani, and Saeid Kaviani Cherati. Survey of agreement between raters for nominal data using krippendorff's alpha, 2020.

[42] Olga Shumilo and Tanel Kerikmäe. The european approach to building ai policy and governance: a haven for bureaucrats or innovators? *Revista de Internet, Derecho y Politica*, 34(34):1–14, 2021. ISSN 16998154. doi: 10.7238/idp.v0i34.387744.

[43] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. A survey on methods and metrics for the assessment of explainability under the proposed ai act. *Frontiers in Artificial Intelligence and Applications*, 346:235–242, 2021. ISSN 09226389. doi: 10.3233/FAIA210342.

[44] A.Zh. Stepanyan. European Artificial Intelligence Act: Should Russia Implement the Same? *Kutafin Law Review*, 8(3):403–422, 2021. doi: 10.17803/2313-5395.2021.3.17.403-422.

[45] Kees Stuurman and Eric Lachaud. Regulating AI. A label to complete the proposed Act on Artificial Intelligence. *Computer Law and Security Review*, 44:105657, 2022. ISSN 02673649. doi: 10.1016/j.clsr.2022.105657. URL `https://doi.org/10.1016/j.clsr.2022.105657`.

[46] Gijs van Dijck. Predicting recidivism risk meets ai act. *European Journal on Criminal Policy and Research*, 28(3):407–423, 2022. ISSN 15729869. doi: 10.1007/s10610-022-09516-8. URL `https://doi.org/10.1007/s10610-022-09516-8`.

[47] Ida Varošanec. On the path to the future: mapping the notion of transparency in the EU regulatory framework for AI. *International Review of Law, Computers and Technology*, 36(2):95–117, 2022. ISSN 13646885. doi: 10.1080/13600869.2022.2060471. URL `https://doi.org/10.1080/13600869.2022.2060471`.

[48] P Verschuren and H Doorewaard. *Designing a Research Project (2nd ed.)*. Eleven International Publishing, 2010.

[49] R Wasko. Data sourcing: Pros and cons of desk research. `https://en.predictivesolutions.pl/data-sourcing-pros-and-cons-of-desk-research,en`, 2019.

[50] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16, 8 2016. ISSN 14712288. doi: 10.1186/s12874-016-0200-9.

Appendices

A

Interview Setup

A.1. Consent Form

Opening Statement

You are being invited to participate in a research study titled Enhancing AI Systems Classification Framework: A Study in the EU's Proposed AI Act. This study is being done by Hilmy Hanif from TU Delft, Complex Systems Engineering Management as part of Master thesis project.

The purpose of this research study is to enhance the AI systems classification framework under AI Act to be able to accommodate potential AI systems use cases. It will take approximately 60-75 minutes to complete. The data will be used for Master's thesis and/or academic publications. We will give you 16 use cases where you have to classify those use cases into several classifications. You have to determine only ONE classification of AI systems for each use case.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by performing this experiment without sensitive data is collected. The experimental study will be recorded, and the data will be stored for the analysis purpose in a TUD institutional storage, accessible only to Hilmy Hanif and Yury Zhauniarovich. The raw data will be archived for up 2 years after the end of the project (estimated September 2023) so it can be used for future research and learning on the topic of AI classification in EU AI Act. Should these results be added in additional publication, you will remain anonymous in those as well. After this period, the data will be deleted.

Your response/views/other input can be quoted anonymously in research outputs. The data analysis will result in an anonymous summary of our conversation. The summary, as well as your input on classification of AI system will be included in the MSc thesis will be made publicly available.

Your participation in this study is entirely voluntary, and you can withdraw at any time. For any further inquiries, please refer to Hilmy Hanif (hilmyhanif@student.tudelft.nl). If you have any questions regarding your personal data after the research, contact Yury Zhauniarovich (y.zhauniarovich@tudelft.nl).

A.2. Interview Protocol

Instruction:

1. This experiment is expected to take approximately 60-75 minutes to complete and should be followed in the following order:
   (a) Introduction (3-5 minutes)
   (b) Phase 1: Explanation (2,5 minutes) + Experiment without Proposed Framework (20 minutes)
   (c) Phase 2: Explanation (2,5 minutes) + Experiment with Proposed Framework (20 minutes)
   (d) Phase 3: Follow-up Interview (10-15 minutes)
2. The interviewer will provide an overview of the research goal and explain the tasks to the respondent.
3. The experiment consists of three sections:
   (a) Section 1: First experiment (using the framework given by the AI Act),
   (b) Section 2: 2nd experiment (using a decision tree (proposed framework) provided by the interviewer),
   (c) Section 3: Semi-structured interview (follow-up questions based on the previous sections)
4. In Section 1, each respondent will be presented with 8 use cases of AI systems. The respondent's task is to classify these AI systems into the four categories specified in the AI Act. To classify the AI systems, the respondent will be provided with references to Title II, Title III, Title IV, and Title IX of the EU AI Act.
5. In Section 2, each respondent will be given another set of 8 use cases of AI systems. The task remains the same: to classify these AI systems into the four categories mentioned in the AI Act. However, for this section, respondents are required to use a decision tree framework to determine the categorization of the AI systems.
6. In Section 3, the interviewer will ask several questions related to the results of Sections 1 and 2, or ask follow-up questions based on the respondent's choices.

A.3. Interview Board on Miro

Below are the visualization of the interview session in Miro. Figure A.1 is the visualization of the first session of the interview where the interviewees have to classify given use-cases with their understanding of the AI Act Article. The AI Act article is in Appendix B. Meanwhile, Figure A.2 is the visualization of the second session of the interview where the interviewees have to classify the remaining use-cases with the decision tree framework.

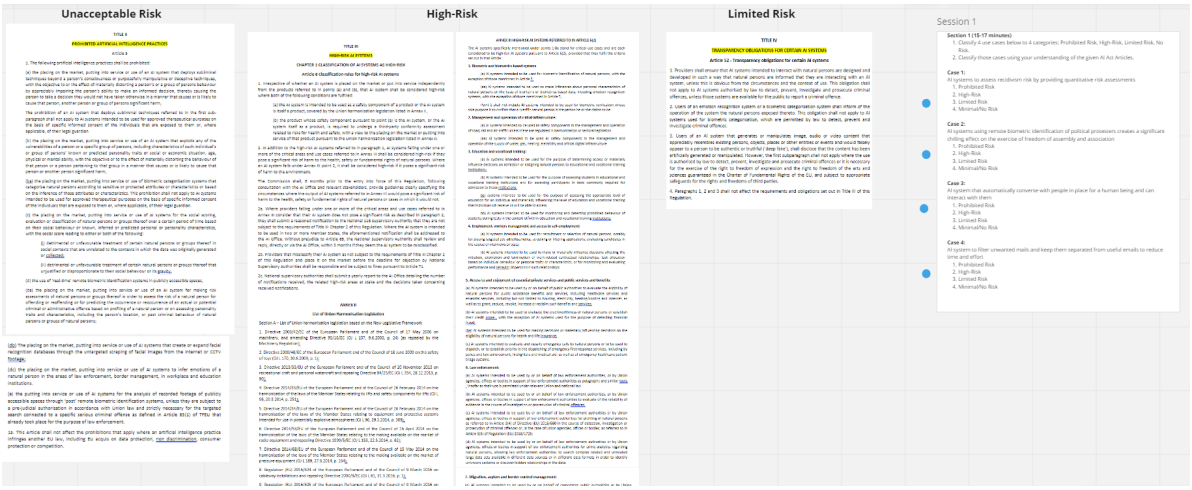

Figure A.1: First Section Board

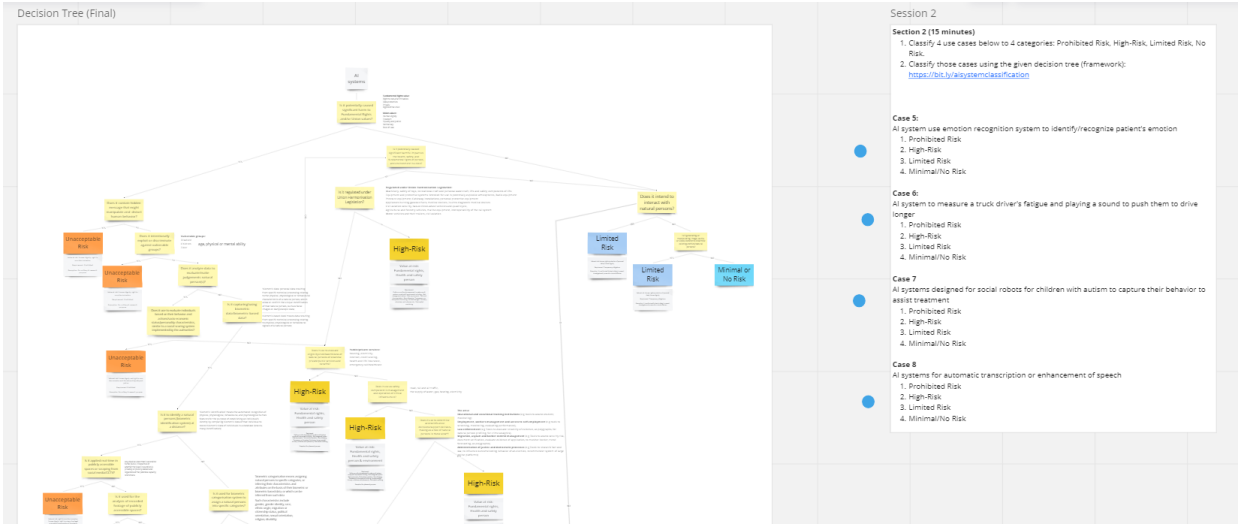

Figure A.2: Second Section Board

B

AI Act Articles

In this Appendix, all AI Act Articles mentioned in the report are presented: Article 5, Article 6, Article 52, Article 69, Annex II and Annex III refer to latest amendment of the AI Act on May, 2023 [11].

B.1. Article 5 - Prohibited Artificial Intelligence Practices

1. The following artificial intelligence practices shall be prohibited:

(a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective to or the effect of materially distorting a person's or a group of persons behaviour by appreciably impairing the person's ability to make an informed decision, thereby causing the person to take a decision they would not have taken otherwise in a manner that causes or is likely to cause that person, another person or group of persons significant harm;

The prohibition of an AI system that deploys subliminal techniques referred to in the first sub-paragraph shall not apply to AI systems intended to be used for approved therapeutical purposes on the basis of specific informed consent of the individuals that are exposed to them or, where applicable, of their legal guardian

(b) the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a person or a specific group of persons, including characteristics of such individual's or group of persons' known or predicted personality traits or social or economic situation, age, physical or mental ability, with the objective or to the effect of materially distorting the behaviour of that person or a person pertaining to that group in a manner that causes or is likely to cause that person or another person significant harm;

(ba) the placing on the market, putting into service or use of biometric categorisation systems that categorise natural persons according to sensitive or protected attributes or characteristics or based on the inference of those attributes or characteristics. This prohibition shall not apply to AI systems intended to be used for approved therapeutical purposes on the basis of specific informed consent of the individuals that are exposed to them or, where applicable, of their legal guardian.

(c) the placing on the market, putting into service or use of AI systems for the social scoring, evaluation or classification of natural persons or groups thereof over a certain period of time based on their social behaviour or known, inferred or predicted personal or personality characteristics, with the social score leading to either or both of the following:

- (i) detrimental or unfavourable treatment of certain natural persons or groups thereof in social contexts that are unrelated to the contexts in which the data was originally generated or collected;
- (ii) detrimental or unfavourable treatment of certain natural persons or groups thereof that unjustified or disproportionate to their social behaviour or its gravity;

(d) the use of 'real-time' remote biometric identification systems in publicly accessible spaces,

(da) the placing on the market, putting into service or use of an AI system for making risk assessments of natural persons or groups thereof in order to assess the risk of a natural person for offending or reoffending or for predicting the occurrence or reoccurrence of an actual or potential criminal or administrative offence based on profiling of a natural person or on assessing personality traits and characteristics, including the person's location, or past criminal behaviour of natural persons or groups of natural persons;

(db) The placing on the market, putting into service or use of AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage;

(dc) the placing on the market, putting into service or use of AI systems to infer emotions of a natural person in the areas of law enforcement, border management, in workplace and education institutions.

(e) the putting into service or use of AI systems for the analysis of recorded footage of publicly accessible spaces through 'post' remote biometric identification systems, unless they are subject to a pre-judicial authorisation in accordance with Union law and strictly necessary for the targeted search connected to a specific serious criminal offense as defined in Article 83(1) of TFEU that already took place for the purpose of law enforcement.

1a. This Article shall not affect the prohibitions that apply where an artificial intelligence practice infringes another EU law, including EU acquis on data protection, non discrimination, consumer protection or competition.

B.2. Article 6 - Classification Rules for High-Risk AI Systems

1. Irrespective of whether an AI system is placed on the market or put into service independently from the products referred to in points (a) and (b), that AI system shall be considered high-risk where both of the following conditions are fulfilled:

- (a) the AI system is intended to be used as a safety component of a product or the AI system is itself a product, covered by the Union harmonisation legislation listed in Annex II,
- (b) the product whose safety component pursuant to point (a) is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment related to risks for health and safety, with a view to the placing on the market or putting into service of that product pursuant to the Union harmonisation legislation listed in Annex II.

2. In addition to the high-risk AI systems referred to in paragraph 1, AI systems falling under one or more of the critical areas and use cases referred to in Annex III shall be considered high-risk if they pose a significant risk of harm to the health, safety or fundamental rights of natural persons. Where an AI system falls under Annex III point 2, it shall be considered high-risk if it poses a significant risk of harm to the environment.

The Commission shall, 6 months prior to the entry into force of this Regulation, following consultation with the AI Office and relevant stakeholders, provide guidelines clearly specifying the circumstances where the output of AI systems referred to in Annex III would pose a significant risk of harm to the health, safety or fundamental rights of natural persons or cases in which it would not.

2a. Where providers falling under one or more of the critical areas and use cases referred to in Annex III consider that their AI system does not pose a significant risk as described in paragraph 2, they shall submit a reasoned notification to the National 118 Supervisory Authority that they are not subject to the requirements of Title III Chapter 2 of this Regulation. Where the AI system is intended to be used in two or more Member States, the aforementioned notification shall be addressed to the AI Office. Without prejudice to Article 65, the National Supervisory Authority shall review and reply, directly or via the AI Office, within 3 months if they deem the AI system to be misclassified.

2b. Providers that misclassify their AI system as not subject to the requirements of Title III Chapter 2 of this Regulation and place it on the market before the deadline for objection by National Supervisory Authorities shall be responsible and be subject to fines pursuant to Article 71.

2c. National supervisory authorities shall submit a yearly report to the AI Office detailing the number of notifications received, the related high-risk areas at stake and the decisions taken concerning received notifications.

B.3. Article 52 - Transparency Obligations for Certain AI Systems

1. Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use. This obligation shall not apply to AI systems

authorised by law to detect, prevent, investigate and prosecute criminal offences, unless those systems are available for the public to report a criminal offence.

2. Users of an emotion recognition system or a biometric categorisation system shall inform of the operation of the system the natural persons exposed thereto. This obligation shall not apply to AI systems used for biometric categorisation, which are permitted by law to detect, prevent and investigate criminal offences.

3. Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated. However, the first subparagraph shall not apply where the use is authorised by law to detect, prevent, investigate and prosecute criminal offences or it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties.

4. Paragraphs 1, 2 and 3 shall not affect the requirements and obligations set out in Title III of this Regulation.

B.4. Article 69 - Codes of Conduct

1. The Commission and the Member States shall encourage and facilitate the drawing up of codes of conduct intended to foster the voluntary application to AI systems other than high-risk AI systems of the requirements set out in Title III, Chapter 2 on the basis of technical specifications and solutions that are appropriate means of ensuring compliance with such requirements in light of the intended purpose of the systems.

2. The Commission and the Board shall encourage and facilitate the drawing up of codes of conduct intended to foster the voluntary application to AI systems of requirements related for example to environmental sustainability, accessibility for persons with a disability, stakeholders participation in the design and development of the AI systems and diversity of development teams on the basis of clear objectives and key performance indicators to measure the achievement of those objectives.

3. Codes of conduct may be drawn up by individual providers of AI systems or by organisations representing them or by both, including with the involvement of users and any interested stakeholders and their representative organisations. Codes of conduct may cover one or more AI systems taking into account the similarity of the intended purpose of the relevant systems.

B.5. Annex II - List of Union Harmonisation Legislation

B.5.1. Section A - List of Union Harmonisation Legislation based on the New Legislative Framework

1. Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC (OJ L 157, 9.6.2006, p. 24) [as repealed by the Machinery Regulation];

2. Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys (OJ L 170, 30.6.2009, p. 1);

3. Directive 2013/53/EU of the European Parliament and of the Council of 20 November 2013 on recreational craft and personal watercraft and repealing Directive 94/25/EC (OJ L 354, 28.12.2013, p. 90);

4. Directive 2014/33/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to lifts and safety components for lifts (OJ L 96, 29.3.2014, p. 251);

5. Directive 2014/34/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to equipment and protective systems intended for use in potentially explosive atmospheres (OJ L 96, 29.3.2014, p. 309);

6. Directive 2014/53/EU of the European Parliament and of the Council of 16 April 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of radio equipment and repealing Directive 1999/5/EC (OJ L 153, 22.5.2014, p. 62);

7. Directive 2014/68/EU of the European Parliament and of the Council of 15 May 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of pressure equipment (OJ L

189, 27.6.2014, p. 164);

8. Regulation (EU) 2016/424 of the European Parliament and of the Council of 9 March 2016 on cableway installations and repealing Directive 2000/9/EC (OJ L 81, 31.3.2016, p. 1);

9. Regulation (EU) 2016/425 of the European Parliament and of the Council of 9 March 2016 on personal protective equipment and repealing Council Directive 89/686/EEC (OJ L 81, 31.3.2016, p. 51);

10. Regulation (EU) 2016/426 of the European Parliament and of the Council of 9 March 2016 on appliances burning gaseous fuels and repealing Directive 2009/142/EC (OJ L 81, 31.3.2016, p. 99); 121

11. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (OJ L 117, 5.5.2017, p. 1;

12. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (OJ L 117, 5.5.2017, p. 176).

B.5.2. Section B - List of Other Union Harmonisation Legislation

1. Regulation (EC) No 300/2008 of the European Parliament and of the Council of 11 March 2008 on common rules in the field of civil aviation security and repealing Regulation (EC) No 2320/2002 (OJ L 97, 9.4.2008, p. 72).

2. Regulation (EU) No 168/2013 of the European Parliament and of the Council of 15 January 2013 on the approval and market surveillance of two- or three-wheel vehicles and quadricycles (OJ L 60, 2.3.2013, p. 52);

3. Regulation (EU) No 167/2013 of the European Parliament and of the Council of 5 February 2013 on the approval and market surveillance of agricultural and forestry vehicles (OJ L 60, 2.3.2013, p. 1);

4. Directive 2014/90/EU of the European Parliament and of the Council of 23 July 2014 on marine equipment and repealing Council Directive 96/98/EC (OJ L 257, 28.8.2014, p. 146);

5. Directive (EU) 2016/797 of the European Parliament and of the Council of 11 May 2016 on the interoperability of the rail system within the European Union (OJ L 138, 26.5.2016, p. 44).

6. Regulation (EU) 2018/858 of the European Parliament and of the Council of 30 May 2018 on the approval and market surveillance of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles, amending Regulations (EC) No 715/2007 and (EC) No 595/2009 and repealing Directive 2007/46/EC (OJ L 151, 14.6.2018, p. 1); 3. Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users, amending Regulation (EU) 2018/858 of the European Parliament and of the Council and repealing Regulations (EC) No 78/2009, (EC) No 79/2009 and (EC) No 661/2009 of the European Parliament and of the Council and Commission Regulations (EC) No 631/2009, (EU) No 406/2010, (EU) No 672/2010, (EU) No 1003/2010, (EU) No 1005/2010, (EU) No 1008/2010, (EU) No 1009/2010, (EU) No 19/2011, (EU) No 109/2011, (EU) No 458/2011, (EU) No 65/2012, (EU) No 130/2012, (EU) No 347/2012, (EU) No 351/2012, (EU) No 1230/2012 and (EU) 2015/166 (OJ L 325, 16.12.2019, p. 1);

7. Regulation (EU) 2018/1139 of the European Parliament and of the Council of 4 July 2018 on common rules in the field of civil aviation and establishing a European Union Aviation Safety Agency, and amending Regulations (EC) No 2111/2005, (EC) No 1008/2008, (EU) No 996/2010, (EU) No 376/2014 and Directives 2014/30/EU and 2014/53/EU of the European Parliament and of the Council, and repealing Regulations (EC) No 552/2004 and (EC) No 216/2008 of the European Parliament and of the Council and Council Regulation (EEC) No 3922/91 (OJ L 212, 22.8.2018, p. 1), in so far as the design, production and placing on the market of aircrafts referred to in points (a) and (b) of Article 2(1) thereof, where it concerns unmanned aircraft and their engines, propellers, parts and equipment to control them remotely, are concerned.

B.6. Annex III - High-Risk AI Systems Referred to in Article 6(2)

The AI systems specifically mentioned under points 1-8a stand for critical use cases and are each considered to be high-risk AI systems pursuant to Article 6(2), provided that they fulfil the criteria set out in that Article:

1. Biometric and biometrics-based systems

(a) AI systems intended to be used for biometric identification of natural persons, with the exception of those mentioned in Article 5;

(aa) AI systems intended to be used to make inferences about personal characteristics of natural persons on the basis of biometric or biometrics-based data, including emotion recognition systems, with the exception of those mentioned in Article 5;

Point 1 shall not include AI systems intended to be used for biometric verification whose sole purpose is to confirm that a specific natural person is the person he or she claims to be

2. Management and operation of critical infrastructure:

(a) AI systems intended to be used as safety components in the management and operation of road, rail and air traffic unless these are regulated in harmonisation or sectoral legislation.

(aa) AI systems intended to be used as safety components in the management and operation of the supply of water, gas, heating, electricity and critical digital infrastructure

3. Education and vocational training:

(a) AI systems intended to be used for the purpose of determining access or materially influence decisions on admission or assigning natural persons to educational and vocational training institutions;

(b) AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to those institutions;

ba) systems intended to be used for the purpose of assessing the appropriate level of education for an individual and materially influencing the level of education and vocational training that individual will receive or will be able to access.

bb) AI systems intended to be used for monitoring and detecting prohibited behaviour of students during tests in the context of/within education and vocational training institutions;

4. Employment, workers management and access to self-employment:

(a) AI systems intended to be used for recruitment or selection of natural persons, notably for placing targeted job advertisements, screening or filtering applications, evaluating candidates in the course of interviews or tests;

(b) AI systems intended to be used to make or materially influence decisions affecting the initiation, promotion and termination of work-related contractual relationships, task allocation based on individual behaviour or personal traits or characteristics, or for monitoring and evaluating performance and behavior of persons in such relationships.

5. Access to and enjoyment of essential private services and public services and benefits:

(a) AI systems intended to be used by or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, including healthcare services and essential services, including but not limited to housing, electricity, heating/cooling and internet, as well as to grant, reduce, revoke, increase or reclaim such benefits and services;

(b) AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score , with the exception of AI systems used for the purpose of detecting financial fraud;

(ba) AI systems intended to be used for making decisions or materially influencing decisions on the eligibility of natural persons for health and life insurance;

(c) AI systems intended to evaluate and classify emergency calls by natural persons or to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by police and law enforcement, firefighters and medical aid, as well as of emergency healthcare patient triage systems.

6. Law enforcement:

(a) AI systems intended to be used by or on behalf of law enforcement authorities, or by Union agencies, offices or bodies in support of law enforcement authorities as polygraphs and similar tools ; insofar as their use is permitted under relevant Union and national law

(b) AI systems intended to be used by or on behalf of law enforcement authorities, or by Union agencies, offices or bodies in support of law enforcement authorities to evaluate of the reliability of evidence in the course of investigation or prosecution of criminal offences;

(c) AI systems intended to be used by or on behalf of law enforcement authorities or by Union agencies, offices or bodies in support of law enforcement authorities for profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of detection, investigation or prosecution of criminal offences or, in the case of Union agencies, offices or bodies, as referred to in Article 3(5) of Regulation (EU) 2018/1725;

(d) AI systems intended to be used by or on behalf of law enforcement authorities or by Union agencies, offices or bodies in support of law enforcement authorities for crime analytics regarding natural persons, allowing law enforcement authorities to search complex related and unrelated large data sets available in different data sources or in different data formats in order to identify unknown patterns or discover hidden relationships in the data.

7. Migration, asylum and border control management:

(a) AI systems intended to be used by or on behalf of competent public authorities or by Union agencies, offices or bodies as polygraphs and similar tools insofar as their use is permitted under relevant Union or national law

(b) AI systems intended to be used by or on behalf of competent public authorities or by Union agencies, offices or bodies to assess a risk, including a security risk, a risk of irregular immigration, or a health risk, posed by a natural person who intends to enter or has entered into the territory of a Member State;

(c) AI systems intended to be used by or on behalf of competent public authorities or by Union agencies, offices or bodies for the verification of the authenticity of travel documents and supporting documentation of natural persons and detect non-authentic documents by checking their security features;

(d) AI systems intended to be used by or on behalf of competent public authorities or by Union agencies, offices or bodies to assist competent public authorities for the examination and assessment of the veracity of evidence in relation to applications for asylum, visa and residence permits and associated complaints with regard to the eligibility of the natural persons applying for a status. 125

(da) AI systems intended to be used by or on behalf of competent public authorities or by Union agencies, offices or bodies in migration, asylum and border control management to monitor, surveil or process data in the context of border management activities, for the purpose of detecting, recognising or identifying natural persons

(db) AI systems intended to be used by or on behalf of competent public authorities or by Union agencies, offices or bodies in migration, asylum and border control management for the forecasting or prediction of trends related to migration movement and border crossing

8. Administration of justice and democratic processes:

a) AI systems intended to be used by a judicial authority or administrative body or on their behalf to assist a judicial authority or administrative body in researching and interpreting facts and the law and in applying the law to a concrete set of facts or used in a similar way in alternative dispute resolution.

aa) AI systems intended to be used for influencing the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda This does not include AI systems whose output natural persons are not directly exposed to, such as tools used to organise, optimise and structure political campaigns from an administrative and logistic point of view.

(ab) AI systems intended to be used by social media platforms that have been designated as very large online platforms within the meaning of Article 33 of Regulation EU 2022/2065, in their recommender systems to recommend to the recipient of the service user-generated content available on the platform.

C

List of Borderline (Non-Obvious) Cases

Below are some borderline (non-obvious) use cases according to several reasons as described in Chapter 4.

Table C.1: Some Borderline (Non-Obvious) Cases

Use Case	Classification	Source
AI systems to assess recidivism risk by providing quantitative risk assessments	Prohibited/High-Risk	Van Dijck (2022) [46]
AI systems for robo-advisors in insurance	Prohibited/High-risk/Limited	Marano & Li (2023) [28]
AI systems for dating counseling service	High-Rik/Prohibited	Lim et. al (2022)* [26]
AI systems as an interactive chatbot service where the expose data without filtering when it was asked for an address or account (e.g Iruda service)	High-Risk/Prohibited	Lim et. al (2022)* [26]
AI systems to provide recommendation for consumers	Limited/Minimal Risk	De Cooman (2022) [12]
AI systems to detect cartel (competition law enforcement authorities)	Prohibited/Limited/Minimal Risk	De Cooman (2022) [12]
Measuring a truck driver's fatigue and playing a sound that pushes them to drive longer	Prohibited Risk	PWC (2022) [39] **
surveillance (video surveillance at human level using e.g bodycam)	Prohibited/High-RIsk	Hupont et. al (2022) [18]
Unconstrained face identification	Prohibited/High-RIsk	Hupont et. al (2022) [18]
Person re-identification	Prohibited/High-RIsk	Hupont et. al (2022) [18]
Person tracking with drones	Prohibited/High-RIsk	Hupont et. al (2022) [18]
Control of attendance	High-Risk/Minimal Risk	Hupont et. al (2022) [18]
Mobile surveillance robots	Prohibited/High-RIsk	Hupont et. al (2022) [18]
Person search by facial apperance	Limited RIsk/Minimal Risk	Hupont et. al (2022) [18]
Face mask detection	Limited RIsk/Minimal Risk	Hupont et. al (2022) [18]
Clinical syndrome assessment	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Student proctoring and tutoring	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Job interviews	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Driver monitoring and warning	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Driver monitoring for autonomous vehicles	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Affective robots as companions for elderly	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Social robots for children with autism	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Pain detection	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Police interrogations	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Emotion estimation in groups or crowds	High-Risk/Limited Risk	Hupont et. al (2022) [18]
Visual lifelogging as memory aid	High-Risk/Minimal Risk	Hupont et. al (2022) [18]
Automatic transcription or enhancement of speech	High-Risk/Minimal Risk	Hupont et. al (2022) [18]
Speech recognition for voice impaired	High-Risk/Minimal Risk	Hupont et. al (2022) [18]
Face-guided communication and interaction	High-Risk/Minimal Risk	Hupont et. al (2022) [18]

* may influence by high-risk, but regarding AI regulation is not High-Risk but prohibited
** For the evaluation in this research, this system is assumed as part-of the Non-Obvious case due to the slight contravening value of 'measuring fatigue' and 'push to drive longer'.

D

Abstraction of Each Risk Class

This section presents the abstraction of each risk level: Unacceptable Risk, High-Risk, and Limited Risk.

Table D.1: Abstraction of Unacceptable Risk

Protected Values	Objective/Intention	Affecting…	Technology/Use Case	How	Risk to …	Intertwined regulation	Input Data	Exemption	Requirement
Contradict Fundamental Rights & Union values	Materially distorting human behaviour	Physical or psychological harms	Neuro-technologist assisted by AI systems that are used to monitor, use, or influence neural data gathered trough brain-computer interfaces	Deploy subliminal components individuals cannot perceive or exploit vulnerabilities of individuals and specific groups of persons		Directive [Unfair Commercial Practice Directive 2005/29/EC, as amended by Directive (EU) 2019/216]	Neural data, biometric data	Research practice, with consent form (ethical standards) and not expose natural persons to harm	Prohibited
	Exploit vulnerabilities of individuals and specific groups of persons	Significant harm to groups or persons	AI systems to categorise natural persons by assigning them to specific categories	inferred sensitive or protected characteristics are particularly intrusive, violate human dignity, and hold great risk of discrimination	human dignity, risk of discrimination	Article 21 of the EU Charter of Fundamental Rights; Article 9 of Regulation (EU)2016/769	Sensitive data, including Gender, gender identity, race, ethnic origin, migration or citizenship status, political orientation, sexual orientation, religion, disability or any other grounds	Research practice, with consent form (ethical standards) and not expose natural persons to harm	
	Social control practices	-	AI systems providing social scoring of natural persons for general purpose	Discriminatory outcomes and the exclusion of certain groups	Right to dignity, and non-discrimination and the values of equality and justice	-	Gender, gender identity, race, ethnic origin, migration or citizenship status, political orientation, sexual orientation, religion, disability or any other grounds	-	
		-	AI systems evaluate or classify natural persons or groups based on multiple data points and time occurrences related to their social behavior in multiple contexts or known, inferred or predicted personal or personality characteristics	-	-	-	Social behavior	-	
	Law enforcement	-	AI systems used by law enforcement authorities to make predictions, profiles or risks assessment based on profiling of natural persons or data analysis based on personality traits and characteristics, including a person's location, or past criminal behavior	-	Risk of discrimination, right to dignity, key legal principle of presumption of innocence	-	Personality trait; Profiling of natural persons; person's location, past criminal behavior	-	
		-	The indiscriminate and untargeted scraping of biometric data from social media or CCTV footage to create or expand facial recognition databases add to the feeling of mass surveillance	-	Right to privacy, gross violations of fundamental rights	-	Scrapping biometric data	-	
	Detect emotional state of individuals	-	Detect emotions, physical or physiological features (facial expressions, movements, pulse frequency or voice) in the areas of law enforcement, border management, in workplace and education institutions	-	-	-	-	Application outside law enforcement, border management, workplace and education institutions.	
	Use of real-time remote biometric identification systems in publicly accessible spaces		AI systems for real-time remote biometric identification of natural persons in publicly accessible spaces	-	Rights and freedoms of person, rule of law, private life of large part of the population	Connected to a specific curious criminal offense (Article 83(1) of TFEU; Article 6a of Protocol No. 21	Biometric data	United Kingdom, Ireland, Denmark	
	Post-remote biometric identification system	-	AI systems for the analysis of recorded footage of publicly accessible spaces through 'post' remote biometric identification systems		-	Connected to a specific curious criminal offense (Article 83(1) of TFEU; Articles 2 and 2a of Protocol No 22	-	Denmark	

Table D.2: Abstraction of High-Risk (1)

Protected Values	Type	Domain	Objective/Intention	Benefit	Potential Risk	Intertwined regulation	Exemption	Requirement
Alignment with sectoral legislation, impact on health, safety, fundamental rights of persons, environment	safety component of products or product falling within regulation	Mentioned in Annex II A	-	-	-	Mentioned in Annex II AI Act	-	third party conformity assessment & mandatory requirements
	stand-alone AI systems	biometric and biometric-based systems	AI systems intended to be used for biometric identification of natural persons	to confirm that a specific natural person is the person he or she claims to be and to confirm the identity of a natural person for the sole purpose of having access to a service, a device or premises (one-to-one verification)	personal data (sensitive data), safety	-	Biometric and biometrics-based systems which are foreseen under EU law to enable cybersecurity and personal data protection measures	mandatory requirements: risk management measures (Article 9), the utilization of high-quality data and appropriate data governance practices (Article 10), thorough documentation and record-keeping to ensure traceability (Article 11 and 12), transparency in the system's (Article 13), the presence of human oversight (Article 14), and the assurance of accuracy, robustness, and cybersecurity of AI systems (Article 15)
			AI systems intended to be used to make inferences about personal characteristics of naturals persons on the basis of biometric or biometrics-based data, incl. emotion recognition systems (except prohibited)			-	biometric verification, includes authentication	
		management and operation of critical infrastructure	AI systems intended to be used as safety components in the management and operation of (1) the supply of water, gas, heating electricity and critical digital infrastructure, of (2) road, rail, and air traffic	Directly protect the physical integrity of physical infrastructure or health and safety of persons and property	risk of health and safety of persons and property	-	Component intended to be used solely for cybersecurity purposes	
		education institution	AI systems intended to be used for the purpose of determining access or materially influence decisions on admission or assigning natural persons to educational and vocational training institutions	to help modernise entire education systems, to increase educational quality, both offline and online and to accelerate digital education, thus also making it available to a broader audience .	ability to secure their livelihood; right to education and training; right not to be discriminated; perpetuate historical patterns of discrimination	-	-	
			AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to those institutions;			-	-	
			AI systems intended to be used for the purpose of assessing the appropriate level of education for an individual and materially influencing the level of education and vocational training that individual will receive or will be able to access.			-	-	
			AI systems intended to be used for monitoring and detecting prohibited behaviour of students during tests in the context of/within education and vocational training institutions;			-	-	
		employment, workers management and access to self-employment	AI systems used in employment, workers management and access to self-employment, notably for the recruitment and selection of persons, for making decisions or materially influence decisions on initiation, promotion and termination and for personalised task allocation	help to make decisions	ability to secure their livelihood; worker's rights, future career prospects; right not to be discriminated; perpetuate historical patterns of discrimination	-	-	
			AI systems used to monitor the performance and behaviour of these persons		data protection and privacy	-	-	
		access to and enjoyment of certain essential private and public services	AI systems used to evaluate the credit score or creditworthiness of natural persons	help to fully participate in society or to improve one's standard	right not to be discriminated; perpetuate historical patterns of discrimination	-	cybersecurity purpose (detecting fraud in the offering of financial services)	
			AI systems are used for determining whether such benefits and services should be denied, reduced, revoked or reclaimed by authorities		impact on livelihood; fundamental rights (right to social protection, non-discrimination, human dignity, effective remedy)	-	-	
			AI systems intended to be used to make decisions or materially influence decisions on the eligibility of natural persons for health and life insurance			-	-	
			AI systems used to evaluate and classify emergency calls by natural persons or to dispatch or establish priority in the dispatching of emergency first response services	make decisions in very critical situations for the life and health of persons and their property	impact on livelihood; fundamental rights (right to social protection, non-discrimination, human dignity, effective remedy)	-	-	

Table D.3: Abstraction of High-Risk (2)

Protected Values	Type	Domain	Objective/Intention	Benefit	Potential Risk	Intertwined regulation	Exemption	Requirement
Alignment with sectoral legislation, impact on health, safety, fundamental rights of persons, environment	stand-alone AI systems	law enforcement	AI systems to support of law enforcement authorities, as polygraphs and similar tools insofar	for the evaluation of the reliability of evidence in criminal proceedings, for profiling in the course of detection, investigation or prosecution of criminal offences, as well as for crime analytics regarding natural persons	fundamental rights	-	for administrative proceedings by tax and customs authorities	mandatory requirements: risk management measures (Article 9), the utilization of high-quality data and appropriate data governance practices (Article 10), thorough documentation and record-keeping to ensure traceability (Article 11 and 12), transparency in the system's (Article 13), the presence of human oversight (Article 14), and the assurance of accuracy, robustness, and cybersecurity of AI systems (Article 15)
			AI systems intended to support of law enforcement authorities to evaluate of the reliability of evidence in the course of investigation or prosecution of criminal offences;		-	-		
			AI systems intended for profiling of natural persons		-	-		
			AI systems intended to be used for crime analytics regarding natural persons, allowing law enforcement authorities to search complex related and unrelated large data sets available in different data sources or in different data formats		-	-		
		migration, asylum and border control management	AI systems intended to be used as polygraphs and similar tools	-	fundamental rights; rights to free movement; rights not to be discriminated; protection to personal data and private life; international protection; good administration	-	-	
			AI systems intended to be used to assess a risk, including a security risk, a risk of irregular immigration, or a health risk, posed by a natural person who intends to enter or has entered into the territory of a Member State	-		-	-	
			AI systems intended to be used for the verification of the authenticity of travel documents and supporting documentation of natural persons and detect non-authentic documents by checking their security features	-		-	-	
			AI systems intended to assist competent public authorities for the examination and assessment of the veracity of evidence in relation to applications for asylum, visa and residence permits and associated complaints with regard to the eligibility of the natural persons applying for a status.	-		-	-	
			AI systems intended to monitor, surveil or process data in the context of border management activities, for the purpose of detecting, recognising or identifying natural persons	-		-	-	
			AI systems intended for the forecasting or prediction of trends related to migration movement and border crossing	-		-	-	
		administration of justice and democratic processes	AI systems intended to be used researching and interpreting facts and the law and in applying the law to a concrete set of facts or used in a similar way in alternative dispute resolution	-	impact on democracy; rule of law; individual freedoms ; right to an effective remedy and to a fair trial	-	for purely ancillary administrative activities that do not affect the actual administration of justice in individual cases, such as anonymisation or pseudonymisation of judicial decisions, documents or data, communication between personnel, administrative tasks or allocation of resources.	
			AI systems intended to be used to influence the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda	-		-	AI systems whose output natural persons are not directly exposed to, such as tools used to organise, optimise and structure political campaigns from an administrative and logistic point of view.	
			AI systems intended to be used by social media platforms that have been designated as very large online platforms in their recommender systems to recommend to the recipient of the service user-generated content available on the platform.	-		-	-	

Table D.4: Abstraction of Limited Risk

Protected Values	Intention	Technology/Use Case	Exemption	Requirement
Trustworthiness; transparency	Interact with natural persons	Converse to a natural person	AI systems authorised by law to detect, prevent, investigate and prosecute criminal offences, unless those systems are available for the public to report a criminal offence	designed in a way that natural persons are informed they are interacting with an AI system
	-	Emotion recognition system / a biometric categorisation system	Biometric categorisation, which are permitted by law to detect, prevent and investigate criminal offences	inform the operation of the system the natural persons exposed thereto
	Generates/manipulates image, audio or video content that appreciably resembles existing persons, objects/places, appear to be a person	Deep fake technology	AI systems authorised by law to detect, prevent, investigate and prosecute criminal offences, unless those systems are available for the public to report a criminal offence	shall disclose that the content has been artificially generated or manipulated

E

Decision Tree Framework Evaluation

E.1. Confusion Matrix - Python Code

Below is the python code to visualize confusion matrix and calculate the accuracy, precision, recall, and F-1 scores.

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Confusion matrices
confusion_matrix_obvious = np.array([[8, 0, 0, 0],
                                     [2, 4, 2, 0],
                                     [1, 3, 4, 0],
                                     [1, 1, 1, 5]])

confusion_matrix_obvious_tree = np.array([[7, 1, 0, 0],
                                          [1, 3, 4, 0],
                                          [1, 0, 7, 0],
                                          [1, 0, 1, 6]])

# Class labels
labels = ["Unacceptable Risk", "High-Risk", "Limited Risk", "No/Minimal Risk"]

# Function to calculate performance metrics
def calculate_metrics(confusion_matrix):
    true_positives = np.diag(confusion_matrix)
    false_positives = np.sum(confusion_matrix, axis=0) - true_positives
    false_negatives = np.sum(confusion_matrix, axis=1) - true_positives

    precision = true_positives / (true_positives + false_positives)
    recall = true_positives / (true_positives + false_negatives)
    f1_score = 2 * (precision * recall) / (precision + recall)
    accuracy = np.sum(true_positives) / np.sum(confusion_matrix)

    return precision, recall, f1_score, accuracy

# Calculate performance metrics for obvious case without decision tree
precision_obvious, recall_obvious, f1_score_obvious, accuracy_obvious =
calculate_metrics(confusion_matrix_obvious)

# Calculate performance metrics for obvious case with decision tree
precision_obvious_tree, recall_obvious_tree, f1_score_obvious_tree,
accuracy_obvious_tree = calculate_metrics(confusion_matrix_obvious_tree)
```

```
# Print performance metrics for both cases
print("All Respondents")
print("Performance Metrics for Obvious Case (Without Decision Tree)")
print("Precision:", precision_obvious)
print("Recall:", recall_obvious)
print("F1-Score:", f1_score_obvious)
print("Accuracy:", accuracy_obvious)
print()

print("Performance Metrics for Obvious Case (With Decision Tree)")
print("Precision:", precision_obvious_tree)
print("Recall:", recall_obvious_tree)
print("F1-Score:", f1_score_obvious_tree)
print("Accuracy:", accuracy_obvious_tree)

# Plot confusion matrices using seaborn
plt.figure(figsize=(12, 6))

# Obvious Case without Decision Tree
plt.subplot(1, 2, 1)
sns.heatmap(confusion_matrix_obvious, annot=True, cmap="Blues", fmt="d",
\xticklabels=labels, yticklabels=labels)
plt.title("Confusion Matrix - Obvious Case (Without Decision Tree)")
plt.xlabel("Predicted Labels")
plt.ylabel("True Labels")

# Obvious Case with Decision Tree
plt.subplot(1, 2, 2)
sns.heatmap(confusion_matrix_obvious_tree, annot=True, cmap="Oranges",
fmt="d", xticklabels=labels, yticklabels=labels)
plt.title("Confusion Matrix - Obvious Case (With Decision Tree)")
plt.xlabel("Predicted Labels")
plt.ylabel("True Labels")

plt.tight_layout()
plt.show()
```

E.2. Time Performance - Python Code

Below is the python code to visualize the time performance of the decision tree.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Data organization (replace commas with periods for consistent float representation)
data = {
    'Case': ['Case 1'] * 8 + ['Case 1'] * 8 + ['Case 2'] * 8 + ['Case 2'] * 8 +
            ['Case 3'] * 8 + ['Case 3'] * 8 + ['Case 4'] * 8 + ['Case 4'] * 8 +
            ['Case 5'] * 8 + ['Case 5'] * 8 + ['Case 6'] * 8 + ['Case 6'] * 8 +
            ['Case 7'] * 8 + ['Case 7'] * 8 + ['Case 8'] * 8 + ['Case 8'] * 8,
    'Approach': (['Without DT'] * 8 + ['With DT'] * 8) * 8,
    'Duration': [
        8, 97, 129, 60, 46.5, 184.25, 22.25, 121.25,
        52.25, 125, 88, 62, 54.75, 309, 61, 71.25,
        6, 97, 91, 156, 46.5, 184.25, 22.25, 31,
        52.25, 162, 30, 150, 54.75, 301, 61, 30,
        22, 97, 75.5, 106, 46.5, 184.25, 22.25, 52,
        35, 58, 321, 150, 54.75, 102, 52.25, 139,
        21, 317, 75.5, 128, 46.5, 105.25, 22.25, 105,
        52.25, 62.75, 573, 150, 54.75, 92, 11, 53,
        47.25, 105.25, 22.25, 40, 63, 176, 75.5, 170,
        16, 62.75, 252, 150, 98.75, 92, 95, 68,
        121.25, 47.25, 105.25, 311, 93, 86, 75.5, 22.25,
        60, 62.75, 440, 71.25, 98.75, 92, 39, 55,
        22.25, 121.25, 117, 313, 93, 150, 47.25, 105.25,
        17, 62.75, 16, 202, 98.75, 92, 61, 71.25,
        184.25, 97, 230, 58, 93, 22.25, 121.25, 47.25,
        20, 98.75, 80, 56, 161, 53, 61, 71.25
    ]
}

# Create a pandas DataFrame
df = pd.DataFrame(data)

# Set the color palette
colors = {'Without DT': 'blue', 'With DT': 'orange'}
sns.set_palette(colors.values())

# Filter data for different cases
cases_1_2_7_8 = df[df['Case'].isin(['Case 1', 'Case 2', 'Case 7', 'Case 8'])]
cases_3_4_5_6 = df[df['Case'].isin(['Case 3', 'Case 4', 'Case 5', 'Case 6'])]

# Create separate figures for the vertical boxplots
fig, axes = plt.subplots(2, 1, figsize=(10, 12))
sns.boxplot(data=cases_1_2_7_8, x='Case', y='Duration', hue='Approach', ax=axes[0])
axes[0].set_title('Duration Comparison Obvious Cases')
axes[0].set_xlabel('Case')
axes[0].set_ylabel('Duration (second)')
axes[0].legend(title='Approach')

sns.boxplot(data=cases_3_4_5_6, x='Case', y='Duration', hue='Approach', ax=axes[1])
axes[1].set_title('Duration Comparison for Non-Obvious Cases')
axes[1].set_xlabel('Case')
```

```
axes[1].set_ylabel('Duration (second)')
axes[1].legend(title='Approach')

plt.tight_layout()

plt.show()
```