Framework for the application of Explainable Artificial Intelligence techniques in the service of Democracy

Marta Sofia Marques da Encarnação

Dissertation presented as partial requirement for obtaining the Master’s degree in Information Management
Framework for the application of Explainable Artificial Intelligence techniques in the service of Democracy

By

Marta Sofia Marques da Encarnação

Master Thesis presented as partial requirement for obtaining the master’s degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Supervisor: Prof. Vitor Santos, Ph.D.
Supervisor: Prof. Maria Anastasiadou, Ph.D.

November 2023
Acknowledgements

To my family for always supporting me throughout all my academic journey and being present in every step of the way.

To Professor Dr. Vitor Santos and Professor Maria Anastasiadou, for agreeing on being my supervisors in the development of this Master Thesis and for all the guidance, assistance and contribution.

To Professor Dr. Mijail Naranjo, Professor Dr. Filipe Montargil, Professor Dr. Ana Cláudia Albergaria and Catarina Ferraz for the insights and feedback given, which helped improve the expected outcome of the present Master Thesis.

To all my former Professors at Nova IMS, for the knowledge and experience transmitted during the pursuit of studies in this master’s degree.

To all my colleagues and friends for the support and motivation given throughout this journey which indirectly or directly contributed to the development of the present Master Thesis.
Abstract

As artificial intelligence is increasingly deployed in democracies around the world, there is growing recognition of the need for transparency and accountability in Artificial Intelligence decision-making. Explainable Artificial Intelligence (XAI) offers a solution to the problem of AI "black boxes" by making the underlying reasoning and decision-making processes of Artificial Intelligence models more transparent to humans. This approach can help ensure that AI systems are aligned with democratic values and principles, and that they are not being used in ways that violate individual rights or compromise personal privacy.

This document studies the two above-mentioned scopes: Explainable Artificial Intelligence (XAI) and Democracy, which led to the development of an artifact. All the required steps are comprehensively explained as to how the proposed framework was constructed in order to be able to be applied in the service of Democracy through the application of Explainable Artificial Intelligence (XAI) techniques. The ultimate goal is to mitigate or even eliminate some threats and challenges Democracies face today and how Artificial Intelligence with the addition of an explainability trait could be of great help.

This study includes a thorough literature review on Explainable Artificial Intelligence (XAI) and its various concepts, approaches, algorithms and techniques relevant to democracy. This is followed by a study of the democratic environment, as well as the challenges and opportunities posed by using Artificial Intelligence in the democratic context. By considering both Explainable Artificial Intelligence (XAI) and the broader implications of Artificial Intelligence in democratic societies, this study seeks to advance our understanding of how to build trustworthy, transparent Artificial Intelligence systems that meet the demands of democratic governance while respecting individual rights and values.

After conducting a comprehensive literature review and other studies, a set of assumptions were established as foundation for the development of the artifact. These assumptions guided the creation of a framework, which is then described in detail with specific examples for each step.

The effectiveness of the framework is assessed through interviews conducted with specialists, and the resulting feedback was then analyzed and discussed. This section of the study also includes a conclusion regarding the value and validity of the framework as a useful tool for the practical application of Explainable Artificial Intelligence (XAI) techniques in the context of democracy.

Keywords

Explainable Artificial Intelligence; Democracy; Explainability; Transparency; Framework; Design Science Research
# Index

1. **Introduction** ........................................................................................................................................... 10
   1.1 Background and Problem Identification ......................................................................................... 10
   1.2 Objectives ........................................................................................................................................ 11
   1.3 Study Importance and Relevance ................................................................................................. 11
2. **Methodology** ......................................................................................................................................... 13
   2.1 Design Science Research (DSR) .................................................................................................. 13
   2.2 Research Strategy ......................................................................................................................... 15
3. **Literature Review - Explainable AI** ............................................................................................... 18
   3.1 Obscurity and bias in AI .............................................................................................................. 18
   3.2 XAI concepts ............................................................................................................................... 20
   3.3 Approaches & Algorithms ........................................................................................................... 21
   3.4 XAI techniques in Democracy .................................................................................................... 23
   3.5 Applications of XAI techniques .................................................................................................. 24
4. **Environment - Democracy** ............................................................................................................ 27
   4.1 Definition ....................................................................................................................................... 27
   4.2 History ........................................................................................................................................... 28
   4.3 Challenges and Problems in Democracy ..................................................................................... 30
   4.4 Challenges and opportunities of using AI in Democracy ............................................................ 31
      4.4.1 Challenges ............................................................................................................................... 31
      4.4.2 Opportunities ......................................................................................................................... 32
5. **Conceptual Framework** .................................................................................................................... 34
   5.1 Assumptions .................................................................................................................................... 34
   5.2 Proposal .......................................................................................................................................... 35
      5.2.1 Democracy challenges Assessment ....................................................................................... 35
      5.2.2 Identify ways to mitigate challenge ....................................................................................... 38
      5.2.3 Identify Requirements ............................................................................................................ 39
      5.2.4 Selection of XAI Technology .................................................................................................. 41
   5.3 Evaluation ........................................................................................................................................ 42
   5.4 Discussion of Results ...................................................................................................................... 43
6. **Conclusion** ........................................................................................................................................... 45
6.1 Limitations ........................................................................................................................................... 45
6.2 Future Work........................................................................................................................................... 46
References .................................................................................................................................................... 47
Appendix..................................................................................................................................................... 54
A - PowerPoint presentation of the framework for the application of XAI Techniques in the service of democracy ................................................................................................................................. 54
B - Interviews.............................................................................................................................................. 59
B.1 - Specialist 1........................................................................................................................................... 59
B.2 - Specialist 2........................................................................................................................................... 59
B.3 - Specialist 3........................................................................................................................................... 60
B.4 - Specialist 4.......................................................................................................................................... 60
List of Figures

Figure 1: Explainable AI Concept (Defense Advanced Research Projects Agency, 2021) .................. 12
Figure 2: DSR Methodology Process Model (Adapted from Peffers et al. (2008)) ......................... 14
Figure 3: Design Science Research Cycle (Hevner et al. 2004) .................................................. 15
Figure 4: Purposes of Explainability and how it can serve distinct types of publics ...................... 20
Figure 5: Framework for the selection of XAI techniques in the service of Democracy .................. 35
Figure 6: Flow chart for the identification and assessment of threats to Democracy ..................... 36
Figure 7: Flow chart for threat management and measure definition ........................................... 38
Figure 8: Flow Chart for requirements identification ........................................................................ 39
List of Tables

Table 1: Example of possible applications of XAI Techniques.......................................................... 26
Table 2: Guide table for Challenges' description by type (Froomkin & Shapiro, 2021) ...................... 36
Table 3: Probability/Impact matrix for risk analysis (Protect Democracy, 2023) ............................... 37
Table 4: Example of threat evaluation and risk assessment................................................................. 38
Table 5: Example of measure Definition............................................................................................ 39
Table 6: Example for data requirements ............................................................................................. 40
Table 7: Example for Model requirements ......................................................................................... 40
Table 8: Example of XAI Technology selection.................................................................................. 41
Table 9: Additional recommendations with challenges, possible measure and technology .......... 42
List of Abbreviations and Acronyms

AI Artificial Intelligence
Ax Assumption x
BC Before Christ
CAVs Concept Activation Vectors
DSR Design Science Research
DSRM Design Science Research Methodology
DSS Decision Support Systems
GDPR General Data Protection Regulation
ICE Individual Conditional Expectation
LIME Local Interpretable Model-agnostic Explainer
ML Machine Learning
NGO Non-governmental Organizations
NIST US National Institute of Standards and Technology
PDPs Partial Dependence Plots
SHAP SHapley Additive exPlanations
TCAV Testing with Concept Activation Vectors
XAI Explainable Artificial Intelligence
1 Introduction

1.1 Background and Problem Identification

Artificial Intelligence (AI) is the application of human-like intelligence in machines for simulating human traits such as learning and reasoning. The main goal is for a machine to easily execute tasks, simple or complex, by mimicking human cognitive capabilities (Jake Frankenfield, 2022). A subset of Artificial Intelligence is Machine Learning (ML) which is based on the idea that computer programs can automatically learn from and adapt to new data without human assistance (Jake Frankenfield, 2022). This automatic learning is enabled through deep learning algorithms by ingesting vast quantities of unstructured data, including text, photos, and video (Jake Frankenfield, 2022).

These days, almost every company has already incorporated AI in its operations or intends to do it. Naturally, the demand for transparency into how these models generate conclusions rises as more businesses integrate AI and advanced analytics into business processes and automate decisions (Nicklas Ankarstad, 2020). Even though these AI machines possess the ability to “think”, to make decisions they are programmed for, they do not have a consciousness, free-will or self-aware self (Glassman, 2019). In the running process of a ML algorithm, the concept of “black box” refers exactly to the impossibility of explaining what exactly happens inside them or how it arrived a specific result. This lack of understanding of AI inner workings and need for more transparency led to the necessity of a new AI state of the art: XAI – Explainable Artificial Intelligence.

The existence of the XAI concept was considered crucial since AI algorithms operate by taking inputs and providing outputs without a proper understanding for the user of the inner workings (Mike McNamara, 2022). Therefore, the aim of XAI is to turn the process behind the output of the algorithm, deep learning, and neural networks understandable for humans (Jake Frankenfield, 2022). XAI is a way to put the social right of explanation into practice while helping to understand the input variables impact on the output.

This concept is of great relevance not only for a matter of understanding what happens in the process of running an algorithm, but also to give enough credibility and power for AI to enter other social/scientific areas. It is crucial to understand how AI makes decisions since they have business impacts, positive and negative. Many organizations still feel uncomfortable on letting AI models make more impactful decisions because they do not trust it enough (Jake Frankenfield, 2022). Therefore, providing insights on how models reach certain conclusions and make decisions will help mitigate this current problem.

A major inherent issue to AI technologies is the tension between it and good public policy. Many democratic governments and agencies are incrementally relying on AI technologies such as facial recognition, process automation, virtual assistants, etc (Marsh, 2019). Some concerns are being raised - privacy and security, understanding how the AI models work exactly and the challenge of humans to trust automated decisions – the major one being the issue of bias. If a ML algorithm is trained on biased datasets, the model will reflect an output obviously biased, and this is a problem for the correct and fair use of AI (Diane Coyle, 2020).

Will explainability techniques help mitigate all these concerns? That is the hope and goal of XAI – Explainable Artificial Intelligence.
1.2 Objectives

The goal of the research is to build a comprehensive framework for the application of Explainable Artificial Intelligence techniques in the service of Democracy.

In order to achieve the main goal, the following intermediate objective were defined:

- Perform a literature review on technology topics
- Perform a systematic literature review on the topic XAI state of the art
- Perform a study about the environment in analysis: Democracy
- Build the framework
- Display the challenges and advantages associated with the framework
- Validate and justify
- Gather all research findings
- Draw some conclusions

1.3 Study Importance and Relevance

This paper’s primary goal is to study the global potential of XAI, where this technology will be of major help, how it will influence the future of business and society and what the cons and pros are as a result of it. The project aim is to build a well-detailed framework for the application of XAI (Explainable Artificial Intelligence) and all its possible uses and advantages. This technology’s major goal is to add explainability to the models, to be able to explain its own output. This project’s outcome will, hopefully, contribute to a wide range of sectors from science to economy by adding explainability into AI models.

The relevance of the study lies in the many purposes of adding an explainability trait to ML algorithms, such as: improving system understanding, offer better behavior predictability and, naturally, increase trust in the system. Moreover, it can help build some bridges between AI and other social sectors such as health, justice and the Public Sector, for example. XAI has the potential to achieve all this and be a step closer for moving towards trustworthy AI (Markus et al., 2021). This paper intends to construct a framework that mirrors all these points plus the way it would serve Democracy.

There are many advantages to understanding how an AI-enabled system leads to a certain output. Not only to help developers guarantee that the algorithm is working as expected, to meet regulatory standards, or even to let the ones affected by an algorithm’s decision trial or modify that outcome for better fitting or understanding. Humans have the primary need of comprehending the decisions that directly impact them and the same happens with decision support systems (DSS) (van der Waa et al., 2021). To trust and rely on an algorithm or DSS, the system has to explain its decision to the user. Consequently, when accomplished, it maximizes the potentiality of AI to its fullest and can assist humans in many tasks/decisions in a much more efficient and fast-paced way than any other traditional decision tool would.

Essentially, Explainable AI has shown to be crucial for any organization to build trust and confidence. It can adopt a responsible approach to AI development with the aid of AI explainability trait. This would help solve the core issue that AI and other decision support systems (DSS) possess, the “black-box” nature,
where the user cannot explain and does not understand what happens in between the input inserted and the output obtained (Shin, 2021).

A positive outcome would essentially mean that the project was clear and helpful enough to support further studies regarding XAI and its potential for influencing future ways of handling data and make decisions by following the considerations done throughout the project. In sum, the development of these guidelines and best practices regarding Explainable Artificial Intelligence wish to offer this study field a clear and rich contribute with useful insights for boosting XAI state of the art.

*Figure 1: Explainable AI Concept* (Defense Advanced Research Projects Agency, 2021)
2 Methodology

For this project, the methodology chosen was the Design Science Research (DSR) to enable design theory and problem' identification about XAI state of the art. Thereby, find and gather crucial information in order to build a framework which aims to solve or help solving issues and gaps where XAI can be of major influence especially in the service of democracy.

The reason this approach was chosen is because:

• Design Science Research (DSR) is a research paradigm (Muntean et al., 2021) and involves the creation of an artifact and/or design theory as a mean to improve the current state of practice as well as existing research knowledge (Vaishnavi & Kuechler, 2021). For this project, the artifact is the conceptual framework.

2.1 Design Science Research (DSR)

Design Science Research (DSR) is a research approach that produces innovative artifacts and/or design theories as its research outcomes (Brendel et al., 2021). When it is focused on the creation of an artifact it involves two primary activities: (1) new knowledge creation through design or improvement of artifacts and (2) an analysis of the artifact’s utility and performance (Vaishnavi & Kuechler, 2021) within the problem context. This artifact being developed in DSR include algorithms, human or computer interfaces and system design methodologies/languages (Vaishnavi & Kuechler, 2021).

Design Science Research (DSR) is an approach mostly used in study areas such as engineering and computer science. Nevertheless, these days, this research paradigm is a rapidly evolving field and has extended to information systems science (Muntean et al., 2021). In information system science the design of an artifact can be systems, algorithms, methods, data models, data visualizations and others that could potentially contribute to the efficacy of IS in organizations (Peffers et al., 2018).

The main focus of Design Science Research (DSR) is to learn through the construction of an artifact. The artifact being constructed throughout this project is a conceptual framework. The DSR framework comprises three research cycles: relevance, rigor, and design (Brendel et al., 2021). These cycles enable a connection between design activities, related fields and practical environment, stimulating the convergence of design with requirements of real-world problems (Brendel et al., 2021).
The most referenced DSR model is the one proposed by Ken Peffers, Tuure Tuuanen, Marcus A. Rothenberger and Samir Chatterjee (2008). This design science research methodology (DSRM) process model is depicted above in Figure 2. This DSR methodology includes six steps: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication. Additionally, there are four possible research entry points: problem-centered initiation, objective-centered solution, design & development-centered initiation and client/context initiation.

A brief description of each DSR activity previously mentioned is given below:

- **Problem Identification and Motivation**: In this step, the research problem is identified and the relevance/importance of the study is defined. This is a crucial step since the problem definition is going to be used to construct the artifact that will reach a potential solution. Thus, giving a proper explanation of the solution’s value helps both the researcher to follow that solution and the audience of the research to accept it and better understand its purpose.

- **Definition of the Objectives for a Solution**: To define the objectives of a solution from the problem definition, previously done, and the knowledge gathered of what is achievable. We can have quantitative objectives, e.g., conditions where a desirable solution would be better than the present one; or qualitative objectives, e.g., having a new artifact’s description of how it is expected to support solutions for problems never addressed before. In sum, the objectives must follow the problem specification.

- **Design and Development**: Beginning the creation of the artifact in this activity. Conceptually, a design research artifact can be any designed object in which contribution is embedded in the design (Peffers et al., 2007). In this activity, it is supposed to establish the artifact’s functionality.
and architecture followed by the development of the actual artifact. Prior knowledge gathering is required, which can be applied in the solution.

- **Demonstration:** In this activity a demonstration of the artifact’s application is done to solve the problem or instances of it. This demonstration can be done in many ways such as simulations, experimentations, case studies or any other appropriate method. It is required to have effective expertise to use the artifact to solve the problem.

- **Evaluation:** The evaluation activity evaluates how the artifact is offering a solution to the problem defined. Evaluation includes comparing the objectives of the solution to the observed results of applying the artifact in the Demonstration activity. This activity can take many forms such as, simulations, feedback, interviews with experts, surveys, or any other suitable form of evaluation. In the end, the researcher is able to decide if it is needed to go back to the second or third activity to improve or complement the artifact to be more effective, or to pursue to the last activity of communicating the study and make some considerations for future projects.

- **Communication:** In this last activity the project is communicated to all relevant audience, and thus may include researchers or practicing professionals. The form of communicating the project depends on the audience and on the project goal.

### 2.2 Research Strategy

![Figure 3: Design Science Research Cycle (Hevner et al. 2004)](image)

Figure 3 depicts the research framework found in (Hevner et al. 2004) and it shows how it focuses on three research cycles and three spheres: the environment; the Design Science Research and Knowledge Base. The Relevance Cycle serves as bridge between the contextual environment of the research project and the design science research activities. The Rigor Cycle then connects design science research activities with the knowledge base in study. The Design Cycle iterates between the core activities of building and evaluating the design artifacts and processes of the research (Hevner, 2014).

In this project, the Knowledge base is focused on XAI Technologies and Techniques and the Environment it will be focused on is the Democratic System.
The following points further explain how each of the process steps were used throughout the development of this project.

- **Problem Identification and Motivation**: This step is fully based on the performance of a literature review on technology topics, such as Artificial Intelligence and Machine Learning and likewise about comprehension of the Democratic environment. This can be also seen as a step for problem recognition and help clarify at what length can the use of XAI be of service in Democracy as well as identify common ground.

  Literature review can broadly be described as a more or less systematic way of collecting and synthesizing previous research (Snyder, 2019). It is basically the way to start building the research while relating it with already existing knowledge and keep-up with the state-of-the-art research (Snyder, 2019). The main goal of doing a literature review is to gather all valuable information that already exists about the study area and identify ideologies, concepts, and theories that may be important for the development of the project. It was as well applied when searching for research methodologies and strategies to reach the most suitable one according to the type of work, problems and questions the project holds.

- **Definition of the Objectives for a Solution**: It is important to define a way to assess and match the XAI technologies with the democratic challenge it would help mitigate. Then, a conceptual framework is developed to reach the goal of being able to classify and map each democracy challenge with the suitable XAI technique.

- **Design and Development**: It is when the actual solution is developed, which means, when the creation of the artifact starts. To start, the artifact’s functionality is defined, as well as how the framework will be helpful, in practice: how XAI will be of service for democracy and which democratic/ethical problems the developed artifact may solve. A conceptual framework is designed, according to the previously defined architecture, to classify and map all identified issues in Democracy with the corresponding XAI solution. This enables the creation of new relations and concepts between democracy, its underlying problems and offers Artificial Intelligence as a technology solution by applying explainability.

- **Demonstration**: After the construction of the artifact it is necessary to evaluate it. The demonstration is made through several simulations and case-studies to where the framework would be of major help or where it could solve relevant issues by offering a logical and effective solution.

- **Evaluation**: For the evaluation of the proposed framework a set of interviews were conducted using pre-defined questions enabling us to focus on the important concerns. There were 4 inquiries who are specialists inside the many relevant scopes related with the topics addressed throughout this master thesis. It helped evaluate the utility, validity and improvements of the framework. With the interviews it was possible to obtain a range of different opinions and points of view which made the feedback richer.
• *Communication*: The project is communicated to the public, the consequent results are presented and further analyzed. Some considerations and limitations are also offered and are identified in order to contribute for future projects related to XAI state of the art.
3 Literature Review - Explainable AI

Machine learning models are often perceived as a mysterious and complex process, leading to the common term "black box" and many AI models employ advanced techniques to generate solutions that are difficult to understand or explain (Krishnan, 2020). As we see AI moving into all aspects of our personal and business environment, the transparency conversation is growing in importance (Sathe, 2021). Decisions derived using AI models will need to be traced, challenged and immune to bias. In other words, organizations need to embrace XAI and ensure that model output can be explained appropriately to all stakeholders (Sathe, 2021).

3.1 Obscurity and bias in AI

Technology is evolving at an extremely fast pace and with it, AI technologies are inherently growing the same way. This evolution can be observed from now having self-driving cars, to intelligent houses commanded by voice or virtual assistants (Steve Nouri, 2021). Many other things can be made possible through AI potentialities and so many are already in study. However, as ML models are becoming more widely adopted, there are also gaps being found in these technologies (Anupam Datta, 2021).

Some notorious issues are facial recognition technologies, health care and discrimination. All caused because of bias in AI systems (Mary Reagan, 2021). There are many ways of having biased data or algorithmic AI bias, and while there are cases where it is unintentional, other cases may be intentional or at least caused by natural assumptions and behaviors - Societal AI bias (Vitor Santos, 2022). Human bias is an issue heavily studied in many areas and forecasted to continue to be inherently natural to humans in an unconscious way (Steve Nouri, 2021). Bias does not appear in a dataset alone, it has many external or internal factors contributing to it such as problem formulation, model trained based on bias data, lack of data, unsupervised learning, many others could be added (Anupam Datta, 2021).

There are several reasons for reaching biased data. Some of the main types of bias are as described:

- As previously mentioned, there are AI algorithms in which unsupervised machine learning is applied. This enables AI to self-learn and looks for trends and conclusion training the algorithm using raw unlabeled data. By contrast, supervised learning trains AI from pre-labeled and structured data and tests if it can reach the correct output (Aminah Aliu, 2021). Giving AI the power to draw its own conclusions from unsorted data leaves many spaces for error and vulnerabilities. Bias can be created through other users’ interaction reflecting it on the data collected and then on its output (Joe McKendrick & Andy Thurai, 2022).

A real-world example of this type of cases is when Twitter users taught Microsoft’s chatbot Tay to be racist, vile, and offensive in less than 24h after its release (James Vincent, 2016). This happened in 2016, when Microsoft designed Thay in an attempt for it to engage, entertain people and learn from them in an experiment that combined machine learning, natural language processing and social networks (Ken Bryden, 2022). However, it quickly backfired from the point where Tay started to tweet highly offensive and abusive things from what it had learnt which was a huge scandal that forced Microsoft to suspend the account (Oscar Schwartz, 2019).
• In the example above, the presence of another form of bias is observed - **historical bias** - which is a type of bias that exists naturally in society and that is then reflected in data collected. This does not mean poor input data or feature selection, but it tends to show more on historically disadvantages groups/regions/products/concepts (Mary Reagan, 2021).

• Still in the first example **Algorithmic bias** is the second form of bias observed. This type is not present in the actual data but is created by the algorithm (Belenguer, 2022). Here, due to the high existence of hate speech on social media, logically the algorithm innocently learnt from it and replicated it. However, there are several causes that may affect the algorithm, either inadvertently or deliberately generating disparities between groups or categories. Some are as simple as Bias in online ads, in word association or in facial recognition, and the 2 major causes are historical human biases and incomplete/unrepresentative training data (Nicol Turner Lee et al., 2019). Other inherent type of biases such as discrimination towards certain races, salary disparities that frequently associate men with higher-paying jobs, and the disbelief/judgment of people with disabilities are what are contributing for the existence of distorted and unfair applications of algorithms, observed from language translators to online recruitment filtering tools (Mitra Best & Anand Rao, 2022).

• Even though it is positive that machines have the power to enter and learn from human knowledge it is at the same time negative as it might be naively contributing for the perpetuation of harmful human biases (Aminah Aliu, 2021). In line with this thought, **Representation bias** happens due to how the universe of data needed to create a dataset is defined and sampled (Mary Reagan, 2021). Many studies have shown how representative bias is so prevalent in AI. Technologies like facial recognition work almost perfectly with any citizen with a light skin tone while with Black, Hispanic, and Middle Eastern it frequently fails to do recognition (Kyle Wiggers, 2020). Another example would be collecting data to create a dataset online or through a smartphone, this would end up giving biased data since not all the universe of people would be able to respond such as lower-income people with no power to own a smartphone or elder people who don’t know how to function so well with those technologies (Mary Reagan, 2021).

While it is a huge challenge to completely overcome and eliminate bias in AI, there are some preventive actions that can be made already (Kyle Wiggers, 2021). One of the very first steps should be to fully understand all types of bias and apply all best practices to recognize and overcome them. Subsequently we will be a step closer to reach the so called “algorithmic hygiene” (Nicol Turner Lee et al., 2019).
3.2 XAI concepts

Understanding the logic behind AI inner workings is of great urgency for the continued expansion of AI in many relevant decision-making processes. This is where Explainable Artificial Intelligence (XAI) enters - this recent method allows users to better understand and trust the outputs of an AI powered system (Vinothkumar Venkataraman, 2022). As stated in Medium from a study made by PWC, the majority (82%) of CEOs agree that for any AI based decision to be trusted, it must be explainable (Tech in 3, 2022).

The main purpose of XAI is to apply its techniques and processes to help users, developers and organizations reach a new level of transparency in the results obtained by justifying its decision/prediction (Ajitesh Kumar, 2022). For better understanding concepts commonly used inside XAI scope, it follows a brief clarification of the most common terms:

- **Explainability**: The system itself can provide an explanation or evidence for a reached decision/prediction (Phillips et al., 2021).
- **Transparency**: If a model itself is understandable then it is transparent. There are three different transparency degrees – simulatable models, decomposable models, and algorithmically transparent models (Barredo Arrieta et al., 2020).
- **Understandability**: A user can understand a learning model’s function or, in other words, how the model works but without any additional explanation about learning inner workings (Meet Gandhi, 2020).
- **Comprehensibility**: How a Machine Learning algorithm can represent what it learns in a way understandable for humans, this level of comprehensibility is related to the valuation of the model complexity (Barredo Arrieta et al., 2020).
- **Interpretability**: Allows to understand what the model is learning and the motives to why it reached a certain decision (Ajitesh Kumar, 2022) and to the extent where it makes sense to the user (Meet Gandhi, 2020).
According to the US National Institute of Standards and Technology (NIST) there are four principles for Explainable AI:

1. **Explanation** – Every output the AI system produces, must come accompanied by an explanation or, at least, provide supporting evidence. There are several types of explanation (Nadejda Alkhaldi, 2021).

2. **Meaningful** – The explanation received must be meaningful so that it helps individual users understand what to do (Giri, 2022) which can mean providing several explanations if there are users with different expertise.

3. **Accuracy** – The explanation must correctly translate the system’s manner for reaching the output (Vitor Santos, 2022), in other words: clear and accurate.

4. **Knowledge Limits** – The system should only function within the conditions for which it was designed or when a certain confidence level in the output is reached (Giri, 2022).

These principles intend to reach a wide set of perspectives and reasons in use cases of AI. As mentioned above in principle 1. Explanation, there are distinct types of explanations that put XAI inside multiple dimensions. Those are:

- **Explanation that benefits the end-user**: When it explains to an individual user why a decision/prediction is made by the ML algorithm that concerns them (Vitor Santos, 2022).

- **Explanation for social acceptance**: Is designed to gain trust and social acceptance in the system and in the decisions made by it (Nadejda Alkhaldi, 2021).

- **Explanation for meeting regulatory and compliance requirements**: It responds to the need of providing, for example, to a safety regulator or to an audit, significant details about the process behind the machine (Vitor Santos, 2022).

- **Explanations to help system development**: to be able to find aspects to improve, debug and maintain to encourage algorithm/system development and maintenance. Technical staff, product managers and executives are examples of possible consumers (Jonathon Phillips et al., 2020).

- **Explanation for owner benefit**: simply for the system’s owner to gain user trust in the machine while benefitting from the model (Vitor Santos, 2022).

### 3.3 Approaches & Algorithms

The explain ability in XAI is bringing many Machine Learning models a contextual reasoning (Vitor Santos, 2022). But how AI works depends on the type of approach that is used. According to the NIST there are three broad algorithms to explain AI:

**Self-Explainable Models** – The most common where models themselves are the provided explanation (Vitor Santos, 2022). Some simple examples are Rule-based methods such as Decision Trees and Linear and Logistic Regression models. Even though these are self-explanatory models, they are not always accurate especially if there is a mismatch between inputs and the required statistical properties. Due to these factors, there are already studies in progress to reach models that can be self-explanatory and simultaneously accurate.
To improve standard decision trees there is a work called decision lists that is presented through nested sequence of “if-then-else” rules. It can be a simple concept but considered hard to interpret and still inaccurate. To overcome the issue of the above technique, Decision sets are developed as an improvement and simplification of Decision lists. These are a sequence of “if-then” rules with a single “else” statement at the end, where each clause is a conjunction of conditions. This last “else” ensures every instance is not left to classify. Here it is shown to improve accuracy and make it easier to interpret. Regardless, Bertsimas and Dunn further improve accuracy while remaining transparent producing a variant of decision trees named optimal classification trees.

Global Explainable AI Algorithms – This works by querying an AI algorithm as if it is a black-box to then produce a separate model able to explain the algorithm. One that is commonly used is SHAP (SHapley Additive exPlanations). Based on research by Nobel Prize-winning economist Lloyd Shapley, SHAP works by applying the principles of game theory (Giri, 2022). In a regression model, each feature can be considered to interact with other features in the model in competition for the model’s output (Giri, 2022). In sum, in SHAP the regression output is treated as a “coalition” game and each feature is a player that can either participate or not against the other features in each row. It then computes the Shapley values to explain the contribution of each player to the model’s output. In more complex systems like deep neural networks, the most widely used approach is TCAV (Testing with Concept Activation Vectors). It basically works in a simpler way, representing neural network state as a linear collection of Concept Activation Vectors (CAVs) and has been showing reliable results when applied to explain image classification algorithms. Additionally, Global explanations can use visualization techniques that are Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE). In the first one, when a feature is changed it shows marginal changes in the model’s output, whereas in the second marginal changes are noticed at a more granular level (Giri, 2022).

Per-Decision Explainable AI Algorithms – Provide a separate explanation for each decision and are considered local explanations (Vitor Santos, 2022). Similarly, to Global XAI algorithms, these are also queried as a black-box system this time looking to explain one decision output of the model. One well known example is Local Interpretable Model-agnostic Explainer (LIME), it explains the model’s features for a single decision, querying the model’s output. Thus, LIME uses the a priori built decision-specific representation of the model to then deliver explanations. Another popular approach of local explanations is Counterfactual explanations. In this approach, if the inputs are changed it studies the impact of it on model outputs. The name explains exactly what it does, if inputs are changed to an extent that contradicts the observed outcome, it then goes explore and look for insights about model’s input and outputs relationship for that specific decision. It basically works to evaluate the existence of casual relationships (Giri, 2022). A huge advantage of this approach is that it tends to be understandable for the general audience.

Additionally, in Machine Learning, the tasks can be categorized into supervised, unsupervised learning and semi-supervised learning (Loukas, 2020). The different types of learning are as followed:

Supervised Learning – These models learn from the labeled dataset and are then used to predict future events (Loukas, 2020). Supervised models can be further classified as either regression or classification cases, depending on the type of output variable being used. If the output variable is a real number, whether it is discrete or continuous, the problem is a regression case. If the output variable is categorical,
then the problem is a classification case (Hshan.T, 2020). It is important to choose the right algorithm to match the specific type of learning problem that is being worked on.

**Unsupervised Learning** – This method involves studying how a system can identify hidden patterns and relationships from unlabeled data in order to infer a useful function or representation (Hshan.T, 2020). Unsupervised models can be further grouped into clustering and association cases. Clustering is one method used to group together tuples based on their shared characteristics or similarities. Association, on the other hand, focuses on discovering hidden relationships or rules that exist between tuples (Loukas, 2020).

**Semi-Supervised** - Is a merge of supervised and unsupervised learning, it uses both labeled and unlabeled data for training (Loukas, 2020). The majority of data available today is unlabeled due to the high cost, time, and expertise required to manually label large datasets. In the case of unlabeled data, unsupervised learning techniques are typically applied first to uncover patterns and group data points. The labeled data can then be used for supervised learning (Hshan.T, 2020).

Overall, XAI techniques are a useful tool for increasing the transparency and interpretability of AI models. By providing explanations on how AI models work, XAI techniques can help build trust in AI and ensure it is used in a way that is fair, ethical, and accountable. The approach choice will depend on the specific use case and the nature of the AI model being explained.

### 3.4 XAI techniques in Democracy

Explainability is a crucial aspect of AI that allows systems to be understood and audited. There are various techniques to achieve explainability in AI systems, such as rule generation, model interpretation, sensitivity analysis, and others. A brief description of the mentioned systems is as followed:

- **Big data analysis, Cognitive services**: Through the analysis of massive amounts of data, it is possible to infer useful information about trends and preferences. This tool is being deployed in both business and governmental entities. An example is the use of past criminal activity to predict future criminal activity which is known as predictive policing (Anastasiadou, 2018).
- **Textual data, Automated sentiment analysis**: Unstructured information present in various free-form text fields such as blog posts or product reviews (Anastasiadou, 2018).
- **Anomalies detection, Fraud detection**: A technique that uses available data to establish a baseline and alert of deviations of normal behavior. These are used to flag uncoherent or uncommon datasets for further analysis. It is a standard tool being deployed by various tax enforcement agencies (Anastasiadou, 2018).

In democratic institutions, the application of explainability techniques can help increase transparency and accountability of decisions made by algorithms. This is particularly relevant in areas like judicial/legal decision-making, where decisions must be fair and transparent. Some other possible applications of XAI techniques in the service of Democracy would be:
1. Voting systems: XAI techniques can be used to provide transparent explanations for how votes are counted, and results are determined. For example, XAI can be used to explain how electronic voting machines work and how they ensure the integrity of the voting process (Fowler, 2013). This can help to prevent errors and fraud and ensure that voters trust the integrity of the voting system.

2. Predictive policing: XAI can be used to provide transparent explanations for how predictive policing algorithms make decisions. This can help to reduce bias and ensure that police decisions are based on accurate and fair information (Vestby & Vestby, 2019). For example, XAI can be used to explain how predictive policing algorithms identify crime hotspots and how police use this information to prevent crime.

3. Government decision-making: XAI can be used to provide explanations for how government decisions are made. This can help to promote transparency and accountability in the decision-making process and ensure that decisions are based on accurate and reliable information (Leben, 2023). For example, XAI can be used to explain how government agencies make decisions about public health policy or environmental regulations.

4. Social media moderation: XAI can be used to provide explanations for how social media platforms moderate content. This can help to promote transparency and fairness in content moderation decisions and ensure that users understand why their content was removed or flagged (Gongane et al., 2022). For example, XAI can be used to explain how social media platforms use machine learning algorithms to identify hate speech or misinformation.

5. Ethics in AI decision-making: XAI techniques can be used to ensure that ethical considerations are incorporated into AI decision-making. By providing explanations for how AI algorithms make decisions and identifying potential biases, XAI can help ensure that AI systems are fair and transparent (Daffner, 2021).

### 3.5 Applications of XAI techniques

There are several different XAI techniques that can be applied to explain how AI models make decisions. By providing explanations for the behavior of these systems, XAI techniques can help address a variety of issues, such as bias, discrimination, and errors.

One of the main benefits of XAI techniques is that they can help identify the features/inputs that are driving the output of the ML model. This can be particularly useful in cases where the model is making predictions which are difficult to interpret or explain.

In the following table are some of the most common techniques with the identification of relevant industries it could be applied followed by a specific example:
<table>
<thead>
<tr>
<th>Approach</th>
<th>Description</th>
<th>Industry</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local interpretable model-agnostic explanations (LIME)</td>
<td>Particularly useful to explain predictions made by complex models like black-box models (Holzinger et al., 2020a).</td>
<td>Healthcare, Recommendation systems</td>
<td>Individual patient outcomes need to be explained to doctors and other medical professionals. Provide explanations for why certain products or services were recommended to a user.</td>
</tr>
<tr>
<td>Partial dependence plots (PDP)</td>
<td>Useful for identifying the contributions of individual features to the predictive value of a black-box model (A Brief Overview of Methods to Explain AI (XAI), 2021).</td>
<td>Finance, Healthcare, Transportation</td>
<td>A credit risk model may use PDP to identify which factors, such as age or income, are most important in determining a person's creditworthiness.</td>
</tr>
<tr>
<td>Integrated Gradients</td>
<td>Particularly useful for identifying key features in deep neural network-based models (Holzinger et al., 2020a).</td>
<td>Image classification, Natural language processing, Drug discovery</td>
<td>An image classification model may use Integrated Gradients to identify which pixels in an image are most important in determining the object in the image.</td>
</tr>
<tr>
<td>Shapley Additive explanations (SHAP)</td>
<td>Can provide model-agnostic local explainability for tabular, image, and text datasets (Aditya Bhattacharya, 2022).</td>
<td>Healthcare, Finance, Transportation</td>
<td>Given a patient's medical history and test results, SHAP can generate an explanation that highlights the factors that contributed to a certain diagnosis, such as the patient's age, blood pressure, and cholesterol levels.</td>
</tr>
<tr>
<td>Counterfactual Explanations</td>
<td>Generate explanations for why a particular outcome was not achieved and help identify areas for improvement in a model, such as identifying the features that would need to be</td>
<td>Finance, Healthcare</td>
<td>Given a patient's medical history and treatment data, Counterfactual Explanations can be used to generate an explanation that shows what changes in treatment would have</td>
</tr>
</tbody>
</table>
changed to achieve a desired outcome (Byrne, 2019).

Anchors  
Generates "if-then" rules that describe the conditions under which a particular prediction is made. It can be used to generate explanations for any model, including black-box models like deep neural networks (Holzinger et al., 2020b).

- Law
- Healthcare
- Finance

been necessary for the patient to respond.

Given a loan application and the model's decision, Anchors can generate an explanation that highlights the factors that influenced the decision, such as the applicant's credit score, income, and employment history.

| Anchors | Generates "if-then" rules that describe the conditions under which a particular prediction is made. It can be used to generate explanations for any model, including black-box models like deep neural networks (Holzinger et al., 2020b). | Given a loan application and the model's decision, Anchors can generate an explanation that highlights the factors that influenced the decision, such as the applicant's credit score, income, and employment history. |

Table 1: Example of possible applications of XAI Techniques
4 Environment - Democracy

4.1 Definition

The current and literal meaning of Democracy is as followed: “the belief in freedom and equality between people, or a system of government based on this belief, in which power is either held by elected representatives or directly by the people themselves” (Cambridge University Press, 2023). However, since Democracy was born, in the time of the ancient Greeks, it has suffered huge changes from its primary form both in theory and practice (Robert A. Dahl, 2022). It is widely interpreted only as a political movement or form of government.

Democracy is rather hard to define since there are actually so many different democratic governments worldwide. There are two most common types of democracies: direct and representative, but there are other variants that can be observed nowadays such as: participatory, liberal, parliamentary, pluralist, constitutional, and socialist democracies (Robert Longley, 2021). Nonetheless, in theory, Democracy should be for everyone including minorities and therefore, should not be exactly the majority ruling if that means that minorities’ interests are ignored completely (Council of Europe, 2023). Additionally, all forms of democracies are based on the ideology of competitive elections, freedom of expression and protection of individual civil liberties and human rights (Robert Longley, 2021) translated in the six foundational elements common to every democracy:

- **Popular sovereignty**: The concept that the maintenance of the governance is made through election of representatives chosen by the people.
- **An Electoral System**: According to the previous principle, people hold the power of political choice. Therefore, it is necessary to have a clear, fair, and free electoral system. Which translates in the following rights (Civics Academy, 2015):
  - All adult citizens have the right to vote.
  - Elections are held at regular, known intervals.
  - Elections are free and fair.
  - Majority rule.
- **Public Participation**: The only way Democracy stays alive in a society is by the active participation of its people. A good and healthy democratic system should encourage and offer every means needed for people to participate and/or take part in political and social activities such as elections, initiatives, and unions.
- **Separation of Powers/Multi-party system**: The main concept of having a Democracy, besides the people being able to choose its representative, is to avoid power concentration in one single person as in a monarchy. It allows different representations inside the government among several parties.
- **Human Rights**: Along with all the freedoms Democracy offers and defends, it also protects all citizens human rights. This means rights’ protection for all human beings no matter the sex, nationality, color, religion, or any other aspect. These rights include (Civics Academy, 2015):
  - Right to vote.
  - Freedom of speech and assembly.
  - Freedom of religion.
- Right to equality.

- A Rule of Law: It is the natural necessity of order and harmony in society, which means, having a principle that citizens are to be accountable to laws and rules. This is crucial to guarantee no individual crosses or oversteps other peoples’ rights and liberties. These laws “(...) are publicly created and equitably enforced in a manner consistent with human rights by an independent judicial system.” (Robert Longley, 2021).

4.2 History

Even though there are some earlier examples of primitive democracy in other parts of the world, Greece is known as the country where Democracy was first observed. This first Democratic Greek model was established in the fifth century BC, in Athens (Council of Europe, 2023) and it had three different institutions:

- The Ekklesia:

  “A political assembly of citizens for conducting public business and for considering affairs proposed by the council.” (Caudle, 2020) It is also the body that wrote the laws and ruled foreign policy (History.com Editors, 2019). The meetings were held forty times per year and any member of the demos (adult male citizen) was welcome to attend it. At this meeting, decisions about wars and foreign policy were made, laws were written and/or approved and, when in case, the sentence of expelling a citizen from Athenian city-state for 10 years - ostracism - was also conducted. The assembly made decisions through a simple majority vote (History.com Editors, 2019).

- The Boule:

  “An advisory citizen body” (N.S. Gill, 2018) where members had to be over 30-year-olds. It was composed of representatives from all ten Athenian tribes equally distributed (History.com Editors, 2019). The meetings happened daily and had a very “hands-on” work in terms of governance. The members of these institutions were the ones deciding which affairs came before The Ekklesia and how the whole democratic system worked (History.com Editors, 2019).

  The idea of a pure democracy could be observed even in the way members were chosen to be part of The Boule: they were elected randomly by lottery and not by election. They believed this way the chosen ones would be selected without any bias such as money or popularity. Despite that, it was registered that some individuals from influential/wealthy families were chosen far more frequently than it would be in a true random lottery (History.com Editors, 2019).

- The Dikasteria:

  The Dikasteria is the definition of a popular court where citizens defended their cases before a council of randomly selected jurors. These jurors were selected daily and had to be male citizens older than 30 years old. Back in these days there were no police or judicial control, therefore, it was the
people themselves that brought cases to the Dikasteria, defended them, and reached verdicts through the majority votes rule (History.com Editors, 2019).

The word “Democracy” is originated from the Greek words: “demos” and “kratos”, meaning people and power, respectively. It can be translated as “power of the people”, a system that is fully conditioned by the will of its people (Council of Europe, 2023). This definition was later redefined by the U.S. former President, Abraham Lincoln, in 1863 as a “…government of the people, by the people, for the people…” (Robert Longley, 2021).

During this time and for several centuries, since tribes and cities remained small, the only form of democracy they would practice would be direct democracy. It was only when cities developed and grew into larger cities/states more populated and civilized, democracy gave space for a new form: representative democracy. This evolution brought some necessary changes to the existing political institutions into new and different ones – legislatures, parliaments, and some political parties - to match the cultural character of the city/state/country (Robert Longley, 2021).

It was not until the end of the 17th century, during the English Civil Wars, that some more drastic changes were conducted and transformed into the existing Democracy model. What happened was that Members from a radical movement started demanding more representation inside the Parliament and the right to vote was extended to all male citizens.

As the evolution of society followed a fast pace, political systems began to shift too. Two democratic models arose – presidential model mainly in Americas, which was governed by an executive; and a parliamentary model mainly in Europe, led by the legislature - and countries shifted between these two systems (Patrick Manning, 2016). A swing between monarchies and democracies was also observed, where this first one gradually fell into disuse even though, nowadays, there are still some monarchies. However these have changed from an exclusive source of power under hereditary rulers to a system more like democracy where the monarch is now more of a “ceremonial head of state” (Patrick Manning, 2016).

The increase in people’s awareness about social matters, made the fight for “human rights” grow exponentially, beginning in the 19th century. After that, all debates lacking in the initial concept of democracy start rising, such as some about: fairness and inclusivity and all types of discrimination, whether by race, ethnicity, nationality, or religion (Patrick Manning, 2016). All these values started to be included in a democratic model as intrinsic.

As previously seen, Democracy is more than a political system with political institutions/parties, there are different interpretations of it and ways of applying it.
4.3 Challenges and Problems in Democracy

While democracy can take many forms, its core principles remain the same: transparency, accountability, participation, and the protection of individual rights and freedoms. While democracy is a highly desirable form of government, it also holds many challenges. In the following paragraphs some challenges are pointed out and briefly explained.

One of the main challenges is the potential for majority rule to turn into "majority tyranny", or tyranny of the masses, where the majority group imposes its will above the one of minority groups or individuals (MasterClass, 2022). Maintaining the balance between majority rule and protection of minority rights is a huge challenge which can potentially lead to the oppression of those who do not share the same views as the majority and, thus, go against principles Democracy stands for such as equality.

Another great challenge for democracy is Political polarization, as different political parties and interest groups compete for power. If the political climate becomes too polarized, it can lead to gridlock and an inability to make decisions that benefit the majority of citizens. (Carothers & O’Donohue, 2019). It also contributes to the deterioration in everyday interactions and social relationships by creating a heavy separation between people with different political ideas. As such, it is also an entry door for the rise of hate crimes and political violence (Carothers & O’Donohue, 2019).

Perhaps the most obvious threat to democracy lies in the existence of corruption - if there is corruption within the political system, elected officials may prioritize their own interests or those of special interest groups over the needs of the citizens. It can, therefore, undermine the legitimacy of democratic institutions and processes, deteriorate trust in government, and weaken public confidence in democracy as a viable form of governance. It can also prevent effective policymaking and lead to the misallocation of resources, further intensifying social and economic inequalities. (International, 2021). Overall, corruption poses a significant threat to the health and sustainability of democratic governances.

A factor that is both a challenge and a consequence of other drivers is Inequality. While democracy is based on the principles of equality and justice, inequalities can still persist in a democratic system. This can include economic, social, or racial inequalities that limit the ability of some groups to participate fully in the political process (Cox et al., 2017). People may be excluded from democracy in indirect or informal ways - some groups in society are denied the protections and resources necessary to participate. Examples of informal limitations can include: voter intimidation or harassment of particular groups, unequal access to justice, and lack of access to resources that are necessary for participation, such as time, money, healthcare, or education (LINDBERG, 2019). These factors can make it difficult or impossible for certain individuals or communities to have a meaningful say in the political process and can result in their voices being marginalized or ignored.

Another major issue is voter apathy, where citizens may not exercise their right to vote or participate in the political process. When a significant share of the electorate does not participate in elections or other political activities, it can undermine the legitimacy of the political system and weaken the representation of diverse perspectives and interests. It can as well lead to a disengaged and uninformed citizenry. (University of Essex, 2021). There are many potential reasons for voter apathy, including a lack of trust in government or political institutions, feelings of disenfranchisement or powerlessness, and a belief that individual votes do not matter or will not make a difference. (University of Essex, 2021).
Climate change more than ever is affecting people’s lives globally. The climate crisis is truly a global challenge as it is caused by and affects the entire human population (Lindvall, 2021). It is exacerbating inequality, increasing poverty and food scarcity, displacing populations and exacting a direct toll on countries’ democratic aims. Strong responses to the climate crisis require strong democratic processes (Inter-Parliamentary Union, 2022). Therefore, this represents a significant threat to Democracy and to its application in most affected areas by climate changes.

While democracy may not be perfect, it remains one of the most effective forms of government for promoting individual freedoms, protecting human rights, and promoting social justice. It is important to address these challenges in order to strengthen democratic institutions and ensure that they truly serve the needs and interests of all citizens.

4.4 Challenges and opportunities of using AI in Democracy

“AI and automated-decision making (ADM) systems are already being deployed all over Europe and are actively influencing citizens’ lives.” (Laureline Lemoine, 2020).

Artificial Intelligence is already a hugely relevant tool for humanity and for the future. In such that it can be considered as being one of the best ones to be able to preserve our freedom (Khari Johnson, 2019). At the same time, this type of technology is already transforming individuals lives and communities (Rik Daems, 2020). It is natural that AI presents challenges and opportunities, some of which will be discussed below.

4.4.1 Challenges

It is undeniable that the rapid evolution of Artificial Intelligence and its inherent social influence, turned out to be crucial for the future of democracies and how they will adapt in the forthcoming years. It is essential to debate and consider the positive outcomes but also the challenges intrinsic to this type of technology. With this comes several different questions – Is Artificial Intelligence a danger for democracy? Can Artificial Intelligence be a useful tool for improving or enhancing democratic models? (School of Transnational Governance, 2022) Will Artificial Intelligence ever be used to cause bias and lead people wrongly into a decision? – and many others could be added.

One of the biggest challenges or even threats to democracy, lies in the power held by AI to “shape individuals’ decision-making processes” (Serbanescu, 2021) or, in other words, to manipulate them. It can be seen as a huge challenge since it would directly affect democratic processes by potentially forcing an influence over citizens. This influence can be intentional or unintentional depending on how it is programmed and the data on which it is based. A wrongly programmed algorithm can create biased data and therefore offer misleading information that will lead to wrong assumptions/decisions.

Worldwide technologies’ use has risen along the years and is continuing to become part of our daily routines. This means we are increasingly consuming more internet content, and it inevitably shapes our
way of thinking and our perception of the world. With this increasing use of AI technologies, we are also exposing and sharing more private and personal information to companies and applications, which is therefore facilitating the manipulation of our decisions and shaping our choices into the desired ones (European Parliament, 2022). With this out sharing of information, there are also concerns about privacy and data protection of individuals. Nowadays we are even giving our facial features and biometric data to AI applications for unlocking purposes for example, or to save passwords and register faster in an app - either through fingerprint or facial recognition. All this data is stored and used by all companies we give consent to that can then track and profile us even better, to later influence our choices and use the data gathered to lead us to behave in a certain manner.

The above mentioned is also connected with another challenge, since based on data gathering and collection, companies target us to certain products that they found, after profiling and register a search or consumption patterns, which would be more to our liking. This way they increase the chances of an effective purchase, making our lives easier by only showing us what we seem to like or be more interested in. Even though this sounds convenient, it is also a threat to democracy and equal access to information and products, narrowing an individual’s choice at first. Adding this to all other capabilities of technologies, such as creating fake content like images and videos that represent a serious risk to decision making, such as frauds (European Parliament, 2022).

General Data Protection Regulation (GDPR) represents a huge challenge for all AI tools. While it does not explicitly reference AI, many of the regulations outlined in the GDPR are applicable to AI, and some may face challenges due to the new methods of processing personal data made possible by AI (Sartor & European Parliament. European Parliamentary Research Service. Scientific Foresight Unit., 2020). There is a noticeable conflict between traditional data protection principles and the unrestricted utilization capabilities of AI and big data. When personal data is being processed by AI systems, data protection authorities enforce regulations such as the GDPR, EUDPR, LED, and other sector-specific legislation essentially to protect individuals’ fundamental rights to privacy and the safeguarding of personal data, which are often intertwined with other essential rights such as the right to dignity, fair trial, and an effective judicial remedy (European Data Protection Supervisor, 2023).

4.4.2 Opportunities

It is normal to doubt the “unknown” and highlight all the possible threats and challenges an innovative technology might present. But it is also necessary to remember that the technology itself is not “inherently harmful (Polonski, 2017)”. The same algorithm that could be used to generate biased and misleading information can also be used to leverage democracy. Moreover, it is undeniable that AI’s ability to manage vast amounts of data and information, to identify patterns and to have trackable inputs makes this an increasingly crucial tool for this technology driven world (Thiel, 2022). Other key opportunities to highlight are explained below.

  
  Fairness and transparency
  
  When used and programmed correctly, AI is not biased and will base its output on the input data, which should always be true and correspond to reality. This way, data-driven decision-making processes
would enable better and fairer decisions and conclusions (Anastasiadou, 2018). Similarly, it can also enable a better access to information, education, and training to the public (European Parliament, 2022).

Moreover, since an AI algorithm is not susceptible to corruption or to being dishonest like humans are, it will always be a more reliable tool to apply all defined rules constantly with no exception. Besides, the outcome is always trackable through the parameters that were set. This means, if the set of rules or parameters are made public, any citizen has the power to know and understand the decision reached by that algorithm and therefore offer a transparent decision-making process (Anastasiadou, 2018).

**Strengthening democracy**

With such an enormous potential, AI can even support Democracy and make it stronger. It could be used to run better electoral campaigns, for example, by better informing and serving the citizens (Polonski, 2017). By using data-based analysis it is possible to prevent misinformation and even cyber-attacks while guaranteeing reliable information (European Parliament, 2022). This would result in a well-informed population and therefore, lead to better and more conscious decision making.

Elections are an easy and direct example of where AI could be supportive. Since it has an historical capacity to then understand an individual's preferences, it can easily help citizens decide on their vote. While making the process easier, it encourages and hopes to increase participation (Frey, 2020) which would only help strengthen democracy and drive people to exercise this right.

While this ethical use of AI can be used to inform and serve citizens, it may also be applied to fight misinformation by blocking or flag articles, websites or posts that contain false information (Polonski, 2017). It can then prevent the spread of incorrect information that may influence citizen's opinion and decision making.

**Security and safety**

AI potentiality in the service of security is boundless as it could help prevent crime and assist criminal justice system. How? By processing in a much efficient way huge data sets, assess risk more accurately and even by applying predictive methods to avoid and estimate crimes to happen (European Parliament, 2022). This capability is already used for cyber-security reasons to detect vulnerabilities and protect entities from hacking or phishing attacks. This means stronger and safer communications and information systems security (Sanchez, 2017).

AI is already being used to avoid dangerous and repetitive tasks once performed by humans. In several industries, many risky tasks previously executed by a human are now replaced by a robot or an automated machine programmed to repeat that action (Wizata, 2022). This has been helping decrease the risk of injuries at work and prejudice of human health, while being more efficient and safer.
5 Conceptual Framework

XAI (Explainable Artificial Intelligence) techniques are increasingly used in decision-making processes in several domains, including politics and governance. The application of XAI techniques in the service of democracy is going to prioritize the use of transparent and interpretable models, promote fairness and reduce discrimination, facilitate collaboration between humans and AI systems, protect data privacy, and provide explanations for deep learning models. By doing so, XAI techniques can help support and enhance democratic decision-making processes while upholding democratic principles such as fairness, accountability, and citizen participation.

5.1 Assumptions

Based on the literature review about Democracy and XAI technologies and its applications, explored in the previous chapter, here are drawn some assumptions that will be then considered in the construction of the framework:

- A1 – XAI technology’s focus is to add explainability to the models in order to explain its own output. By adding an explainability trait to ML algorithms it can potentially improve system understanding, offer better behavior predictability and increase trust in the system consequently.

- A2 – Democracy can take many forms, but its core principles lie always in: transparency, accountability, participation, and protection of individual rights and freedoms. It is a highly desirable form of government however it also holds many challenges. Some may benefit or even be overcome from the application of XAI.

- A3 – Artificial Intelligence is already a hugely relevant tool for humanity and for the future. In such that it can be considered as being one of the best ones to be able to preserve our freedom (Khari Johnson, 2019). At the same time, this type of technology is already transforming individuals lives and communities (Rik Daems, 2020).

- A4 – There are various reasons for biased data. Some of the main types of bias are unsupervised machine learning, historical bias, Algorithmic bias, and Representation bias.

- A5 – The explain ability in XAI is bringing many Machine Learning models a contextual reasoning (Vitor Santos, 2022). But how AI works depends on the type of approach that is used. According to the NIST there are three broad algorithms to explain AI: Self-Explainable Models, Global Explainable AI Algorithms and Per-Decision Explainable AI Algorithms.

- A6 – In democratic institutions, the application of explainability techniques can help increase transparency and accountability of decisions made by algorithms, some examples: Voting systems, Predictive policing, Government decision-making, Social media moderation and Ethics in AI decision-making.
5.2 Proposal

The assumptions mentioned in the previous section were used to conduct the construction of a framework proposition to help apply XAI Technologies in the service of Democracy.

The framework aims to support the identification of ways to mitigate democracy challenges and threats, pointing out eventual conditions and restrictions, preceding the selection of all suitable XAI techniques for a certain context/problem.

For a possible application of the developed framework, it is crucial to gather a multidisciplinary team that will define collaboratively every step until the desired outcome. This means having professionals from the scope of technology, but also politics, governments, Human rights, NGO’s, decision makers or researchers in the area of democracy for an accurate assessment of every phase.

Figure 5 represents the key phases for the construction of the artifact.

![Figure 5: Framework for the selection of XAI techniques in the service of Democracy](image)

5.2.1 Democracy challenges Assessment

As previously explored, during the literature review, there are some fundamentally important threats to Democracy as we live it. To acknowledge those key challenges that may be potentially threatening Democracy is the first step in this framework. These challenges vary and can be differently classified depending on which component of Democracy it is challenging.

In this step, decision makers, democracy specialists or governments should make a clear assessment either through public opinion, questionnaires or previous market research to define the challenges to address.

The flow chart depicted in Figure 6 aims to help draw a well-established path in order to try to mitigate those threats, depending on which foundational element it may be endangering.
As previously explored, the foundational elements shared by every Democracy are (Robert Longley, 2021):
- Popular sovereignty
- An Electoral System
- Public Participation
- Separation of Powers/Multi-party system
- Human Rights
- A Rule of Law

For the construction of the artifact, these are the elements that will be considered, as they represent the very root of a Democratic system.

To develop a complete description of the threats it is necessary to consider the type of challenge it represents and the sources and inherent consequences resulting from it. This way, it is easier to segment all possible challenges and point a path for a fitting solution. Table 2 illustrates this idea, being divided by the diverse spheres the threats fit in – Social, Economic, Geographic, Politic, Technologic, Individual – followed by the challenge itself, one possible cause and consequence(s).

<table>
<thead>
<tr>
<th>Type of Challenge</th>
<th>Challenge</th>
<th>Cause</th>
<th>Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td>Voter apathy</td>
<td>Disengagement</td>
<td>Weaken representation</td>
</tr>
<tr>
<td>Economic</td>
<td>Inequalities</td>
<td>Lack of resources</td>
<td>Political inequality</td>
</tr>
<tr>
<td>Geographic</td>
<td>Climate crisis</td>
<td>Threat to humanity</td>
<td>Poverty, food scarcity, inequalities, etc.</td>
</tr>
<tr>
<td>Politic</td>
<td>Corruption</td>
<td>Greed</td>
<td>Deteriorates public trust in government</td>
</tr>
<tr>
<td>Technologic</td>
<td>Online disinformation</td>
<td>Social media</td>
<td>Citizens make wrong decision/Are manipulated</td>
</tr>
<tr>
<td>Individual</td>
<td>Preserve Privacy and data protection</td>
<td>Technology evolution</td>
<td>Cyberattacks, surveillance, biometric data more accessible</td>
</tr>
</tbody>
</table>

Table 2: Guide table for Challenges’ description by type (Froomkin & Shapiro, 2021)
When it comes to evaluating the challenges, it must measure the relevance of the rights being violated and the odds of it happening. This way, it is possible to establish a priority order between verifiable threats and apply XAI Technologies accordingly. For this assessment to be as accurate as possible, it is necessary for the right specialists to be involved at this stage.

Table 3 illustrates a qualitative risk analysis using a probability/impact ranking matrix, constructed specifically for the application in this Master thesis, which is divided as followed:

- (L): Low
- (M): Moderate
- (S): Severe
- (C): Critical

<table>
<thead>
<tr>
<th>Probability/Impact</th>
<th>Meaningless (Within range of a normally functioning consolidated democracy)</th>
<th>Not significant (Moderate violations atypical of a consolidated democracy, but that do not yet threaten breakdown)</th>
<th>Significant (Violations that signal significant erosion of democracy quality and warn of high potential for breakdown in future)</th>
<th>Critical (Critical violations that seriously threaten near-term survival)</th>
<th>Extreme (Violations severe enough to make system non-democratic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rare</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Improbable</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Severe</td>
<td>Severe</td>
</tr>
<tr>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Severe</td>
<td>Critical</td>
<td>Critical</td>
</tr>
<tr>
<td>Probable</td>
<td>Moderate</td>
<td>Severe</td>
<td>Critical</td>
<td>Critical</td>
<td>Critical</td>
</tr>
<tr>
<td>Very Probable</td>
<td>Moderate</td>
<td>Severe</td>
<td>Critical</td>
<td>Critical</td>
<td>Critical</td>
</tr>
</tbody>
</table>

Table 3: Probability/Impact matrix for risk analysis (Protect Democracy, 2023)

Once this analysis is made, it is expected that the application of these first steps of the framework results in an artifact such as is illustrated in Table 4. It is expected for this to be the starting point for the definition of viable ways to mitigate identified threats and which are the most suitable XAI Techniques to apply for each case.
### Table 4: Example of threat evaluation and risk assessment

<table>
<thead>
<tr>
<th>Foundational Element</th>
<th>Type</th>
<th>Name</th>
<th>Cause</th>
<th>Consequence</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Participation</td>
<td>Technologic</td>
<td>Online disinformation</td>
<td>Social media</td>
<td>Citizens make wrong decision/Are manipulated</td>
<td>S</td>
</tr>
</tbody>
</table>

#### 5.2.2 Identify ways to mitigate challenge

In this phase, in which viable measures are identified, as well as how and where they should be applied to mitigate the subsequent threat previously identified. The following Figure 7 shows the flow chart steps associated:

![Flow chart for threat management and measure definition](image)

The application of measures can be divided into 2 spheres:

- **Internal Factors**: The measure may be defined if the threat can be identified as an intrinsic/internal factor in a Democratic System.

- **External Factors**: The measure can also be defined according to a certain external occurrence that may affect a Democratic system.

The definition of a measure is a crucial step as it leads to the ultimate goal, to mitigate possible threats to Democracy through the application of XAI Techniques. We may observe:

- Preventive measures: e.g. Analyze crime data and predict the likelihood of future events
- Reactive measures: e.g. analyze job postings and resumes, to ensure that hiring processes are fair and unbiased.

The definition of measures must always consider its complexity, urgency and probability of failure since it can be mitigating one or more threats at a time and may benefit from the application of different XAI Technologies. As an example, XAI techniques may be used to reduce discrimination and improve fairness in predictive policing models, which are used to identify areas and individuals at substantial risk of
committing crimes. At the same time, it allows law enforcement agencies to allocate resources more effectively.

Depending on the risk assessment and probability of events, it is important to highlight that it may be necessary to add multiple measures and technologies. Low risk events may require simpler measures and technologies, while critical events will demand a more complex measure or set of measures and technologies in order to maximize the chances of effective mitigation of the threat.

The last phase of the threats management and solution definition is a new risk assessment after the implementation of that measure. This allows us to measure the real difference between the previous state and risk and how the application of a measure or measures helped minimize it, or not. And as depicted in the flow chart in Figure 7, as long as the risk remains above “Low”, it should be defined and implemented complementary measures.

To illustrate this phase of the framework, an artifact similar to Table 5 should be accomplished.

<table>
<thead>
<tr>
<th>Threats</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>Technologic</td>
<td>Online Disinformation</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Table 5: Example of measure Definition** |

5.2.3 Identify Requirements

After the definition of a viable way or ways to mitigate the threat, it is necessary to identify requirements and conditions for the selection of the best suitable XAI Technique for each case.

*Figure 8: Flow Chart for requirements identification*

Some relevant requirements and conditions should be analyzed and identified to be able to describe and establish the defined measure. The requirements are based on characteristics about the context, the type of data, model properties, explainability needs and compatibility. All these are fundamental requirements to be recognized before selecting a fitting XAI technology.

- **Data requirements**: Is the data you will work on structured or unstructured, numerical or categorical, high-dimensional or low-dimensional, noisy or clean? This is key to then identify the
different algorithms and techniques for the XAI tool. Another important Data requirement would be Data accessibility and permission, according to GDPR.

- **Model requirements:** Is the model a black-box or white-box model, a simple or complex model, a supervised or unsupervised model, a classification or regression model? Depending on it, the selection of approaches for the XAI tool can differ.

- **Explainability Requirements:** It is important for the XAI tool to understand the explainability needs and type of audience that will perceive it. It is relevant to understand if the measure will reach the general public or not, for example, since the explainability needs may be different from applying a measure and technology inside an organization or institution.

- **Compatibility requirements:** To determine whether it is feasible, in practice, to apply an explanation method to that specific scenario, covering aspects previously identified as requirements, such as the type of measure, the scope of the explanation needed, the model and the training data. Also, compatibility requirements cover the scope of data privacy and data rights. It is crucial to analyze the feasibility of getting access to the required data, or possible terms and also social and IT acceptance.

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>Numerical</td>
</tr>
<tr>
<td>Unstructured</td>
<td>Categorical</td>
</tr>
</tbody>
</table>

*Table 6: Example for data requirements*

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Model</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black box</td>
<td>Complex</td>
<td>Harder to explain with intrinsic or model-specific methods</td>
</tr>
<tr>
<td>White box</td>
<td>Simple</td>
<td>Extrinsic or model-agnostic methods</td>
</tr>
</tbody>
</table>

*Table 7: Example for model requirements*
### 5.2.4 Selection of XAI Technology

After defining the measure, identifying the requirements and defining which conditions and restrictions exist inside the specific context, it is finally possible to define the technology that will materialize those requirements. In this stage, technology specialists may assess the best type of technology/technologies to be implemented for the scenario in study.

<table>
<thead>
<tr>
<th>Threat Type</th>
<th>Name</th>
<th>Risk</th>
<th>Sphere</th>
<th>Measure</th>
<th>Description</th>
<th>XAI Technology</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technologic</td>
<td>Online Disinformation</td>
<td>S</td>
<td>Internal</td>
<td>Preventive</td>
<td>Analyze social media data and patterns to predict the likelihood of future events.</td>
<td></td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reactive</td>
<td>Identify and flag false/misleading information online.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 8: Example of XAI Technology selection*

After the last phase is developed and implemented, it is still not the end of the work. The developed solution must be continuously revisited and evaluated.

It is necessary to evaluate the applied XAI model using metrics such as accuracy, precision, transparency, and consistency to ensure that it is providing accurate and reliable explanations. This evaluation may require extra times, for specialists to make trade-offs between explainability and accuracy. The team must continuously monitor and update the XAI model as needed in order to maintain the desired accuracy, transparency, and fairness.

It is also vital to test the XAI model for the existence of bias to ensure that they are fair and non-discriminatory. Here lies the primary goal of the designed framework, to work on clear and non-biased data handled by specialists who are free of private or ulterior motivations.

Additional recommendations are described in table x as an addition to the example given throughout the explanation of the framework steps:

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Measure Description</th>
<th>XAI Technology</th>
<th>Expected Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explainability in voting systems</td>
<td>Explain the outputs of ML models to detect election fraud and make decision-making process more transparent.</td>
<td>LIME and SHAP</td>
<td>Increase transparency and accountability in voting systems and election processes</td>
</tr>
<tr>
<td>Fairness in predictive policing and enhancing public safety</td>
<td>Reduce discrimination and improve fairness in predictive policing models</td>
<td>Counterfactual Explanations and Neural networks</td>
<td>Identify and correct bias in predictive policing models AND identify areas/individuals at substantial risk of committing crimes</td>
</tr>
</tbody>
</table>
### Table 9: Additional recommendations with challenges, possible measure and technology

<table>
<thead>
<tr>
<th>Access to voting and ensure fair representation</th>
<th>Improve security and accessibility of voting processes and analyze voting patterns.</th>
<th>K-Means Clustering or Hierarchical Clustering</th>
<th>Ensure district boundaries are drawn in a fair and representative manner.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corruption and abuse of power</td>
<td>Identify unusual patterns of behavior in financial transactions, which can help detect potential instances of bribery/embezzlement</td>
<td>Isolation Forest or One-Class Support Vector Machines (SVM)</td>
<td>Increase transparency and accountability in governments and reduce opportunities for corruption.</td>
</tr>
<tr>
<td>Citizen Participation</td>
<td>Identify public opinion on specific issues by analyzing social media and other online sources.</td>
<td>Sentiment Analysis</td>
<td>Facilitate citizen participation in government decision-making processes.</td>
</tr>
</tbody>
</table>

#### 5.3 Evaluation

The evaluation of the proposed framework for the application of XAI Techniques in the Democratic environment to help achieve measures for the mitigation of Democracy challenges were made through interviews to specialists of different scopes in order to obtain a richer and wider feedback and insights.

The construction of the framework results from the assumptions gathered through the comprehensive literature review previously made. Accordingly, the conducted interviews intended to measure the utility and validity of the framework as well as to understand possible improvements.

To support the interviews, a PowerPoint presentation was developed to display the main goal, relevance and explanation of the proposed framework, which can be seen in Appendix A. After the presentation of the framework to the specialists, three questions were asked:

1) Do you consider the proposed framework useful? Why?
2) Do you agree with the proposed framework? If not, please explain.
3) Do you have any recommendation or suggestions for further improvements to the proposed framework?

These questions enabled the assessment of the framework utility inside the democratic scope with the application of XAI technologies, evaluate the structure and design of the artifact, gather some suggestions for improvement or better clarification to include in the present research work and some limitations to consider as well. The analysis and answers made by the four specialists can be found in Appendix B.
5.4 Discussion of Results

After the evaluation made to the proposed framework by the different specialists, a reflection about the feedback gained was made.

When answering the first question, regarding the utility of the framework, all inquiries expressed how useful they found the artifact to be, mostly for decision makers. Specialist 1 found that the framework would work as a guide or a tool to better assess decisions in the scope of Democracy or even for a more standardized process. This last part is in line with the opinion of Specialist 3 who visions the XAI tool being used in processes or procedures where AI is already being applied, or close to being in the future, or where the process won’t be a black box.

Concerning the validity of the framework, there was a general approval of the artifact presented through the conducted interviews (Appendix A). The inquiries agree that the flow and design of the proposed framework makes sense and touches all relevant matters. A highlighted point was the necessity of it being applied by a multidisciplinary team for it to really work and be well and accurately assessed through all the steps.

Some recommendations and suggestions for the framework’s improvement were also obtained:

- Specialist 1 suggested the adjustment of some terminologies since it was currently referenced solution and the specialist considered it should be adapted since what was presented were ways to mitigate problems and not solutions or policies. Another recommendation was to clearly explain throughout the framework, that there should exist a constant involvement of the different type of specialists in all steps.

- Specialist 2 similarly with Specialist 1 recommended the inclusion of different type of stakeholders in the process, namely people with experience in human rights and democratic rights. So that an accurate version and assessment of reality is guaranteed. This specialist also suggested that when selecting the suitable XAI technology, more than assessing if we can have access to the data, to also guarantee responsible innovation.

- Specialist 3 found that a possible recommendation, or work for the future, would be to test the framework in very specific areas where AI is already implemented and keep on testing the components of XAI tool in an incremental way instead of a revolutionary way.

- Specialist 4 emphasized the necessity for a clear mention of GDPR as a challenge for the application of the proposed framework. The specialist found this to be a possible barrier for the outcome desired and recommended adding extra mentions or workarounds to this topic throughout the thesis.

After analysis of the feedback received, some of the recommendations and suggestions were considered and applied. In point 4.4, in the challenges part, the GDPR was already mentioned but was then done in more depth to stress the relevance it holds. Then, point 5.2 was also incremented with a clear statement about the multidisciplinary team that must exist in order to apply the artifact. Many other mentions about this team were also added throughout the proposal.
In the 2\textsuperscript{nd} step of the proposed framework the title was changed from ‘possible solutions’ to ‘Identify ways to mitigate challenges’ and all references to ‘solution’ were changed to ‘measure’ as a way to assure correct terminology for what is being proposed.

In the 3\textsuperscript{rd} step of the proposed framework, in data requirement’s section, it was included once again the necessity of assessment of GDPR as a requirement.
6 Conclusion

The present chapter includes an outlined review of the work developed throughout the Thesis, some conclusions were drawn and finalized with work that can be applied in the future.

In this Master Thesis it was possible to study more in depth the Explainable Artificial Intelligence scope and its techniques as well as the Democratic context, construct an artifact to define challenges and the many types of measures that may be applied in order to mitigate those challenges. To do so, it was crucial to understand XAI concepts, the approaches and algorithms but also the practical application of this type of technologies in many social areas but mainly in the Democratic context. Therefore, more than studying Democracy and its history, it was found crucial to study challenges and problems in Democracy but also challenges and opportunities of using AI in Democracy.

The extensive literature review and Environment study, allowed for the identification of a set of assumptions that led to the constructions of the framework. This framework has four main steps with the purpose of being applied in the service of Democracy through the application of the different possible XAI techniques that may, ultimately, help mitigate existing challenges and risks for the Democratic system.

The proposed artifact intends to display and include all necessary steps to reach the selection of the best suitable XAI technology or technologies. Thus, it is mandatory to follow the proposed workflow, this means, make a comprehensive assessment of Democracy challenges, identify then ways to mitigate them by defining measures that would accomplish that and outline all necessary requirements. To provide a clear understanding of what is being proposed, flow charts and examples were given for every step of the artifact.

An evaluation of the proposed framework was then made through the conduction of interviews with specialists from different areas but related with the topics at study. The interviews were important for the measure of the framework’s utility and validity as well as to obtain valuable feedback and suggestions for improvements which led to some rectifications in the project but also provided inputs for limitations and future work.

6.1 Limitations

- A limitation is the fact that the proposed framework was not applied in a specific case study which could have improved its accuracy or validity. Even though the artifact was developed through an extensive literature review and evaluated by experienced professionals, its practical application would possibly lead to the identification of critical points and other improvements.

- As mentioned previously, GDPR is both a challenge for the application of AI in Democracy and a limitation. In the step of identifying requirements, it will always be necessary to assess the accessibility of the data required. There may be some barriers due to data privacy rules and laws, and some data might even be impossible to obtain.
The proposed framework is conditioned by the gathering of a multidisciplinary team in order for its application to be possible. Yet, it is expected that this team is unbiased and will work on the framework goal without ulterior motives and taking advantage of the multiplicity of ideas and knowledge. All these are assumptions that may or may not happen in practice due to human volatility. Hence, it represents a limitation for the work developed as these premises are crucial for reaching the expected results from the framework.

6.2 Future Work

Some considerations for future work are as followed:

- With the goal of continuously improving the present framework, its application in multiple different study cases would be fundamental. This way, as mentioned in the Limitations chapter, critical points could be identified.

- As a suggestion for a workaround regarding GDPR and the limitations it might raise for some types of sensitive data, is masking or encrypting the data or the part that should not be accessed by everyone. This would allow for the team to still work on the data without violating any GDPR rights and then provide unmasking/decryption permission only for the entity or individual(s) that may access that information.
References


Aminah Aliu. (2021, January 9). A Is for AI, B Is for Bias. Medium. https://medium.com/swlh/a-is-for-ai-b-is-for-bias-8fabc0f4a93b


Inter-Parliamentary Union. (2022, September 14). How to mitigate these five threats to democracy. Inter-Parliamentary Union. https://www.ipu.org/news/news-in-brief/2022-09/how-mitigate-these-five-threats-democracy


Ken Bryden. (2022, November 28). The racist Microsoft bot that lasted less than a day on Twitter by Ken Bryden . Crast.Net. https://crast.net/214929/the-racist-microsoft-bot-that-lasted-less-than-a-day-on-twitter/


Appendix

A - PowerPoint presentation of the framework for the application of XAI Techniques in the service of democracy

Document presented during the conduction of interviews for the evaluation and validation of the proposed framework in the scope of the present Master thesis.
Build a comprehensive framework for the application of Explainable Artificial Intelligence techniques in the service of Democracy

- The framework aims to provide a path that will hopefully help mitigate current and possible problems in the Democratic System through the application of XAI Technologies.

Study relevance

- Lies on the many purposes of adding an explainability trait to ML algorithms, such as: improving system understanding, offer better behavior predictability and, naturally, increase trust in the system. Moreover, it can help build some bridges between AI and other social sectors such as health, justice and Public Sector, for example.

Framework

- The Framework has 4 main steps:

  1. Democracy challenges
  2. Ways to mitigate challenge
  3. Identify Requirements
  4. Suitable XAI techniques

- The application of the framework should be a joined effort between ONG’s, Governments or decision-makers and Technology specialists. It requires multidisciplinary teams throughout the process.

Next slides will have flow charts and examples of expected outputs in each step.
1. Democracy challenges Assessment

(Step 2)

Example of expected output:

<table>
<thead>
<tr>
<th>Foundational Element</th>
<th>Type</th>
<th>Name</th>
<th>Cause</th>
<th>Consequence</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Participation</td>
<td>Technological</td>
<td>Online disinformation</td>
<td>Social media</td>
<td>Citizens make wrong decision/Are manipulated</td>
<td></td>
</tr>
</tbody>
</table>

Types: Social/Economic, Geographic/Political, Technologic, Individual

Obtained through a Probability/Impact matrix for risk analysis

2. Identify ways to mitigate challenge

(Step 3)

Example of expected output:

<table>
<thead>
<tr>
<th>Threats</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Name</td>
</tr>
<tr>
<td>Technologic</td>
<td>Online Disinformation</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Internal Factors: Solution may be defined if the threat can be identified as an intrinsic/internal factor in a Democratic System.

External Factors: The solution can also be defined according to a certain external occurrence that may affect a Democratic system. (e.g., Wars, climate catastrophes, etc.)
3. Identify Requirements

(Step 4) Definition of a measure → Identify Requirements → Select XAI Technology

Requirements:

• Data requirements: Is the data you will work on structured or unstructured, numerical or categorical, high-dimensional or low-dimensional, noisy or clean?

Example:

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>Linear or tree-based models</td>
</tr>
<tr>
<td>Unstructured</td>
<td>Deep learning or natural language processing</td>
</tr>
</tbody>
</table>

• Model requirements: Is the model a black-box or white-box model, a simple or complex model, a supervised or unsupervised model, a classification or regression model?

Example:

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Model</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black box</td>
<td>Simple</td>
<td>Explain or explain in more intuitive or model-specific methods</td>
</tr>
<tr>
<td>White box</td>
<td>Simple</td>
<td>Explain or explain in more intuitive or model-specific methods</td>
</tr>
</tbody>
</table>

• Explainability requirements: It is important for the XAI tool to understand the explainability needs and type of audience that will perceive it.

• Compatibility requirements: To determine whether it is feasible, in practice, to apply an explanation method to that specific scenario.

4. Selection of XAI Technology

Example of expected output:

<table>
<thead>
<tr>
<th>Threat</th>
<th>Measure</th>
<th>XAI Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technologic</td>
<td>Preventive</td>
<td>Analyze social media data and patterns to predict the likelihood of future events.</td>
</tr>
<tr>
<td></td>
<td>Reactive</td>
<td>Identify and flag false/misleading information online.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Threat</th>
<th>Measure Description</th>
<th>XAI Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Explainable online - Predictive models to detect electronic fraud and lower economic crime.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Counterfactual - Explainable online - Predictive models.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explainable online - Predictive models.</td>
<td></td>
</tr>
</tbody>
</table>

Additional Recommendations:

<table>
<thead>
<tr>
<th>Threat</th>
<th>Measure Description</th>
<th>XAI Technology</th>
<th>Expected Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Explainable online - Predictive models to detect electronic fraud and lower economic crime.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Counterfactual - Explainable online - Predictive models.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explainable online - Predictive models.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reliable and accurate in predicting pricing and identifying meaningful information of cotton commodities.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explainable online - Predictive models.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explainable online - Predictive models.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reliable and accurate in predicting pricing and identifying meaningful information of cotton commodities.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explainable online - Predictive models.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reliable and accurate in predicting pricing and identifying meaningful information of cotton commodities.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Diagram of NOVA IMS Information Management Framework]
Interview Questions

1) Do you consider the proposed framework useful? Why?
2) Do you agree with the proposed framework? If not, please explain.
3) Do you have any recommendations or suggestions for further improvements of the proposed framework?

Thank you for your time and expertise!
B - Interviews

Transcription of the interviews conducted in order to evaluate and validate the proposed framework.

B.1 - Specialist 1

Name: Mijail Naranjo

Biography: PhD in Geoinformatics and Assistant Professor at NOVA IMS with wide experience in Information Management.

1) Do you consider the proposed framework useful? Why?

Yes, I would say it’s very useful. Decision makers can have a guide, a tool to better assess decisions in this scope or a more standardized process. This way, the team can decide based on the same image, otherwise everyone can have their own understanding of the problem.

2) Do you agree with the proposed framework? If not, please explain.

I agree, but the term used ‘solution’ must be adjusted.

3) Do you have any recommendation or suggestions for further improvements to the proposed framework?

A bit linked with question 2), adjust some terminologies and another recommendation, since the framework is made to two different types of audience, it’s that it is important for both to work together. If we have only technical people working in the framework it will not be possible to correctly assess what is biased, what is not, what are the implications and so on. So, clearly define in the framework where should be placed the involvement of the type of users and how they can communicate with each other.

GDPR for data privacy as limitation

B.2 - Specialist 2

Name: Ana Albergaria

Biography: PhD in Artificial Intelligence and Human Rights with wide specialization in Sociology and Research.

1) Do you consider the proposed framework useful? Why?

Yes, I find it very useful, mainly because it reaches an almost micro deconstruction of the challenges. They are identified, are seen from different point of views, approaches and scopes and it also offers a preventive or reactive response to it. And it is certainly very important for decision makers to decide alongside technology specialists which technology to use without threatening the foundational elements of Democracy. Therefore, I found it a very interesting exercise, it is a real working tool for decision makers to be able to know which XAI technology may apply in the service of democracy.
2) **Do you agree with the proposed framework? If not, please explain.**

Yes, I agree with the proposal after all clarifications and definition that a multidisciplinary team must exist.

3) **Do you have any recommendation or suggestions for further improvements to the proposed framework?**

A suggestion I would give, is the one we talked about throughout the presentation that is including different type of stakeholders in the process, namely people with experience with human rights and democratic rights who will guarantee an accurate version and assessment of the reality. And also, when selecting the suitable technology, more than assessing if we can have access to the data, having many different perspectives to guarantee responsible innovation.

**B.3 - Specialist 3**

**Name:** Filipe Montargil

**Bibliography:** PhD in Sociology, Professor and Researcher on society and technology

1) **Do you consider the proposed framework useful? Why?**

I think the framework might be useful for procedural matters regarding our system’s functioning. I see the XAI tool being used in processes or procedures where we are already applying or close to applying AI and where the process will not be a black box.

2) **Do you agree with the proposed framework? If not, please explain.**

Yes, globally the framework makes sense.

3) **Do you have any recommendation or suggestions for further improvements to the proposed framework?**

For me it is better to apply in practice the solution, fast and with minimum viable product, something lighter and start learning from mistakes and keep improving. So, one recommendation would be to test the framework in very specific areas where AI is already implemented and keep on testing the components of XAI tool in an incremental way instead of a revolutionary way.

**B.4 – Specialist 4**

**Name:** Catarina Ferraz

**Bibliography:** Master’s degree in information management with specialization in Marketing Intelligence, Anthropologist and Researcher

1) **Do you consider the proposed framework useful? Why?**
Yes, the framework itself is very useful almost ahead of its time and highlights relevant challenges in today’s society that the governments, unfortunately, won’t address.

2) Do you agree with the proposed framework? If not, please explain.

Yes, I agree with the framework.

3) Do you have any recommendation or suggestions for further improvements to the proposed framework?

A recommendation would be to mention GDPR as a challenge for the application of the framework.