ARTICLE TYPE

AI Solutions for Transparent, Explainable and Regulatory Compliant Public Policy Development

Article title

Author 1: Mr. Thanasis Papadakis [1]

Author 2: Dr. Ioannis T. Christou, [1,2]

Author 3: Mr. Charalampos Ipektsidis [1]*

Author 4: Dr. John Soldatos [1]

Author 5: Mr. Alessandro Amicone [3]

[1] Netcompany-Intrasoft, Research and Innovation Development (RID) Department, Luxembourg, Luxembourg

[2]The American College of Greece, Athens, Greece

[3]GFT Italia Srl, CU Innovation, Genova, Italy

*Corresponding author: John.Soldatos@Netcompany-Intrasoft.com

Keywords: Artificial Intelligence; Public Policy; Machine Learning; Explainable AI; AI Act

Abbreviations (if needed): AI: Artificial Intelligence; XAI: Explainable Artificial Intelligence; ML: Machine Learning; DL: Deep Learning; MLP: Multilayer Perceptron Network, ANN: Artificial Neural Network; QARMA: Quantitative Association Rules Mining Algorithm

Abstract

In recent years, public policy makers leverage large amounts of policy-related digital data that are generated through different channels (e.g., e-services, social media) to realize a shift towards data-driven evidence-based policy development. The advent of Machine Learning (ML) and Artificial Intelligence (AI) holds the promise to facilitate and accelerate this shift, through easing and automating the processing of large datasets, while helping policy makers to identify unique, yet previously hidden, policy development insights. Nevertheless, the use of AI for public policy development is also associated with significant technical, political and regulatory challenges. This paper discusses these challenges and suggests a range of technical and technological solutions for overcoming them. The latter solutions include a reference architecture and a blueprint data mining process for AI-based policy making, along with AI algorithms that alleviate AI bias and boost transparency and explainability. Moreover, the paper presents the practical validation and use of these technological building blocks in real-life public policy making cases.

Policy Significance Statement

This paper illustrates a collection of AI solutions that can empower data scientists and policy makers to use AI/ML for the development of explainable and transparent policies in-line with the emerging European regulation for AI. It discusses the challenges of AI based policy making, along with potential solutions at the technical and technological level.

1. Introduction

1.1. AI for Public Policy Making: The Rationale

In an era of rapid digital transformation and technology acceleration, Artificial Intelligence (AI) is one of the most prominent and disruptive enablers of novel digital applications in a variety of sectors such as energy, healthcare, smart cities, public administration, and industry. From a technical and technological viewpoint, AI's potential to drastically improve the efficiency of business processes has been greatly propelled by advances in parallel hardware (e.g., Jouppi (2017)) and scalable software systems e.g., (Gonzalez et. al. 2012). The latter have enabled the development of advanced machine learning frameworks (e.g., (Dean et. al. 2012)) and novel optimization algorithms (e.g., (Kingma and Ba 2014)) that are suitable for large scale problems in realistic settings. From a business perspective, AI is enabling organizations to process vast amounts of information in timely, automated, and cost-effective ways (Leyer 2021). It also facilitates organizations to use the outcomes of this processing in order to optimize and accelerate their decision-making.

One of the most widely used subfields of AI is Machine Learning (ML), which enables computing systems to learn without human instruction, but rather based on historical data and statistical knowledge. Machine learning is very popular in sectors where vast amounts of information are available. Hence, it could become a powerful tool for policy makers. As the volume of big data increases at an exponential rate (Chauhan and Sood. (2021)), policy makers and public administration workers are challenged to collect, read, study, analyze, process and experiment with this information. ML systems could boost automation in the processing of large volumes of data towards providing relevant insights to policy makers. In this way, they can help policy makers to take advantage of growing data volumes in scalable and cost-effective ways. Moreover, ML ensures that policy makers consider all relevant information via correlation and cross-analysis of multiple datasets. In many cases, ML models can also be used to unveil potentially hidden patterns and correlations of policy-related datasets. The latter can drive policy optimizations that are hardly possible based on the legacy methods used to process information.

Overall, ML has the potential to improve the way public policies are developed. Policy makers are interested in new methods of understanding and analyzing data that will help them make better decisions (Edwards and Veale 2018). There is a layer of human ingenuity missing from existing policy making models (Deng et. al. 2020). True machine learning is one way to close this gap. ML technology can provide decision makers with high-level analysis, helping them to connect the dots and arrive at more effective policy making strategies in areas such as crime flighting, ensuring health, and protecting the environment. In recent years, the use of ML as a public policy making tool is increasingly placed in the strategic agenda of public policy organizations such as central, regional, and local governments (e.g., (Lindgren et. al. 2019), (Rosemann, Becker and Chasin 2020)).

1.2. Bias, Explainability and Regulatory Challenges

While ML is clearly going to be useful, there is still a long way to go before it becomes an everyday tool in policy-making scenarios. Policymakers are just beginning to learn how ML can assist them to find answers to some of the world's most pressing questions. Moreover, several research and innovation projects are currently being developed for ML-based policy making systems. However, to develop ML that works for policy making in practice, there is a need to address the following challenges:

- **Biased Systems**: Algorithmic bias is one of the most prominent sources of harm of the ML applications lifecycle (Hao 2019), (Harini and Guttag 2021). Policy makers must ensure that the algorithms used for policy making are unbiased, representative and leave no citizen behind. This can be extremely challenging given the proclaimed lack of representative datasets about citizens and policy making processes. To make things worse, algorithmic bias is often introduced in unintended ways i.e., developers and data scientists working on public policy making systems may unintentionally build biased systems (Harini and Guttag 2019). For instance, using collections of data from electronic channels is likely to ignore the needs of elderly or low-income citizens that don't use such channels. Likewise, the use of historical datasets that correspond to biased processes can transfer human biases to the results of AI algorithms for public policy making.
- **Explainability and Transparency**: Public policy makers must be able to explain their data-driven decisions to citizens. At the same time, the policy making process needs to be transparent and trusted, otherwise it cannot be accepted by citizens and the society at large. Unfortunately, the most efficient ML models and algorithms (e.g., deep neural networks) tend to operate as black boxes with rather limited transparency and explainability (Soldatos and Kyriazis 2021). This makes their use in pragmatic policy making settings very difficult. To remedy this situation, the research community is working on explainable AI (XAI) techniques, which aim to either build AI systems that are themselves transparent, or otherwise be able to explain how black-box algorithms work and to render them more acceptable to citizens. Nevertheless, the use of XAI in

public policy making settings is still in its infancy: even the very concept of what constitutes a good explanation is still under debate currently. Moreover, there are no agreed and proven ways for selecting models that balance performance and explainability in-line with the requirements of policy makers.

- **Regulatory Compliance**: ML-based systems for public policy making must comply with emerging regulations in AI, such as the AI Act of the European Parliament and Council of Europe. The AI Act is globally the first systematic effort to regulate AI systems. It defines stringent requirements for high-risk AI systems that will be used to drive crucial decisions, like most public policy making decisions. These requirements include guarantees for transparency, explainability, data quality, and human oversight. As already outlined, emerging AI technologies like XAI can help meet these requirements. Nevertheless, there is no systematic approach for mapping technology tools into concrete requirements for the AI Act.

Overall, prior to deploying machine learning for public policy making, policymakers must ensure that the models used are unbiased, transparent, and explainable, otherwise governments and organisations could face serious risks when automating or outsourcing key decisions.

1.3. Related Work

Despite increased interest in the use of AI for data-driven, evidence-based policy making (Gesk and Leyer 2022), the development of practical systems that can be operationalized is in early stages. Some research works can be found in the broader context of data-driven policy making that leverage Big Data (Hochtl, Parycek, and Schollhammer 2016). Some systems for data-driven policy making take advantage of social media information (Bertot, Jaeger and Hansen 2011) based on different techniques, including data mining and machine learning algorithms (Charalabidis, Maragoudakis and Loukis 2016). There are also works on political analysis using statistical techniques, which is a foundation for machine learning (Monogan 2015). More recently, the use of machine learning has been proposed and explored for the analysis and mining of public policy related information, as part of evidence-based policy approaches (Androutsopoulou and Charalabidis 2018). Specifically, ML techniques for public policy related applications have been explored in several areas including taxation (López et. al. 2019), public security and counterterrorism (Lee, Mantari and Roman-Gonzalez 2020), public work design (Eggers, Schatsky and Viechnicki 2017) and healthcare (Qian and Medaglia 2019).

These systems have provided insights on the benefits and the challenges of ML-based policy making. Addressing bias and explainability challenges are acknowledged as being critical requirements for the practical deployment of AI systems by public policy development organizations. The development and use of bias detection toolkits (Bellamy et. al. 2018) and XAI techniques (e.g., (Ribeiro, Singh, and Guestrin 2016), (Fryer, Strümke and Nguyen 2021), (Shrikumar, Greenside, and Kundaje 2017)) are considered as a remedy to bias and explainability issues respectively. However, these toolkits have not been extensively applied, used, and evaluated in public policy making use cases. Likewise, in the light of the emerging AI Act, there is not much research on how explainability techniques could be matched to regulatory requirements in ways that balance performance and explainability. This is particularly important given also concerns about the overall trustworthiness of explanations over back-box models (Rudin 2019), which are usually criticized for their ability to ensure transparent, reliable, and trustworthy AI systems. The present paper unveils the merits of ML and XAI for evidence-based policy development. Moreover, it is one of the first research papers that presents approaches for developing systems that are compliant to the AI Act.

1.4. Article Structure and Contribution

This paper is motivated by existing gaps in the explainability and transparency of public policy use cases, as well as by challenges associated with regulatory compliance. It introduces a data mining process, a reference architecture and ML algorithms for dealing with these issues in the context of policy making. Specifically, we first introduce a reference architecture and a data mining framework that could boost the development of robust and unbiased systems for public policy making. The reference architecture illustrates a set of modules and tools that are destined to support policy makers in adopting, using and fully leveraging AI/ML techniques in their policy making efforts. On top of the reference architecture, a data mining process based on the popular CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is adapted to support public policy making activities. The process makes provisions for Explorative Data Analysis (EDA) to detect and remedy potential biases linked to the used datasets.

From an algorithmic viewpoint the paper presents our work on adapting and applying background algorithms of the authors (Christou 2019, Christou et. al. 2020, Christou et al. 2022) i.e., algorithms produced by the QARMA (Quantitative Associations Rule Mining) ML framework, to public policy making applications for smart cities and communities. The introduced algorithms enable knowledge mining and representation in the form of explainable rules, which boosts the interpretability of the public policy knowledge. This alleviates the limitations of black-box models (e.g., deep neural networks) for policy making without any essential performance penalty; it also allows for a whole new class of explanations for black-box models' decisions.

Another contribution of our paper is the mapping of different algorithmic tools (including QARMA) to different AI-based policy making use cases that feature different risk levels. Specifically, the paper suggests how different algorithms and techniques can boost the regulatory compliance of different classes of AI systems in-line with the risk based classification of the AI Act. This can serve as an early guide for data scientists and other developers of AI/ML systems for public policy making, who wish to develop regulatory compliant systems by design.

The paper is structured as follows:

- Section 2 following this introduction introduces the data mining process and the reference architecture for public policy making. It also illustrates how ML models can drive the evidence-based policy making process.
- Section 3 presents our arsenal of ML algorithms and tools for public policy making use cases. It also discusses how these tools map to the requirements of the different risk levels of the AI Act.
- Section 4 illustrates the use of our data mining process and our explainable algorithms in real-life public policy making use cases. The relative performance of the different algorithms is discussed, along with their suitability for the presented use cases.
- Section 5 is the final and concluding section of the paper.

2. AI Platform for Data Driven Policy Making

To address existing gaps in the development of AI-based use cases for public policy making, we herewith introduce a reference architecture and a data mining framework for the development of robust and unbiased systems for public policy making. The presented reference architecture and the accompanying data mining process are integrated in a single platform. This platform is developed in the AI4PublicPolicy project, which is co-funded by the European Commission in the scope of its H2020 program for research and innovation. In AI4PublicPolicy, policymakers and AI experts are collaborating closely towards unveiling AI's potential for automated, transparent, and citizen-centric development of public policies. The reference architecture of the AI4PublicPolicy platform is specified as a set of software modules and tools that aim at supporting policy makers in using, and fully leveraging AI/ML techniques. At the same time, the data mining process specifies how the various tools of the AI can be used to support the development of end-to-end policy development solutions.

2.1. Platform Reference Architecture and Main Components

The AI4PublicPolicy platform aims to collect and analyze data that are used for automated, transparent and citizen centric development of public policies. The architecture is inspired by the Big Data Value (BDV) Reference Model (Curry et. al. 2021). It is characterized as a *reference architecture* since it is presented at a high-level, abstract, logical form, which provides a blueprint for the implementation of AI-based public policies.

The architecture specifies modules and functionalities for:

- *Data analytics* i.e., the implementation of techniques for understanding and extracting knowledge from data. AI4PublicPolicy specifies and implements AI tools for policy modelling, extraction, simulation, and recommendations. The AI tools include machine learning to extract policy related knowledge from large datasets, including opinion mining and sentiment analysis functionalities.
- *Data protection* i.e., the implementation of technological building blocks for safeguarding sensitive data, such as data anonymization mechanisms.
- *Data processing architectures* i.e., architectural concepts for handling both data-at-rest (e.g., data stored in databases of the policy authorities) and data-in-motion (e.g., data concerning interactions between citizens, the administration, and the e-services of the administration). The architecture enables the handling of streaming data from sentiment analysis and opinion mining technologies that enable the capture of citizens' opinions on social media.
- *Data management* i.e., techniques for dealing with large amounts of data, including management of both structured (e.g., data in tables) and unstructured data (e.g., citizens' opinions in natural language). The employed data management techniques make provisions for handling multilingual data using natural language processing tools and tools for semantic interoperability of policy data sources.
- *Cloud and High-Performance Computing* building blocks that enable the integration of the platform with the portal of the European Open Science Cloud (EOSC) to facilitate access to cloud and High-Performance Computing (HPC) resources. EOSC is a federated system based on a set of existing research infrastructures

that delivers a catalogue of services, software, and data from major research infrastructure providers. The integration of the AI4PublicPolicy platform with the EOSC portal is destined to deliver a complete environment for AI-based policy making. This environment enables the sharing of datasets and models for data driven policies, as part of a cloud-based Virtual Policy Management Environment (VPME).

Moreover, the AI4PublicPolicy platform provides cyber-defense strategies against prominent attacks against AI systems (e.g., poisoning attacks of data) to increase the security of the AI systems in-line with regulatory mandates (e.g., the AI Act) for systems that take critical decisions. Finally, the platform incorporates eXplainable AI (XAI) models that make ML-based policies explainable to humans to enhance transparency and the overall acceptable of policies by citizens and policy makers.

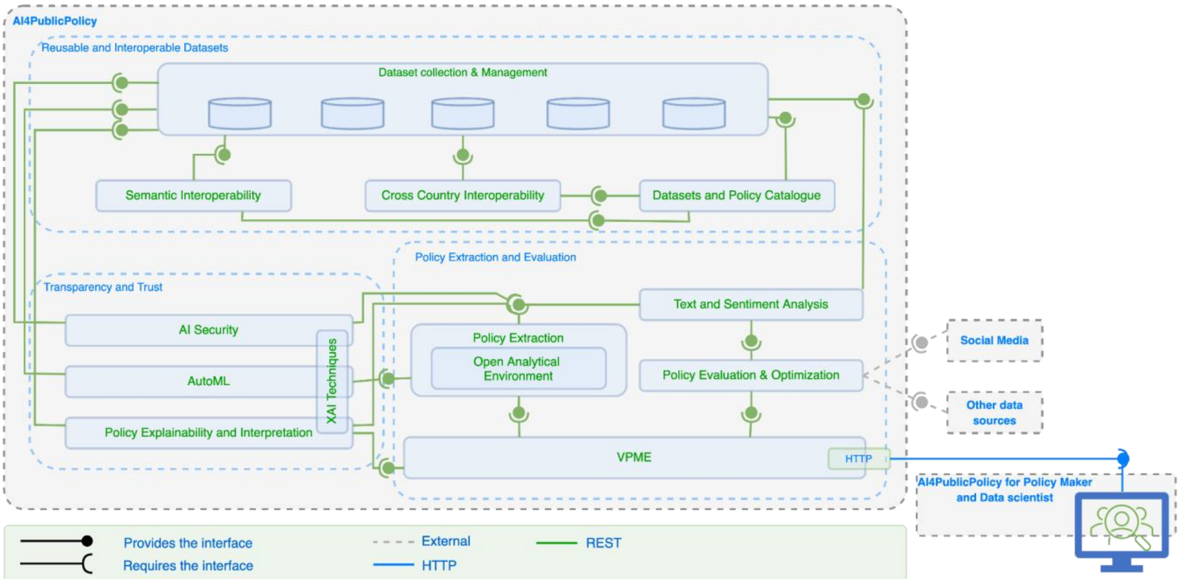

AI4PublicPolicy
Reusable and Interoperable Datasets
Dataset collection & Management
Semantic Interoperability
Cross Country Interoperability
Datasets and Policy Catalogue
Policy Extraction and Evaluation
Transparency and Trust
AI Security
AutoML
Policy Explainability and Interpretation
XAI Techniques
Text and Sentiment Analysis
Policy Extraction
Open Analytical Environment
Policy Evaluation & Optimization
VPME
HTTP
Social Media
Other data sources
AI4PublicPolicy for Policy Maker and Data scientist
Provides the interface
Requires the interface
External
HTTP
REST

Fig. 1. AI4PublicPolicy Architecture Components

Our blueprint architecture for implementing and deploying AI-based policy development is illustrated in Fig. 1. The figure illustrates a logical view of the architecture, including the main components of the cloud-based AI platform and the interactions between them. Specifically, the components of the architecture blueprint include:

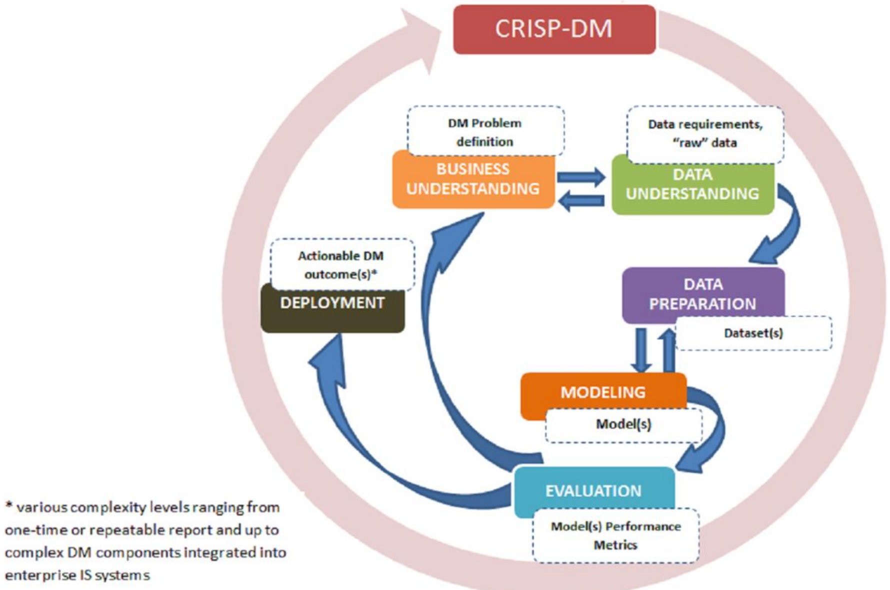
- **Dataset Collection and Management:** Provides the software tools that collect and manage datasets. Datasets are collected through proper Application Programming Interfaces (APIs).
- **Semantic Interoperability:** Provides the interoperability functionalities that enable the mapping between the different data formats available, in-line with an agreed set of formats.
- **Cross Country Interoperability:** Translates data from one language to another target language in order to share data and policies and to boost such sharing.
- **Datasets and Policies Catalogue:** This component is a directory of policies and datasets, which facilitates the dynamic discovery of data and policies towards facilitating reuse.
- **XAI Techniques:** This module is responsible for providing information for analysing and explaining the machine learning models to help humans understand the rationale behind policy decisions.
- **Policy Explainability and Interpretation:** This component is used to build the policy models which will produce the analysis and interpretation of the policy datasets.
- **AI Security:** Incorporates AI-related cyber-defence strategies in order to protect AI systems against attacks (notably data poisoning and evasion attacks).
- **AutoML:** This component facilitates the selection of the optimal algorithms for a specific AI process chain. To this end, it maintains a set of well-established algorithms, which are used to drive the selection of the optimal ones.

- **Text and Sentiment Analysis:** This component provides information about the sentiment of citizens, notably sentiment related to policy decisions. From an implementation perspective it is based on Natural Language Processing (NLP) and text analytics algorithms.
- **Policy Extraction:** Enables the policy maker to choose an AI workflow from a catalogue of machine learning (ML) and deep learning (DL) workflows and apply it.
- **Policy Evaluation and Optimization** allows the simulation and evaluation of policies by making use of the opinions and feedback of local actors to propose new insights and improvements.
- **Virtualized Policy Management Environment (VPME)** is a cloud-based platform that integrates the different components of the platform based on proper APIs.

2.2. Policy Extraction Methodology

On top of the reference architecture that serves as blueprint for integrating and deploying AI-based policies, a data-mining process must be realized to support policy extraction. The methodology leverages data collection and management building blocks of the platform to assemble proper policy making datasets, along with AI/ML techniques for extracting and explaining the policies. In this direction, the Cross Industry Standard Process for Data Mining (CRISP-DM) process (Marban, Mariscal and Segovia (2009)) has been properly adapted. Specifically, the six phases[1] of CRISP-DM are used to support the policy development process as follows (Fig. 2).

1. **Business understanding**: This step is focused on the specification of the policy extraction problem and its framing in the correct policy context.
2. **Data understanding**: This step focuses on exploring the availability of proper datasets for the policy extraction problem at hand, including the availability of data with proper volumes and a representative nature that helps alleviate bias.
3. **Data preparation**: This step is identical to the corresponding dataset of CRISP-DM. It comprises a set of repetitive data preparation tasks to construct the final dataset from the initial raw data.
4. **Modeling**: This step deals with the selection and application of the various modelling techniques followed by calibration of their parameters. In conjunction with the AI4PP architecture, it is facilitated by the catalogue of algorithms and the AutoML components.
5. **Evaluation**: This step evaluates the selected models against their policy development goals.
6. **Deployment**: Once a set of appropriate policies have been extracted, validated, and evaluated, this step deals with the actual deployment of the ML functionalities that will help extract policies and visualize them to policy makers.



* various complexity levels ranging from one-time or repeatable report and up to complex DM components integrated into enterprise IS systems

Fig. 2. CRISP-DM phases and key outputs (Chapman 2020), (Plotnikova, Dumas and Milani (2020))

[1] https://www.datascience-pm.com/crisp-dm-2/

In the context of our reference architecture, the CRISP-DM methodology is used to provide a data-driven, AI-based and evidence-based approach to extracting policies. It also specifies the phases of collaboration between policy makers, data scientists and AI experts. The latter are the stakeholders that participate in the realization of the various phases of the adapted CRISP-DM process.

2.3. ML-Enabled Policy Making Process

Using the AI4PublicPolicy platform and the CRISP-DM process, policy makers (e.g., governmental officials) can benefit from a novel data-driven policy making process, which is illustrated in Fig. 3. The process involves the development of ML models based on data from a variety of sources including citizen feedback. These machine learning models are enhanced with domain specific metadata to enable the production of policy models. The latter reflect real world decisions that can be optimized based on the parameters of the ML model. In this direction, there is a need for collaborative interactions between the policy maker and the data scientist. This collaboration aims at translating the low-level semantics of ML models (i.e., the parameters, hyperparameters and attributes of an ML model) to high-level semantics of policy models (e.g., semantics associated with policy decisions).

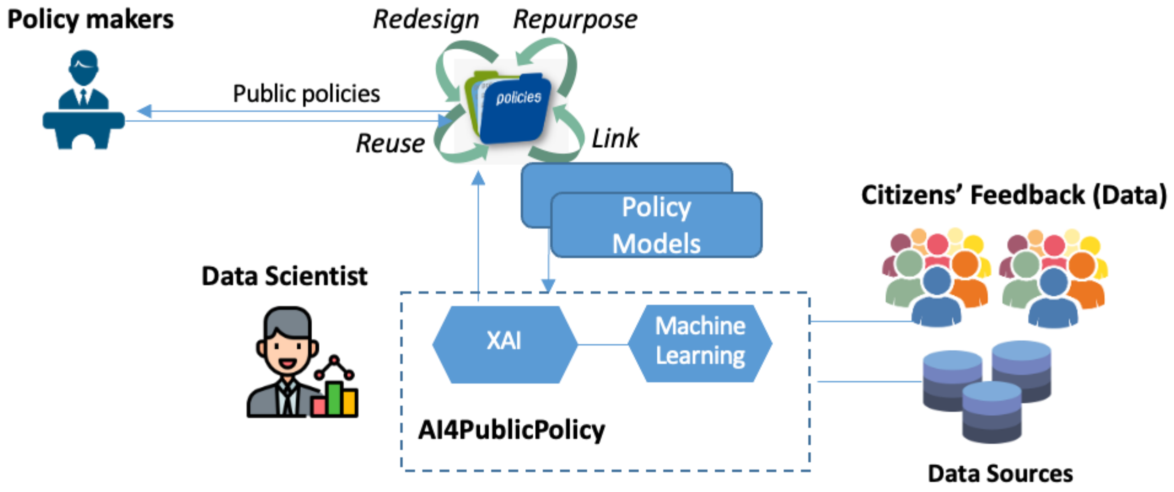


Fig. 3. Policy Making Process using the AI4PublicPolicy Platform

Key to the implementation of the policy making process of Fig. 3 is the XAI modules and capabilities of the platform. These capabilities provide insights on why and how an ML model suggests a particular policy decision. Specifically, they provide insights on what patterns have been learnt by the model, how they interact with each other, as well as what are the main parameters that drive the suggested decisions. Hence, they are a key to explaining how the ML model and its associated policy model operates. In many cases, the XAI modules of the platform provide a human-readable explanation of the model, which facilitates policy makers to understand and to use it. As a prominent example, the QARMA family of explainable machine learning algorithms that are used for the validating use cases in Section 4, expresses policy models in the form of easy-to-understand rules.

3. Machine Learning Methods for Public Policy Making

At the heart of AI-based policy development is the use of ML models to extract policy-related insights from the available datasets. In this direction, different types of ML models can be used. The latter models come with their pros and cons. To ensure transparent and regulatory compliant policy development, it is of uttermost importance to present models that can be explained. The following paragraphs illustrate different types of ML models for policy making, including explainable models and our suggested approach.

3.1. Black-Box Models

In general, Machine Learning algorithms can break down in two major categories, including: (i) traditional Machine Learning models, which are mostly based on classical statistical algorithms; and (ii) Deep Learning (DL) which comprises models based on artificial neural networks with many layers of "perceptrons" that are inspired from models of brain neuron cell operation

Machine Learning algorithms can be further divided into Supervised, Unsupervised and Reinforcement Learning. Supervised Learning algorithms are the most used in policy development problems. They use as training data examples of input vectors along with their corresponding target variables. Their aim is to predict some target variables. Supervised learning is further divided into two major sub-categories called regression and classification depending on whether the target variable is continuous or categorical. Some of the most prominent examples of supervised learning techniques, include linear regression, logistic regression (for classification) K-Nearest Neighbors Algorithms (K-NN) (classification), and Support Vector Machines (SVMs) (works for both regression and classification problems). When the training data examples consist of input vectors without corresponding targets the problem is an unsupervised one. The aim in such a problem is to match the input data to similar groups (clustering problem), or to determine the distribution that generates the input data (estimation). The final goal of unsupervised learning is to transform multi-dimensional data to fewer dimensions which are equally representative of the information present in the initial dataset. A popular technique for unsupervised learning is the K-means algorithm, which is an iterative algorithm that groups the data in clusters around specific centroids.

Artificial Neural Networks (also called multi-layer perceptron) is a subfield of machine learning which consists of algorithms inspired by the structure and the function of the human neuronal networks that comprise the brain. Prominent examples of Artificial Neural Networks (ANNs) are the Deep Learning models, i.e. multi-layer perceptrons of many layers (often more than 80) which are becoming increasingly important due to their scaling ability:performance, measured as any measure of accuracy in unseen test data, tends to increase with input training data volume increases. The increasing accuracy of ANNs contrasts with traditional ML models that often reach a plateau in performance even when very large volumes of data are available. Another great advantage of deep neural networks over traditional ML models is their ability to perform automatic feature extraction from raw data. Furthermore, ANNs can perform very well in classification and in regression tasks, while exceling in a variety of domains and inputs. Also, the inputs to ANNs is not restricted to tabular data. Rather it can expand to unstructured data such as images, text, and audio data. Modern state of art deep neural networks use a training technique based on gradient-based optimization to minimize the error on the training set, and use the fastest possible way to compute the gradient via a technique called back propagation, which is an instance of a more generic technique called "automatic differentiation". MLPs (Multilayer Perceptron Networks), CNNs (Convolutional Neural Networks) and LSTMs (Long Short-Term Memory Recurrent Neural Networks) are some of the most indicative deep neural networks using automatic differentiation for their learning phase (that aims to minimize the overall error on unseen data!). Despite their popularity, DL models have a major disadvantage when it comes to policy extraction: the resulting models are so complex that operate as black boxes which makes them non-transparent and essentially impossible to understand by any human.

3.2. Explainable AI Methods

The rising popularity in DL has led to a need for explaining and interpreting the workings and decisions made by an ANN trained classifier. There is a subtle but important difference in the semantics of the words "explainable" and "interpretable". An interpretable model is a model whose decision-making process is easily understood by a human even if the rationale behind the process might not be clear. For example, a classifier that outputs a decision tree is easy to replicate its decision on a particular input data instance, even without resorting to the use of a computer at all. All a human must do is follow the branches of the tree according to the values in the data instance until they reach a leaf node that is always labelled with a classification label. The process is fully interpretable, even if the reasons why the tree was constructed in this way may be completely incomprehensible to the user. An explainable model on the other hand, is a model that can somehow provide an explanation for its decision given a particular data instance. This begs the question "what constitutes an explanation?". Recently, a lot of research in XAI has focused in two directions: the first, called SHapley Additive exPlanations (SHAP) (Fryer, Strümke and Nguyen 2021), is based on the economic theory of the individual value a colleague brings to a collective effort of a finite number of individuals, pioneered by Nobel prize laureate Lloyd Shapley. The theory says that to measure how much is the fair share of a particular individual to the outcome of a collective effort, one needs to calculate the total outcome of the collaboration for any given subset of the collaborators, compute the gain that any such team will obtain when they add the particular individual to their ranks, and offer as fair share to the individual the average value of all these gains. This fair-share value is known as the Shapley value of the individual for the given collaboration group. Transferring this concept to the ML field, one can imagine the individual features contained in a tabular dataset as the collaborators in the final estimation of a classification label for a given data instance. Using the above theory, one can compute in theorythe importance of each feature in the decision-making process for a given data instance, as the Shapley value of that feature (in practice, due to the large number of computations that this process entails, only a sample of the possible "colleague configurations" is tested and averaged, so the value computed is usually only a statistical estimate of the true Shapley value of the feature). The SHAP methodology provides a framework that covers to a significant degree the second major direction in approaching explanations for black-box models, namely the Locally Interpretable Model-agnostic Explanations method (LIME) (Ribeiro, Singh, and Guestrin 2016). The LIME method essentially seeks to build a model "around" a new data instance and the relevant decision made by the black-box classifier. To this end, it chooses a number of instances "close" (according to some notion of distance) to the given data instance and applies the black-box model on those instances to get <data, black-box-label> pairs for a dataset in the neighbourhood of the original data instance. It then proceeds to build a simple interpretable model, such as a decision tree or a logistic regression model, on this just constructed dataset. Finally, it proposes the resultant model as a locally valid "interpretation" of the black-box behaviour in this neighbourhood.

Both methods have proved very useful in helping people decide the trustworthiness of complex models built on non-tabular data, such as text or images. For example, a trained deep neural network on a text corpus comprised of messages posted on "atheism" and "Christianity" newsgroups achieved a 94% accuracy on held-out test data. However, when questioned about the features that weighted the most on its decisions revealed that the presence of words such as "posting", "hosting" and "Keith", were key to deciding that the message was about atheism, which of course makes absolutely no sense. Indeed, when the same trained model was evaluated on a newer set of messages posted on the same newsgroups achieved an accuracy of approximately 58%, proving that the model was nearly useless. This demonstrates the dangers of blindly accepting non-explainable (black-box) models based just on accuracy measures in some contexts.

3.3. QARMA Family of Models

While the importance of estimating feature weights on the final decision made by a black-box classifier cannot be under-estimated, we claim that such feature weighting does not constitute intuitive, human-friendly explanations of the decisions of a model. In particular, in the case of LIME, every time an explanation is requested for a model decision, a new (interpretable) model must be built from scratch. This makes the process non-reactive as it is almost impossible to always respond in near real-time.

On the other hand, we find that explanations based on rules that hold with high confidence on the dataset are easy for people to grasp and correspond more intuitively to what people expect from an "explanation". For example, consider a case where a customer's bank-loan application is rejected. When faced with the following explanations, which is more likely to be understood by the customer as more appropriate?

1. "The importance of the feature 'declared bankruptcy within the last year' is highest among all other features"
2. "Your application was rejected because you declared bankruptcy within the last year, and from our records, with probability 99%, if an application has the 'declared bankruptcy' check-box ticked, the loan is not paid-back on time"

The second explanation, citing a rule that holds with high accuracy on the dataset provides enough evidence as to the reason behind the rejection. Therefore, it constitutes a sufficient explanation for the user. In all cases, including the rule-based offered explanation, the real workings of the black-box model that made its decision might be completely different from what is offered as an explanation. This must be expected since the black-box model is by definition opaque and we have no way of knowing how it works.

To offer rule-based explanations, we need to know all rules that apply on a given dataset. Restricting our attention to tabular data only, we apply QARMA, a family of highly parallel/distributed algorithms that extract all non-redundant quantitative association rules that hold with at least a certain user-defined minimum support and confidence on the dataset (Christou, 2019). QARMA produces all valid, and non-redundant rules of the forms:

$$a_k \in [l_k, h_k] \wedge \dots a_m \in [l_m, h_m] \rightarrow t \geq L$$

$$a_i \in [l_i, h_i] \wedge \dots a_j \in [l_j, h_j] \rightarrow t \leq H$$

for regression tasks, and

$$a_p \in [l_p, h_p] \wedge \dots a_q \in [l_q, h_q] \rightarrow t = v$$

for classification tasks.

Each of the above rules are guaranteed to hold with minimum support and confidence on the dataset. The variables $a_i, \dots a_q$ are input features in the dataset, and t is the target variable. The rules, once extracted, are permanently stored on a (relational) database.

Having obtained all the rules that hold on the dataset, the system offers explanations given a new pair of a data instance plus black-box prediction on that instance.

For a classification problem, the system scans the entire ruleset in the database and collects all rules for which the <data instance, prediction> pair satisfies both their antecedent conditions as well as the consequent. From this collected set, the rule(s) with maximum confidence (and maximum support, in case of ties) is presented to the user as explanations of the black-box prediction.

For a regression task, we collect again all rules which obey the <data instance, prediction> pair, and from this set we pick as explanations up to two rules, one of which constrains the value of the target variable from above, and the other from below. From the set of all rules that are satisfied by the <data, prediction> pair with the black-box model prediction being the equality "$t = v_c$", we pick the rule that predicts an inequality of the form "$t \geq v_{mx}$" with v_{mx} being the largest value appearing in the consequent of any rule of this form that is still less than or equal to the black-box predicted value v_c; ties are broken in favor of the rule with the highest confidence, then in favor of the rule with highest support. Similarly, we pick the rule that predicts an inequality of the form "$t \leq v_{mn}$" with v_{mn} being the smallest value appearing in the consequent of any rule of this form that is still greater than or equal to the black box predicted value v_c; ties are broken as already mentioned.

The resulting rule(s) are immediately understood by humans and constitute a much more intuitive explanation of the black-box decision. In case of the black-box model erring, such rules can help trace the error in the dataset, and in the statistics of the dataset (the rules that hold on it) that lead to the wrong decision. Furthermore, offering combinations of feature quantifications that lead to the problematic decisions. Hence, we claim that such rules can lead to greater insight into the source of the problem than individual feature importance metrics can offer.

3.4. QARMA Suitability for Regulatory Compliant Policy Development

In the light of the AI Act, policy makers leveraging AI for data-driven policy development and policy extraction, must perform a risk assessment of their systems. The result of the assessment indicates whether a system is of high, low or medium risks This is then used to drive the ML model selection requirements as illustrated in the following table (Table 1), which illustrates a recommendation for the ML model to be used for each one of the different outcomes of the risk assessment.

Risk Assessment Outcome	Explainability Requirement	Recommendation
High Risk	Mandatory Explainability	QARMA, LIME, SHAP
Medium Risk	Optional Explainability	DL Model or QARMA
Low Risk	No Explainability Requirement	Any DL/ML Model

Table 1. Mapping Explainability Requirements to Different ML Models

QARMA's ability to provide both explainable and high-performance policy extraction makes it an ideal choice for high-risk AI use cases that deal with critical decisions. Considering that most public policies concern decisions of high potential risk, QARMA has broad explainability even in cases of DL models that might yield better performance than QARMA in specific datasets. Note also that QARMA can be used in conjunction with DL models to boost their explainability, as illustrated in the following section.

4. Validation in Real Policy Making Cases

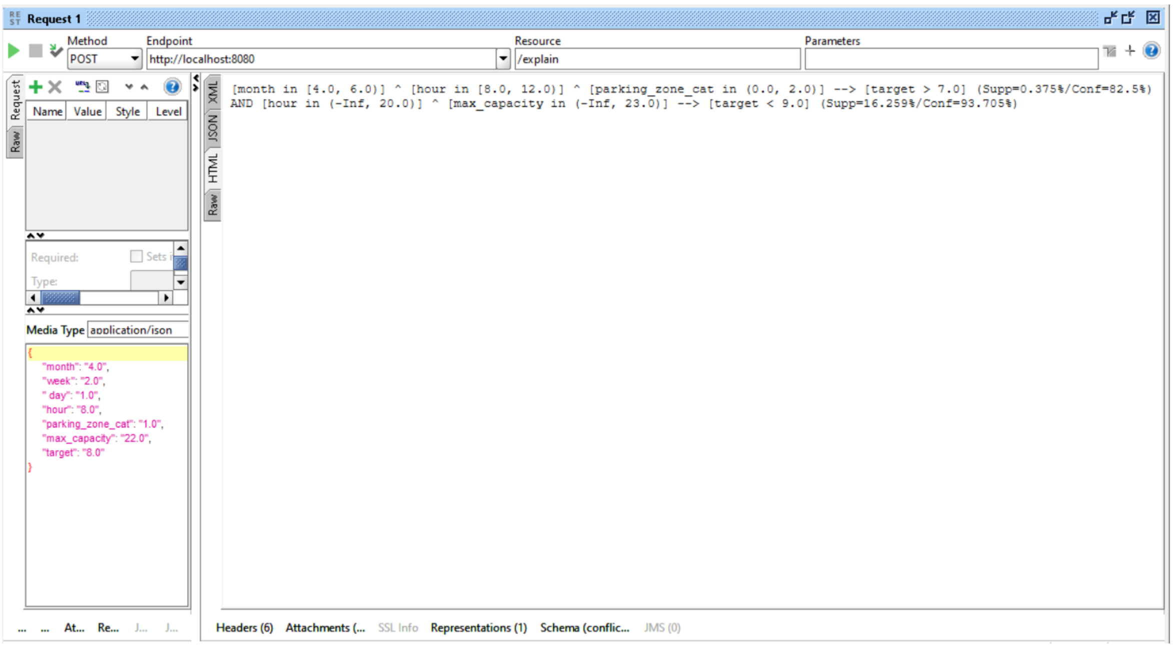
4.1. The AI4PublicPolicy Platform as a Validation Testbed

To validate our explainable and regulatory compliant approach to public policy making, we have leveraged the capabilities of the AI4PublicPolicy platform that has been introduced in Section 2. Specifically, we have used the platform to experiment with different datasets and ML models in a variety of use cases. The platform's data collection and management APIs have been used to acquire datasets from legacy data sources, including systems and file collections used by local governments and other policy making organizations. The various datasets have been registered to the dataset catalogue component and used to test different ML models that led to new policy models. The latter have been also integrated in the platform's catalogue.

To test, validate and evaluate different ML models, we have relied in the CRISP-DM methodology as explained in Section 2. Based on CRISP-DM different ML models (including our QARMA family of algorithms) have been tested in terms of their performance, accuracy, and predictive power over the available datasets. The best performing models have been accordingly explained and presented to end-users (i.e., policy makers) leveraging the policy explainability and interpretation module. Some modules of the platform (e.g., the AutoML and semantic interoperability modules) have not been used in the implementation of the validating use cases that are presented in the following sections.

4.2. Smart Parking Policies

To validate our QARMA approach for public policy making, we applied the algorithm on a parking space availability dataset provided by one of the municipalities (smart cities) that participate in the AI4PublicPolicy project. Specifically, we have ran the QARMA algorithm and extracted all rules that hold on the dataset that have as consequents rules of the form $target \geq v$ or alternatively $target \leq v$ where the variable $target$ represents the number of available parking slots in a particular zone i.e., a well-defined geographic area within the limits of the municipality. At the same time, we also trained a deep neural network to learn to predict this target variable given values for the input features, which are hour-of-day, day-of-month, month-of-year, particular zone etc. We then created a REST (Representational State Transfer) application that listens to resource "/explain" for HTTP POST requests with a JSON body containing the values for a data instance, together with the predicted value for that instance by a black-box deep neural network. Fig. 4 shows a particular REST API call and the response received by this call. The latter shows two rules that perfectly explain the decision of the neural network. Once our REST web-app has loaded the rules from the database (145,224 rules constraining the target value from below, and 129,236 rules constraining the target value from above), it takes approximately 3 seconds to respond over HTTP to any request, making the web-app fully interactive with a human user trying to understand the decisions of the neural network.



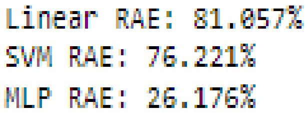
```
{
   "month": "4.0",
   "week": "2.0",
   " day": "1.0",
   "hour": "8.0",
   "parking_zone_cat": "1.0",
   "max_capacity": "22.0",
   "target": "8.0"
}
```

```
[month in [4.0, 6.0)] ^ [hour in [8.0, 12.0)] ^ [parking_zone_cat in (0.0, 2.0)] --> [target > 7.0] (Supp=0.375%/Conf=82.5%)
AND [hour in (-Inf, 20.0)] ^ [max_capacity in (-Inf, 23.0)] --> [target < 9.0] (Supp=16.259%/Conf=93.705%)
```

Fig 4: Calling QARMA as a REST web-app to explain black-box model decisions for Smart Parking Policies

The explanations provided by the proposed rules are providing deeper and clearer intuition than a sorted list of features in order of importance in the case of the SHAP method, or the approximate local (usually linear) model that LIME offers. What is more important, the rules come with support and confidence values associated with them. Therefore, a small confidence value for either of the two rules, or consequent values that are far from the black-box prediction, are strong indicators that the black-box prediction should not be trusted very much. In the worst case, there will be no rule in support of the black-box prediction. In such a case the prediction should not be trusted, especially for high-stake decisions, such as those found in legal or health related use-cases. The latter are also the types of use cases that would be classified as high-risk according to the AI Act of the European Parliament and the council of Europe.

In the available parking spaces prediction problem, a new dataset was also constructed. It included variables capable of providing a possible correlation with the free parking spaces variable. For the problem, three models were developed with their complexity as the main criterion. The models included a simple linear regression, an SVM model modified to work for regression, and an MLP. The relative absolute error was used as an evaluation metric. Since the metric is an error, the less the value of the RAE metric the better the model performs.



```
Linear RAE: 81.057%
SVM RAE: 76.221%
MLP RAE: 26.176%
```

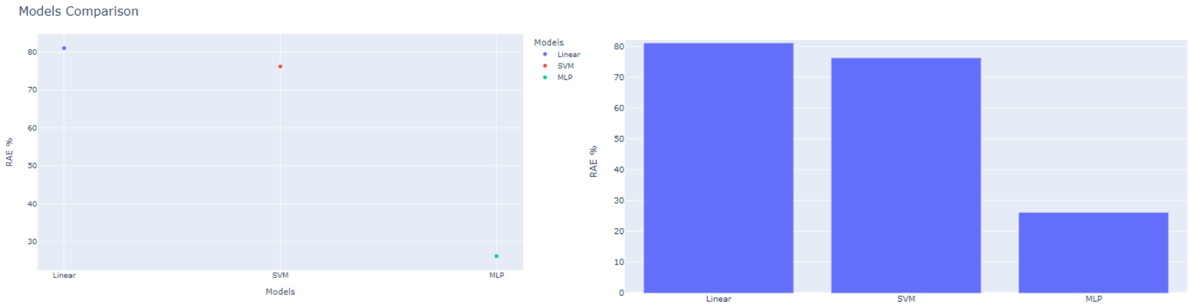
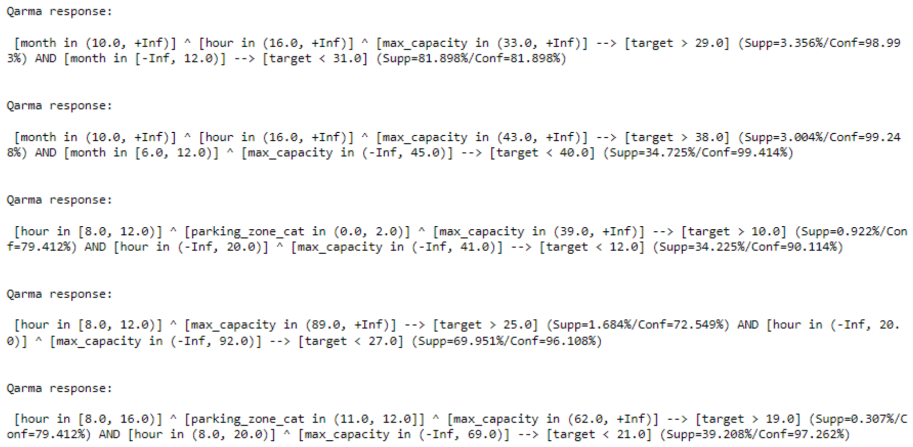


Fig 5: Comparing Alternative ML Models for Smart Parking Policies Extraction

As is evident, the MLP neural network not only outperformed the rest of the models, it achieved a very good RAE that was very close to zero. The results from the solution of the available parking spaces indicate that a future enrichment of the dataset could lead to much higher performing models. For a better understanding of the results a bar plot visualization constructed where the models RAE percentage is more clearly compared (Fig. 5). The MLP neural network gives by far better results than the other models.

Moreover, the QARMA explainable model has been applied to the dataset to explain the deep learning model's predictions, to give a clearer explanation of why the MLP model made a certain prediction, and to indicate which values of the other dataset features played a major role in the decision (Fig. 6). Given our predictions and the test dataset used for the prediction, QARMA produced a set of 2 rules each time in the form of antecedents and consequent. The ranges of the feature values explain that the dataset features range between these values when the model outputs the current value for the target variable. Therefore, QARMA can be also used in conjunction with a DL approach, as a means of explaining a black-box model.



```
Qarma response:

 [month in (10.0, +Inf)] ^ [hour in (16.0, +Inf)] ^ [max_capacity in (33.0, +Inf)] --> [target > 29.0] (Supp=3.356%/Conf=98.99
3%) AND [month in [-Inf, 12.0)] --> [target < 31.0] (Supp=81.898%/Conf=81.898%)


Qarma response:

 [month in (10.0, +Inf)] ^ [hour in (16.0, +Inf)] ^ [max_capacity in (43.0, +Inf)] --> [target > 38.0] (Supp=3.004%/Conf=99.24
8%) AND [month in [6.0, 12.0)] ^ [max_capacity in (-Inf, 45.0)] --> [target < 40.0] (Supp=34.725%/Conf=99.414%)


Qarma response:

 [hour in [8.0, 12.0)] ^ [parking_zone_cat in (0.0, 2.0)] ^ [max_capacity in (39.0, +Inf)] --> [target > 10.0] (Supp=0.922%/Con
f=79.412%) AND [hour in (-Inf, 20.0)] ^ [max_capacity in (-Inf, 41.0)] --> [target < 12.0] (Supp=34.225%/Conf=90.114%)


Qarma response:

 [hour in [8.0, 12.0)] ^ [max_capacity in (89.0, +Inf)] --> [target > 25.0] (Supp=1.684%/Conf=72.549%) AND [hour in (-Inf, 20.
0)] ^ [max_capacity in (-Inf, 92.0)] --> [target < 27.0] (Supp=69.951%/Conf=96.108%)


Qarma response:

 [hour in [8.0, 16.0)] ^ [parking_zone_cat in (11.0, 12.0]] ^ [max_capacity in (62.0, +Inf)] --> [target > 19.0] (Supp=0.307%/C
onf=79.412%) AND [hour in (8.0, 20.0)] ^ [max_capacity in (-Inf, 69.0)] --> [target < 21.0] (Supp=39.208%/Conf=97.262%)
```

Fig 6: Explaining a Deep Neural Network for Smart Parking Policies via QARMA derived rules

4.3. Infrastructure Maintenance Policies

A second validation scenario was run on infrastructures' maintenance datasets. Specifically, the problem that needed to be resolved was the "issue prediction" problem for one of the municipalities that participate in the AI4PublicPolicy project. The municipality receives "issues" for maintenance problems via an automatic online application for its services that are provided by different authorities in the municipality. After preprocessing the dataset with the maintenance cases an attempt was made to predict the number of the issues each authority service would receive within a certain month. To predict several issues for every month, the time series patterns of the dataset were taken into consideration. The three previously outlined models were again developed and tested. The comparison of the models illustrated that a neural network model performs better than the rest. However, although its predictions are relatively good, there is a need to enrich the dataset for the models to solve the problem optimally.

5. Conclusions

In the era of technology acceleration, public policy makers are provided with unprecedented opportunities to collect and manage digital data from a variety of different channels, including citizens' touch points, e-services, and social networks. These data enable a transition to data-driven, evidence-based policy making based on the use of ML and AI technologies. Nevertheless, the use of AI in public policy making is still in its infancy, as challenges associated with transparency, explainability and bias alleviation are not fully addressed yet. Moreover, public policymakers must comply with emerging AI regulations. This paper has presented and discussed these challenges, along with potential solutions at the AI system development and ML modelling levels. Our main value propositions lie in the introduction of blueprints for developing AI systems for policy developments as well as in the application of the QARMA ML framework for the extraction of explainable policies. QARMA provides a very good balance between performance and explainability, which makes it appropriate for use in high-risk policy making decisions. Specifically, it is very good choice for cases where high-performance deep learning models must be explained by means of a surrogate model.

Overall, the paper presented an approach to explainable and regulatory compliance policy development, based on AI technologies. The approach has been already validated using smart parking and infrastructure maintenance datasets. Our validation has proven a dual merit for QARMA i.e., both as high-performance ML model and as an explainability tool during the policy extraction process. We are currently working on expanding the scope of the validation in other public policy domains, such as energy management on smart buildings, handling of citizens' complaints and requests, as well as urban mobility optimization.

Acknowledgments. Part of this work has been carried out in the scope of the H2020 AI4PublicPolicy project, which is titled: "Automated, Transparent Citizen-Centric Public Policy Making based on Trusted Artificial Intelligence". The authors acknowledge support and contributions from all partners of the project.

Funding statement. Part of this work has been carried out in the scope of the AI4PublicPolicy Project, which has received funding from the European Commission under the H2020 programme (contract No. 101004480). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests. None

Data availability statement. None – Data and Code of QARMA are proprietary.

Author contributions. Reference Architecture and Data Mining Process: C. Ipektsidis; A. Amicone; QARMA Specification and Implementation: I. Christou; Introduction and Mapping to Regulatory Requirements: J. Soldatos. Validation on Smart Parking and Infrastructure Maintenance Datasets: T.Papadakis. Writing original draft: All authors. End-to-End Editing of the Paper: J. Soldatos. All authors approved the final submitted draft.

References

A. Androutsopoulou, Y. Charalabidis (2018). A framework for evidence based policy making combining big data, dynamic modelling and machine intelligence. In Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance, Galway, Ireland, 2018.

Bellamy, Rachel KE, et al. (2018). "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." arXiv preprint arXiv:1810.01943 (2018).

Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, and Wirth R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Charalabidis, Y., Maragoudakis, M., Loukis, E. (2015). Opinion Mining and Sentiment Analysis in Policy Formulation Initiatives: The EU-Community Approach. In: , et al. Electronic Participation. ePart 2015. Lecture Notes in Computer Science(), vol 9249. Springer, Cham. https://doi.org/10.1007/978-3-319-22500-5_12

P. Chauhan and M. Sood. (2021). "Big Data: Present and Future," in Computer, vol. 54, no. 4, pp. 59-65, April 2021, doi: 10.1109/MC.2021.3057442.

I.T. Christou (2019). "Avoiding the hay for the needle in the stack: Online rule pruning in rare events detection", in IEEE Intl. Symposium on Wireless Communication Systems (ISWCS), Special Session on IIoT, pp. 661-665, Oulu, Finland.

I. T. Christou, N. Kefalakis, A. Zalonis and J. Soldatos (2020) "Predictive and Explainable Machine Learning for Industrial Internet of Things Applications," IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 213-218, doi: 10.1109/DCOSS49796.2020.00043.

I.T. Christou, N. Kefalakis, J. Soldatos, A.-M. Despotopoulou (2022). “End-to-end industrial IoT platform for Quality 4.0 applications”, Computers in Industry, 137:103591.

Curry, E., Metzger, A., Berre, A., Monzón, A., and Boggio-Marzet, A. (2021). A Reference Model for Big Data Technologies. 10.1007/978-3-030-68176-0_6.

Jefrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marcaurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, and Andrew Y. Ng. (2012). Large Scale Distributed Deep Networks. In NIPS ’12.

Changyu Deng, Xunbi Ji, Colton Rainey, Jianyu Zhang, Wei Lu. (2020). “Integrating Machine Learning with Human Knowledge”, iScience, Volume 23, Issue 11, 2020, 101656, ISSN 2589-0042, https://doi.org/10.1016/j.isci.2020.101656.

Edwards, L.; and Veale, M. (2018). Enslaving the algorithm: From a “Right to an Explanation” to a “Right to Better Decisions”? IEEE Security & Privacy 16(3): 46–54.

Eggers, W. D., Schatsky, D., & Viechnicki, P. (2017). AI-augmented government. Using cognitive technologies to redesign public sector work. Retrieved July,7 2021, from https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/artificial-intelligencegovernment.html.

Fryer, D.V., Strümke, I., & Nguyen, H. (2021). Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. IEEE Access, 9, 144352-144360.

Tanja Sophie Gesk, Michael Leyer (2022). “Artificial intelligence in public services: When and why citizens accept its usage”, Government Information Quarterly, 2022, 101704, ISSN 0740-624X, https://doi.org/10.1016/j.giq.2022.101704

Joseph Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, and Carlos Guestrin (2012). PowerGraph: Distributed Graph-parallel Computation on Natural Graphs (OSDI’12). 17–30.

Karen Hao 2019. “This is how AI bias really happens—and why it’s so hard to fix”, MIT Technology Review, February 4, 2019, available: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

J. Hochtl, P. Parycek, and R. Schollhammer (2016), Big data in the policy cycle: Policy decision making in the digital era, Journal of Organizational Computing and Electronic Commerce. 26, 147–169, 2016.

John Carlo Bertot, Paul T. Jaeger ⁎, Derek Hansen 2011, The impact of polices on government social media usage: Issues, challenges, and recommendations, 2011.

Norman P. Jouppi (2017), et. Al. In-Datacenter Performance Analysis of a Tensor Processing Unit. In Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA ’17). ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3079856.3080246.

Diederik P. Kingma and Jimmy Ba (2014), Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980 (2014).

Huamaní, Enrique Lee, Alva Mantari, and Avid Roman-Gonzalez (2020). “Machine Learning Techniques to Visualize and Predict Terrorist Attacks Worldwide Using the Global Terrorism Database.” International Journal of Advanced Computer Science and Applications 11, no. 4 (2020).https://doi.org/10.14569/IJACSA.2020.0110474

M. Leyer (2021), S. Schneider, Decision augmentation and automation with artificial intelligence: Threat or opportunity for managers?, Business Horizons, 64 (5) (2021), pp. 711-724.

Pérez López, César, María Delgado Rodríguez, and Sonia de Lucas Santos (2019). “Tax Fraud Detection through Neural Networks: An Application Using a Sample of Personal Income Taxpayers.” Future Internet 11, no. 4 (March 30, 2019): 86. https://doi.org/10.3390/fi11040086.

Lindgren, C.Ø. Madsen, S. Hofmann, U. Melin (2019), Close encounters of the digital kind: A research agenda for the digitalization of public services, Government Information Quarterly, 36 (3) (2019), pp. 427-436

Marban, O., Mariscal, G., and Segovia, J. (2009). A Data Mining & Knowledge Discovery Process Model. In J. Ponce, & A. Karahoca (Eds.), Data Mining and Knowledge Discovery in Real Life Applications. IntechOpen. https://doi.org/10.5772/6438

Monogan III, James E. (2015). Political Analysis Using R, Springer. http://link.springer.com/book/10.1007%2F978-3-319-23446-5

Plotnikova V., Dumas M., and Milani F. (2020). Adaptations of data mining methodologies: a systematic literature review. PeerJ Computer Science 6:e267 https://doi.org/10.7717/peerj-cs.267

Tara Qian, and Rony Medaglia (2019). “Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare.” Government Information Quarterly 36, no. 2 (April 2019): 368–83. https://doi.org/10.1016/j.giq.2018.09.008.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI:https://doi.org/10.1145/2939672.2939778.

M. Rosemann, J. Becker, F. Chasin (2020), City 5.0. Business & Information Systems Engineering (2020), pp. 1-7.

Rudin, Cynthia (2019). “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” ArXiv:1811.10154 [Cs, Stat], September 21, 2019. http://arxiv.org/abs/1811.10154.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17). JMLR.org, 3145–3153.

John Soldatos (ed.), Dimosthenis Kyriazis (ed.) (2021), "Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production", Boston-Delft: now publishers, http://dx.doi.org/10.1561/9781680838770.

Suresh, Harini and John V. Guttag (2019), “A Framework for Understanding Unintended Consequences of Machine Learning.” ArXiv abs/1901.10002 (2019): n. pag.

Harini Suresh and John Guttag. (2021), A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, 2021, --, NY, USA. ACM, New York, NY, USA 9 Pages. https://doi.org/10.1145/3465416.3483305