


RESEARCH ARTICLE

Explainable and transparent artificial intelligence for public policymaking

Thanasis Papadakis¹, Ioannis T. Christou^{1,2}, Charalampos Ipektisidis¹, John Soldatos¹  and Alessandro Amicone³

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

²The American College of Greece, Athens, Greece

³GFT Italia Srl, CU Innovation, Genova, Italy

Corresponding author: John Soldatos; Email: John.Soldatos@netcompany.com

Received: 04 August 2023; **Revised:** 15 December 2023; **Accepted:** 28 December 2023

Keywords: AI Act; artificial intelligence; explainable AI; machine learning; public policy

Abstract

Nowadays public policymakers are offered with opportunities to take data-driven evidence-based decisions by analyzing the very large volumes of policy-related data that are generated through different channels (e.g., e-services, mobile apps, social media). Machine learning (ML) and artificial intelligence (AI) technologies ease and automate the analysis of large policy-related datasets, which helps policymakers to realize a shift toward data-driven decisions. Nevertheless, the deployment and use of AI tools for public policy development is also associated with significant technical, political, and operation challenges. For instance, AI-based policy development solutions must be transparent and explainable to policymakers, while at the same time adhering to the mandates of emerging regulations such as the AI Act of the European Union. This paper introduces some of the main technical, operational, regulatory compliance challenges of AI-based policymaking. Accordingly, it introduces technological solutions for overcoming them, including: (i) a reference architecture for AI-based policy development, (ii) a virtualized cloud-based tool for the specification and implementation of ML-based data-driven policies, (iii) a ML framework that enables the development of transparent and explainable ML models for policymaking, and (iv) a set of guidelines for using the introduced technical solutions to achieve regulatory compliance. The paper ends up illustrating the validation and use of the introduced solutions in real-life public policymaking cases for various local governments.

Policy Significance Statement

This paper illustrates a collection of AI solutions that can empower data scientists and policymakers 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 policymaking, along with solutions that alleviate these challenges. The introduced solutions have been validated in the development of various data-driven, evidence-based policies for local governments.

1. Introduction

1.1. Artificial intelligence for public policymaking: the rationale

In recent years, artificial intelligence (AI) is disrupting entire economic sectors such energy, transport, industry, healthcare, smart cities, and public administration. This disruptive power of AI stems from its

potential to improve the automation, efficiency, and speed of business processes in these sectors. The latter improvements are propelled by advances in parallel hardware (e.g., Jouppi et al., 2017) and scalable software systems (e.g., Gonzalez et al., 2012), which have in turn enabled the development of advanced machine learning (ML) frameworks (e.g., Dean et al., 2012) and novel optimization algorithms (e.g., Kingma and Ba, 2014) suitable for large-scale problems. These technical advances enable organizations to process and analyze vast amounts of data in timely and cost-effective ways (Leyer and Schneider, 2021). They also facilitate organizations to use data analytics outcomes to improve, optimize, and accelerate their decision-making processes.

In most cases, the AI-based processing and analysis of large datasets is based on ML models and algorithms. The latter facilitate digital systems to learn without human instruction (e.g., rules) but based on historical information and statistical knowledge. ML is extremely popular in use cases that involve the processing and analysis of vast amounts of information. This is the case for modern policymaking processes, given that policymakers are nowadays provided with more policy-related datasets than ever. In recent years, data volumes increase at an unprecedented rate (Chauhan and Sood, 2021), which challenges policymakers and public administration workers. Specifically, as part of their decision-making processes, policymakers must collect, read, study, analyze, and process vast amounts of data. ML systems can automate the processing of large volumes of data toward extracting policymaking insights. Thus, they can help policymakers to benefit from the growing data volumes of policymaking data, without worrying about data scalability challenges. Furthermore, ML ensures that policymakers consider all relevant information via correlation and cross-analysis of multiple datasets. Therefore, there are also cases when ML models and technologies enable novel capabilities based on the identification of potentially hidden patterns and correlations of policy-related datasets. For instance, ML algorithms can detect non-obvious associations between different policymaking parameters such as citizens' behavioral patterns after specific events or during specific time windows, beyond what is already known or observed by policymakers. The identification of such patterns can enable policy optimizations that are hardly possible based on legacy information processing methods.

Overall, ML can improve the way public policies are developed, which is the reason why this paper presents challenges and methods for ML-based policy development. ML technologies enable new methods for understanding and analyzing data, which can lead policymakers to better decisions (Edwards and Veale, 2018). Moreover, ML models come to provide a layer of human ingenuity, which is currently missing from existing policymaking models (Deng et al., 2020). Therefore, the use of ML as a public policymaking tool is increasingly considered in the strategic agendas of public policy organizations such as central, regional, and local governments (e.g., Lindgren et al., 2019; Rosemann et al., 2020).

1.2. The challenges of ML-based policy development

There is a clear rationale behind using ML models and tools in policymaking scenarios. However, the use of ML in practical policymaking scenarios entails the following challenges:

- **Algorithmic bias:** One of the most prominent problems with the development ML applications is the development of biased algorithms (Hao, 2019; Suresh and Gutttag, 2021). Therefore, policymakers must ensure that their policymaking algorithms are unbiased and lead to representative policies rather than producing policies that favor specific citizen groups and exclude others. However, the development of unbiased systems is very challenging due to the lack of representative policymaking datasets. For instance, digital data collected in social media are not typically representative of the preferences and needs of citizen groups that do not use social media platforms (e.g., elderly or low-income citizens). Hence, training ML algorithms for policy decision-making using such biased data is likely to result in policy decisions that do not account for the needs of such citizen groups. It is also noteworthy that algorithmic bias is often introduced in unintended ways (Suresh and Gutttag, 2019), which asks for bias detection and mitigation processes. As a prominent example, the training of ML-based decision-making systems based on historical datasets that

comprise bias is one of the common ways for injecting human biases within ML-based policy models. Unless there are some bias detection systems in place, such biased systems can be deployed in production and lead to biased policies.

- **Explainability and transparency:** Public policymakers must be able to understand how their data-driven decisions work, while at the same time explaining the rationale behind these decisions to citizens. Moreover, policymaking processes must be transparent and trustworthy to be accepted by citizens. Unfortunately, the many high-performance ML models (e.g., models based on deep neural networks) operate as black boxes and offer very limited transparency and explainability (Soldatos and Kyriazis, 2021). This makes their use in pragmatic policymaking settings very difficult. To remedy this situation, the research community is working on explainable AI (XAI) techniques, which aim at building AI systems that are transparent (i.e., “glass box” models) and at explaining how black box algorithms work. XAI techniques can be used to explain to policymakers why an AI system recommends a specific policy decision. This is particularly important for the practical deployment of AI systems in policy development settings as policymakers won’t use AI systems that cannot understand how they work. XAI is also a key prerequisite for the transparency of ML-based policy decisions, as these decisions cannot be transparent unless stakeholders understand how AI systems recommend specific decisions. Nevertheless, the use of XAI in public policymaking settings is still in its infancy. Even the very concept of what constitutes a good explanation is still under debate currently. Recent works on explainable ML methods for policy development (Amarasinghe et al., 2023) underline the importance of contextualizing ML explanations and highlight the limitations of existing XAI techniques. Furthermore, they highlight the importance of stakeholder’s engagement and the need to prioritize policymakers’ requirements rather than relying on technology experts to produce explanations for ML-based policies (Bell et al., 2023). Moreover, there are no agreed and proven ways for selecting models that balance performance and explainability in line with the requirements of policymakers.
- **Regulatory compliance:** ML-based systems for public policymaking must comply with emerging regulations in AI, such as the AI Act of the European Parliament and Council of Europe (European Commission, 2021). The AI Act is globally one of the first structured and systematic efforts to regulate AI systems. It defines stringent requirements for high-risk AI systems that are used to drive crucial decisions, like most public policymaking decisions. These requirements include guarantees for transparency, explainability, data quality, and human oversight. 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.

1.3. Related work

Despite increased interest in the use of AI for data-driven, evidence-based policymaking (Geske 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 policymaking that leverages Big Data (Hochtl et al., 2016). Some systems for data-driven policies take advantage of social media information (Bertot et al., 2011) based on different techniques, including data mining and ML algorithms (Charalabidis et al., 2015). There are also works on political analysis using statistical techniques, which is a foundation for ML (Monogan III, 2015). More recently, the use of ML 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 areas like taxation (López et al., 2019), public security and counterterrorism (Huamaní et al., 2020), public work design (Eggers et al., 2017), and healthcare (Qian and Medaglia, 2019).

These systems have provided insights on the benefits and challenges of ML-based policymaking. Addressing bias and explainability challenges are crucial 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 et al., 2016; Fryer et al., 2021; Shrikumar et al., 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 policymaking 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 introduces XAI solutions for evidence-based policy development. Moreover, it 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 ML-based public policy use cases, as well as by challenges associated with regulatory compliance. It introduces a data mining process, a reference architecture (RA), a cloud-based ML platform, and a framework for explainable ML algorithms for dealing with these issues in the context of policymaking. Specifically, we first introduce an RA and a data mining framework that could boost the development of robust and unbiased systems for public policymaking. The RA illustrates a set of modules and tools that are destined to support policymakers in adopting, using, and fully leveraging AI/ML techniques in their policymaking efforts. On top of the RA, a data mining process based on the popular CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is adapted to support public policymaking activities. The process makes provisions for explorative data analysis to detect and remedy potential biases linked to the used datasets. Moreover, a virtualized cloud-based platform is presented, which enables the development of ML-based policymaking applications in line with the RA and the CRISP-DM process.

From an algorithmic viewpoint, the paper presents the adaption and application of 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 policymaking applications for smart cities. 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 policymaking without any essential performance penalty. It also allows for a whole new class of explanations for black box models' decisions. The QARMA ML framework forms the core of the XAI module of the presented platform, which enables it to produce explainable rules that are transparent and understandable to policymakers.

Another contribution of paper is the mapping of different algorithmic tools (including QARMA) to different AI-based policymaking 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 policymaking, who wish to develop regulatory compliant systems by design. In the scope of the paper, the risk-driven approach of the AI Act is positioned in the wider landscape of enterprise risk management (ERM) frameworks and their use for assessing the risks of AI applications.

It should be noted that an earlier and reduced version of this paper has been presented in the scope of the "Data for Policy" conference, which was held in Brussels in December 2022. This earlier version has been published as an open access paper (Papadakis et al., 2022). The present version of the paper extends this earlier edition by introducing the virtual policy management environment (VPME), as well as by presenting additional validation use cases.

The paper is structured as follows:

- **Section 2** following this introduction introduces the data mining process and the RA for public policymaking. It also illustrates how ML models can drive the evidence-based policymaking process, along with the virtualized cloud-based platform for developing AI-based policies.

- **Section 3** presents our arsenal of ML algorithms and tools for public policymaking use cases. These algorithms are explainable and are used to empower the operation of the XAI module of the platform. It also discusses how these tools (including the XAI techniques) map to the requirements of the different risk levels of the AI Act, as well as how they could be used in conjunction with popular ERM frameworks.
- **Section 4** illustrates the use of the introduced data mining process and XAI algorithms in real-life public policymaking 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. RA and virtualized policy management environment for ML-based policies

To address existing gaps in the development of AI-based use cases for public policymaking, we herewith introduce a RA and a data mining framework for the development of robust and unbiased systems for public policymaking. The presented RA and the accompanying data mining process are integrated in a single platform, which is also presented in later paragraphs. This platform has been 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 collaborated closely towards unveiling AI's potential for automated, transparent, and citizen-centric development of public policies. The RA of the AI4PublicPolicy platform is specified as a set of software modules and tools that aim at supporting policymakers 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. In this direction, a cloud-based platform for developing such policies has been developed as well.

2.1. Platform RA and main components

The AI4PublicPolicy platform collects and analyzes data that are used for automated, transparent, and citizen-centric development of public policies. The architecture is inspired by the Big Data Value reference model (Curry et al., 2021). It is characterized as a *RA* 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 modeling, extraction, simulation, and recommendations. The AI tools include ML 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 (NLP) 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-

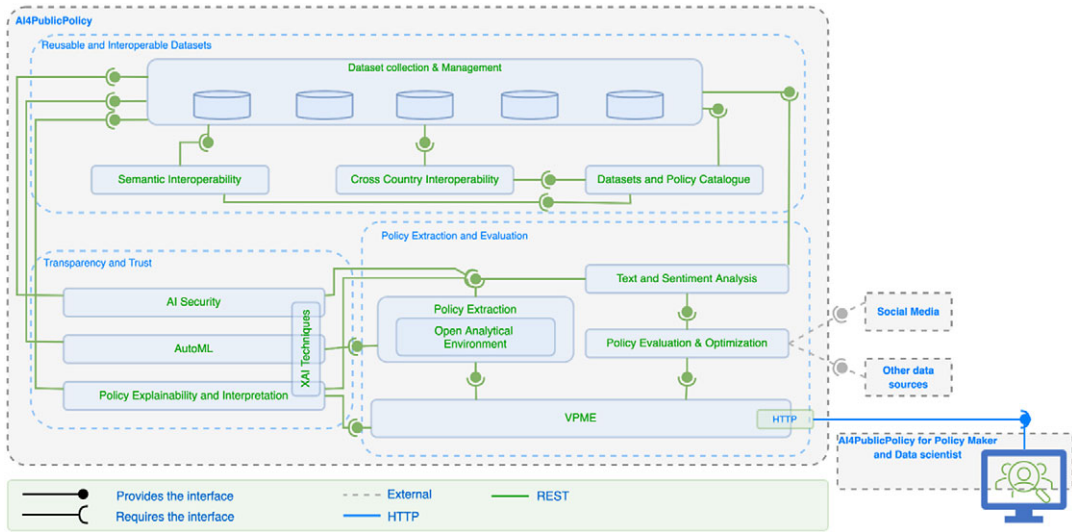


Figure 1. *AI4PublicPolicy architecture components.*

performance computing resources. EOSC is a federated system based on a set of existing research infrastructures that delivers a catalog 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 policymaking. This environment enables the sharing of datasets and models for data-driven policies, as part of a cloud-based 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 XAI models that make ML-based policies explainable to humans to enhance transparency and the overall acceptability of policies by citizens and policymakers.

Our blueprint architecture for implementing and deploying AI-based policy development is illustrated in Figure 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 catalog:** This component is a directory of policies and datasets, which facilitates the dynamic discovery of data and policies toward facilitating reuse.
- **XAI techniques:** This module is responsible for providing information for analyzing and explaining the ML 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-defense 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 NLP and text analytics algorithms.
- **Policy extraction:** Enables the policymaker to choose an AI workflow from a catalog of ML and deep learning (DL) workflows and apply it. The workflows are available in an open analytical environment (e.g., Project Jupyter¹) that enables the development and deployment of ML applications.
- **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.
- **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 RA 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 policymaking 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 et al. 2009) has been properly adapted. Specifically, the six phases² of CRISP-DM are used to support the policy development process as follows (Figure 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 modeling techniques followed by calibration of their parameters. In conjunction with the AI4PP architecture, it is facilitated by the catalog 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 policymakers.

In the context of the presented RA, 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 policymakers, 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 policymaking process

Using the AI4PublicPolicy platform and the CRISP-DM process, policymakers (e.g., governmental officials) can benefit from a novel data-driven policymaking process, which is illustrated in Figure 3.

¹ <https://jupyter.org>.

² <https://www.datascience-pm.com/crisp-dm-2/>.



Figure 2. CRISP-DM phases and key outputs³ (Creator: Kenneth Jensen⁴).

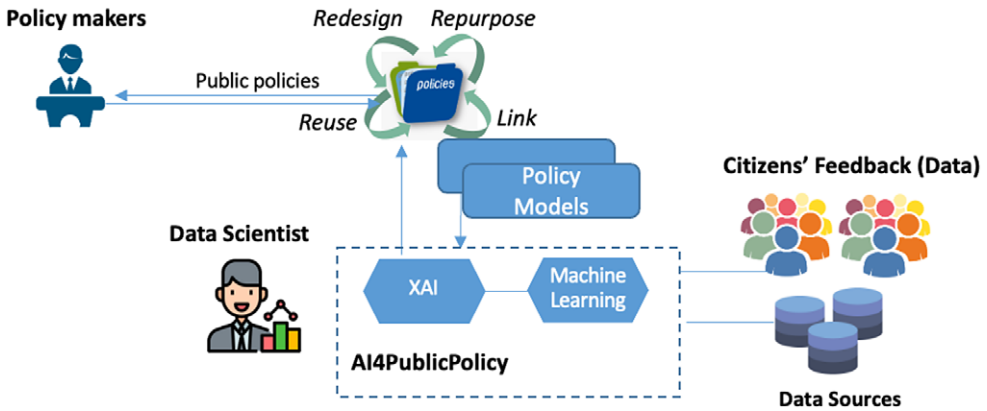


Figure 3. Policymaking process using the AI4PublicPolicy platform.

The process involves the development of ML models based on data from a variety of sources including citizen feedback. These ML models are enhanced with domain-specific metadata to enable the production

³ Licensed under the Creative Commons Attribution-Share Alike 3.0.

⁴ Own work based on: <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf>.

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 policymaker 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).

Key to the implementation of the policymaking process of Figure 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, and 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 policymakers to understand and use it. As a prominent example, the QARMA family of explainable ML algorithms that are used for the validating use cases in Section 4 expresses policy models in the form of easy-to-understand rules.

2.4. The VPME for ML-based policies

The VPME is an integrated cloud-based platform for managing datasets and policies. It is developed in line with the presented RA and supports the CRISP-DM policy development process. Figure 4 illustrates the home page of the platform, which offers the functionalities for datasets and policies management, including uploading of datasets, extraction of policies based on ML-models, and validation of policies based on citizens' feedback. These functionalities match some of the building blocks of the presented RA.

2.4.1. Datasets and policy management

The dataset and policy management functionality of the platform enables data scientists and policymakers to specify and upload datasets from different sources to the platform. Using these datasets, it is possible to define different policies. Specifically, datasets are used to train different AI models and implement policies. It persists datasets and policies within a cloud database. Moreover, it offers APIs that enable the development of policies. A web-based interface is provided to enable non-technical users to define dataset schemas and policies, to upload datasets in the platform, and to visualize policies and datasets.

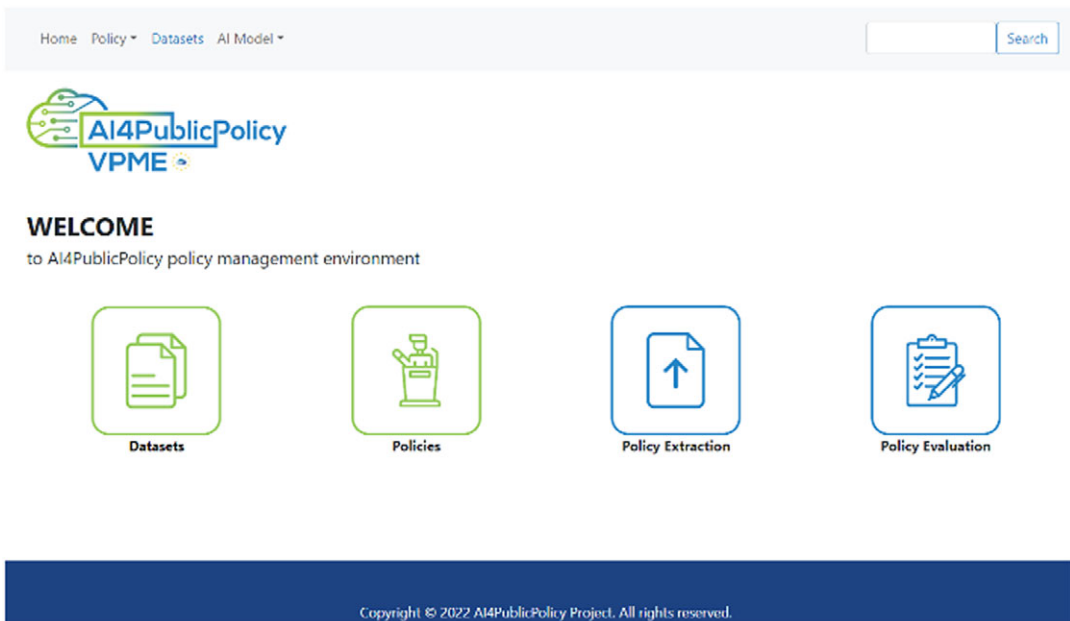


Figure 4. Home page of the VPME—main entities.

The screenshot displays the 'Policy Definition' section of a web application. At the top, there is a navigation bar with links for 'Navbar', 'Home', 'Policy', 'Datasets', and 'AI Model', along with search boxes. Below the navigation bar is the 'AI4PublicPolicy' logo. The main heading is 'Policy Definition', followed by the sub-heading 'Register a new policy'. The form consists of several sections:

- Area:** A dropdown menu with the placeholder text '-- Select an area --'.
- Policy Name:** A text input field.
- Owner:** A text input field.
- Policy description:** A large text area for entering the policy's details.
- Objective:** A large text area for defining the policy's purpose.
- Policy Datasets:** A section titled 'SELECT THE DATASETS TO BE USED IN THIS POLICY' with an 'Add Dataset' button. It includes a 'Dataset' dropdown menu with the placeholder 'Select from available datasets' and a small 'X' button. A note below states: 'Indicate the dataset to be used in the policy from the available ones. This field can be empty and be filled latter.'
- Policy KPIs definition:** A section titled 'DEFINE THE KEY INDICATORS TO ESTIMATE FOR THIS POLICY' with an 'Add KPI' button. It features a dropdown menu with the placeholder 'KPI #1'.

 At the bottom left of the form is a blue 'Submit' button.

Figure 5. Policy definition user interface (UI).

2.4.2. Policy definition and management

The platform offers policy definition functionalities based on a proper form-based interface (Figure 5). A policy is defined in terms of the name and a description of the policy, the owner of the policy, and the purpose of the policy. It also allows end users to associate datasets to policies. Moreover, users can indicate key performance indicators (KPIs) for the policy. Following the definition of a policy, end users can also change some of its properties. Policies can be listed and filtered according to their sector such as energy and transport policies.

2.4.3. Cross-country interoperability and policy sharing

The cross-country interoperability and policy sharing functionalities of the VPME enables the translation of policies and datasets from one language to another (Figure 6). It is destined to facilitate the reuse of different policies and datasets in countries with different languages. Once a translation is made, the information is stored internally to facilitate its repurposing and reuse.

2.4.4. Policy extraction toolkit

The VPME integrates a policy extraction toolkit, which enables the extraction of data-driven policies based on available ML algorithms. The latter are to estimate and recommend the parameters of the policy models based on the available policy datasets. In this direction, a web interface for the policy extraction is



POLICIES

Policies in: english ▾

Optimized parking space allocation [More information](#) Translate to greek ▾ [Delete](#)

Προσδιορίστε περιοχές στάθμευσης στο δρόμο με υψηλά/χαμηλά αιτήματα στάθμευσης, έτσι ώστε οι θέσεις στάθμευσης μεταξύ των κατοίκων και των επισκεπτών να μπορούν να κατανεμηθούν πιο αποτελεσματικά για τη βέλτιστη χρήση του χώρου και τα υψηλά έσοδα για την πόλη.

Datasets:

Maintenance Incidents Prediction for Planning Purposes [More information](#) Translate to english ▾ [Delete](#)

Predict maintenance incidents in the next months to plan material purchases and plan seasonal personnel.

Datasets:

Optimized parking space allocation

Objective: Οι διαθέσιμες θέσεις στάθμευσης στην Αθήνα επισημαίνονται στο δρόμο και κατανέμονται μεταξύ κατοίκων της πόλης (απαιτείται ειδική άδεια) και επισκεπτών (πληρωμή μέσω εφαρμογής στάθμευσης). Ως εκ τούτου, με αυτήν την πολιτική ο Δήμος Αθηναίων θα ήθελε να διαθέσει πιο αποτελεσματικά τους διαθέσιμους χώρους μεταξύ κατοίκων και φιλοξενουμένων, ώστε να μην μένουν αχρησιμοποίητοι ή λείπουν χώροι κάθε εποχή.

KPIs

Name	Description
Ποσοστό κάλυψης θέσεων στάθμευσης	
Έσοδα από θέσεις στάθμευσης επισκεπτών	
Ικανοποίηση πολιτών	

Figure 6. Policies translation user interface.

provided, which offers two main views, a view for AI experts and a view for policymakers. The “AI expert” view empowers AI experts to select a project and work on it (Figure 7). Likewise, the policymaker view enables policymakers to test AI models for a specific policy KPI.

2.4.5. Policy evaluation toolkit

The VPME offers a policy evaluation toolkit, which enables the simulation and evaluation of the developed AI policies. In particular, it integrates different types of surveys based on different electronic channels. The surveys are linked to action strategies that can be implemented at different stages of an ML pipeline. A policymaker view and a stakeholder view are provided. Through their view, policymakers manage and create custom surveys, including application surveys and surveys in social media platforms such as Twitter (Figure 8). Moreover, policymakers can view the results of the surveys using an

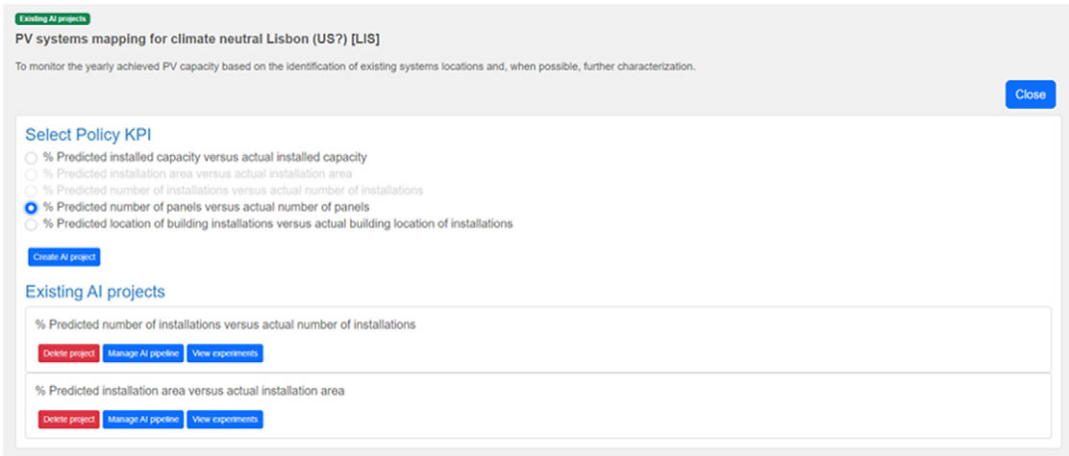


Figure 7. AI expert view in the policy extraction toolkit: projects for policy extraction.

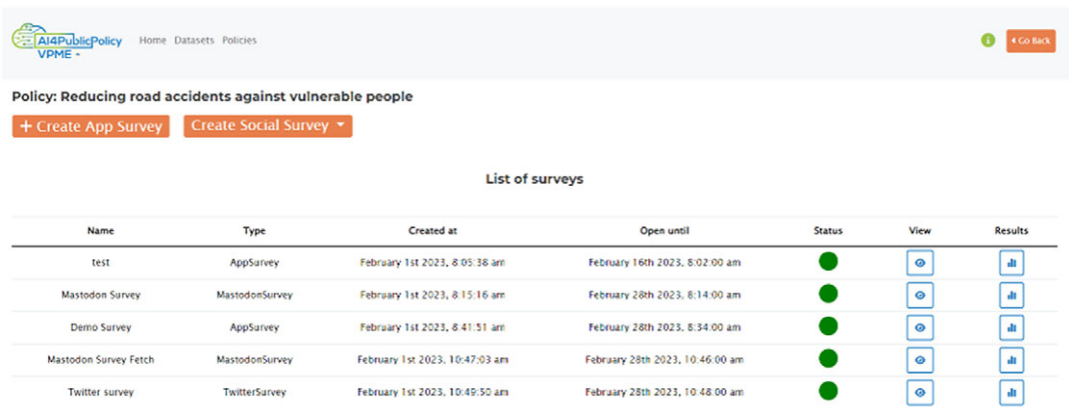


Figure 8. List of surveys in the VPME.

appropriate dashboard. Stakeholders (e.g., citizens) can participate to the surveys in order to provide their evaluation feedback about the policies.

3. ML for public policymaking: From black box to explainable models

AI-based policy development relies on ML models that extract policy-related insights from available datasets. To this end, different types of ML models can be used, each one with its own pros and cons. The AI4PublicPolicy platform leverages XAI techniques for extracting policy insights that are transparent and understandable by policymakers. Specifically, the project’s XAI techniques can be used to explain how black box ML models operate. This enables the platform to integrate black box ML models within its ML module, while at the same time explaining them via XAI techniques that are integrated into the XAI module of the platform (Figure 3).

3.1. Black box models

ML algorithms can be broken down into two major categories: (i) traditional ML models, which are mostly based on classical statistical algorithms, and (ii) DL, which comprises models based on artificial neural

networks (ANNs) with many layers of “perceptrons” that are inspired from models of brain neuron cell operation. ML algorithms can be further divided into supervised, unsupervised, and reinforcement learning (Bonaccorso, 2017). Supervised learning algorithms are the most used in policy development problems. They are used as training data examples of input vectors along with their corresponding target variables. Their aim is to predict some target variables. 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 (i.e., clustering problem) or to determine the distribution that generates the input data.

ANNs is the subfield of ML which consists of algorithms inspired by the structure and the function of the neuronal networks of the human brain. As already outlined, prominent examples of ANNs are the DL models, i.e., multi-layer perceptrons, which are very popular due to their scaling ability. Specifically, their 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 reach a plateau in performance even when very large volumes of training data are used. 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 regression tasks, while exceling in a variety of domains and inputs. The inputs to ANNs are not restricted to tabular data. Rather they can expand to unstructured data such as images, text, and audio data.

Modern state-of-the-art deep neural networks use 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.” Multilayer perceptron networks (MLPs), convolutional neural networks, and long short-term memory recurrent neural networks are some of the most indicative deep neural networks using automatic differentiation for their learning. Despite their popularity, DL models have a major disadvantage when it comes to policy extraction. Specifically, the resulting models are so complex that operate as black boxes. This makes them non-transparent and essentially impossible to understand by any human.

3.2. XAI methods

The rising popularity of DL has driven a need to explain and interpret the workings and decisions made by ANN trained classifiers. 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 labeled 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 direction is based on the so-called SHapley Additive exPlanations (SHAP) (Fryer et al., 2021) approach. SHAP is based on the economic theory of the individual value a colleague brings to a collective effort of a finite number of individuals. The SHAP theory indicates that in order 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 a 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 theory the 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. Hence, 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 (LIME) method (Ribeiro et al., 2016). The LIME method 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 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 neighborhood 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 behavior in this neighborhood.

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 makes absolutely no sense. Indeed, when the same trained model was evaluated on a newer set of messages posted on the same newsgroups, it 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.

3.3. *QARMA ML models for XAI*

While the importance of estimating feature weights on the final decision made by a black box classifier cannot be underestimated, we claim that such feature weighting does not constitute intuitive, human-friendly explanations of a model’s decisions. In particular, in the case of the LIME method, 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, 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” or
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 cites a rule that holds with high accuracy on the dataset and 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 there is no way to know exactly 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:

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

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

for regression tasks and

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

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, QARMA collects again all rules which obey the <data instance, prediction> pair, and from this set, it picks 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 " $t = v_c$," it picks 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, it picks 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 cases black box models, 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, it offers combinations of feature quantifications that lead to problematic decisions. Hence, such rules can provide deeper insight into the source of the problem than individual feature importance metrics can offer.

The XAI module of the AI4PublicPolicy platform (Figure 3) is powered by the QARMA ML framework. This enables the platform to produce policy insights in the form of explainable rules that can be understood and interpreted by policymakers. Overall, the project's platform provides the means for explaining DL techniques. This enables policymakers to leverage the high-performance and accuracy of ANNs in practical settings.

3.4. AI policy development in line with the AI Act

The deployment and use of AI systems is associated with several risks, including, for example:

- Bias and fairness risks which are associated with AI algorithms that lead to unfair and discriminatory outcomes (Srinivasan and Chander, 2021; Suresh and Guttag, 2021), in applications like hiring, loan approvals, and criminal justice.

- Security risks, which are associated with AI systems that are vulnerable to attacks and manipulation. For instance, there are adversarial attacks that can mislead AI models causing them to produce wrong outputs (Apruzzese et al., 2020). Such attacks take place by introducing subtle changes to the input data of the model.
- Privacy risks, which often arise as a result of the unauthorized use of privacy-sensitive data during the training of AI algorithms (Rahman et al., 2023).
- Safety risks, as errors in AI predictions or decisions of autonomous systems (e.g., robots, drones) can have serious health and safety implications for their users (e.g., industrial workers) (Westhoven, 2022).

This is gradually leading organizations to assess the risks of their AI deployments based on risk management frameworks such as the NIST framework (NIST, 2021). In recent years, framework for AI risk assessment have also emerged such as the ISO/IEC 23894:2023 standard (ISO/IEC, 2023). The latter helps organizations to integrate risk management in their AI-related activities. At the regulatory forefront, the AI Act is also adopting a risk-driven approach to the regulatory compliance of AI deployment. Specifically, according to the AI Act, policymakers that use ML for data-driven policy development must perform a risk assessment of their AI systems. The result of the assessment indicates whether a system is of high, low, or medium risk. Systems of different risk levels then subject to varying explainability requirements for their AI models as illustrated in Table 1. The table presents a recommendation for the ML model to be used for each one of the different outcomes of the risk assessment.

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 policy decisions. Most public policies concern high-risk decisions, which means that QARMA has a very broad applicability. Specifically, QARMA can be used in high-risk applications, even in cases where DL models yield better performance than QARMA in specific datasets. This is because DL models are generally not explainable, which hinders their use in high-risk use cases according to the AI Act. Moreover, it is also possible to use QARMA in conjunction with DL models to boost their explainability, as illustrated in the following section.

4. Validation in real policymaking cases

4.1. The AI4PublicPolicy platform as a validation testbed

To validate our explainable and regulatory compliant approach to public policymaking, we have leveraged some of the capabilities of the AI4PublicPolicy platform and the VPME environment that have been introduced in Section 2. Specifically, we have used the platform to experiment with different ML models and datasets 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 policymaking organizations. The various datasets have been registered to the dataset catalog and used to test different ML models that led to new policy models. The latter have been also integrated in the platform's catalog. To test, validate, and evaluate different ML models, we have relied on the CRISP-DM. Different ML models (including QARMA) have been tested in terms of their performance, accuracy, and predictive power over the available datasets. In this direction, the ML and

Table 1. Mapping explainability requirements to different ML models

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

XAI modules of the platform have been leveraged (Figure 3) over real-life datasets of the use cases. The best performing models have been accordingly explained and presented to end users (i.e., policymakers).

4.2. Smart parking policies (Athens, Greece)

To validate our QARMA approach for public policymaking, we applied the algorithm on a parking space availability dataset provided by the city of Athens, Greece, which participates in the AI4PublicPolicy project. Specifically, we have ran the QARMA algorithm and extracted all rules that hold on the dataset that have consequent 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 geographic area of the municipality. 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 (Hypertext Transfer Protocol) POST requests with a JSON (JavaScript Object Notation) body containing the values for a data instance, together with the predicted value for that instance by a black box deep neural network. Figure 9 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 few 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.

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

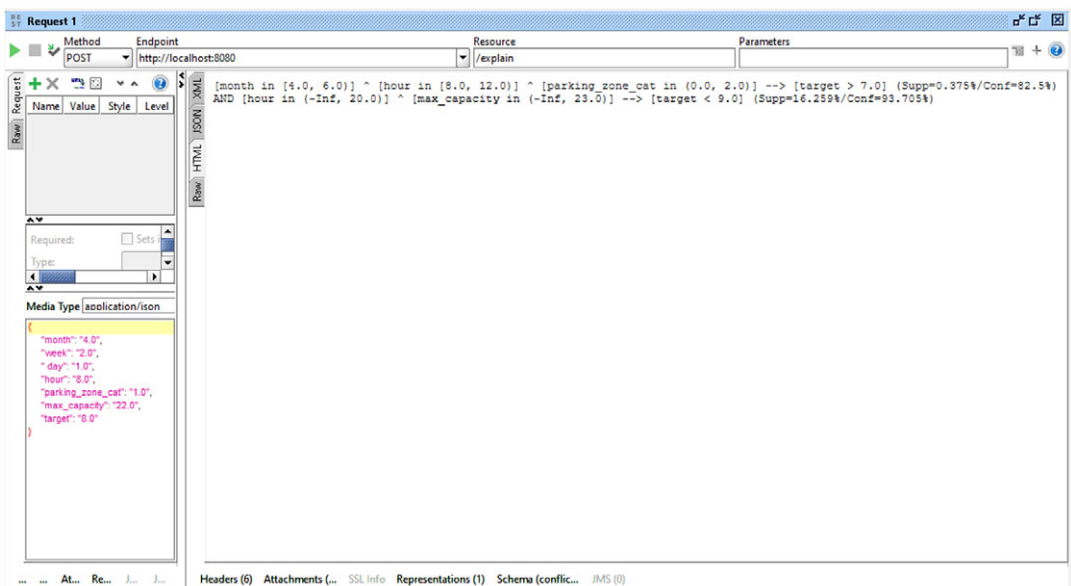


Figure 9. Calling QARMA as a REST web app to explain black box model decisions for smart parking policies.

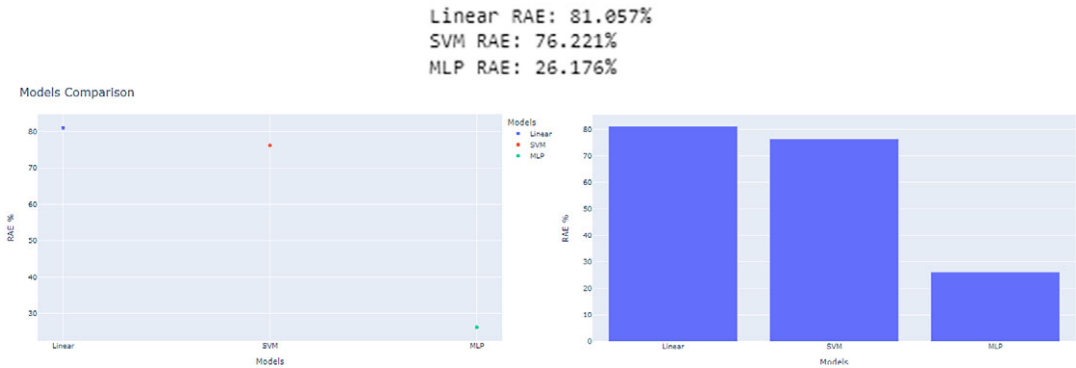


Figure 10. Comparing alternative ML models for smart parking policies extraction.

(e.g., 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 space prediction problem, a new dataset was also constructed. It included variables capable of providing a possible correlation with the free parking spaces variable. For this problem, three models were developed with their complexity as the main criterion. The models included a simple linear regression, a support vector machine model modified to work for regression, and a MLP. The relative absolute error (RAE) 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.

As is evident, the MLP neural network not only outperformed the rest of the models, it also 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 higher performing models. For a better understanding of the results, a bar plot visualization was constructed where the models RAE percentage is more clearly compared (Figure 10). 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 DL 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 (Figure 11). 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 consequents. 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. Hence, QARMA can be also used in conjunction with a DL approach to explain a black box model.

4.3. Infrastructure maintenance policies (Burgas, Bulgaria)

A second validation scenario was run on infrastructures' maintenance datasets. Specifically, the problem that needed to be resolved was the "issue prediction" problem for the municipality of Burgas that participates 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. The first pre-processing step is to cluster the data for both visualization and prediction purposes. The municipality of Burgas set forth the following rules to be obeyed by any clustering process:

- (i) any two issues that are located up to 80 m apart are considered to be "connected," i.e., must belong to the same cluster.
- (ii) a cluster is a maximal set of points that has the property that any two points in the set are connected via at least one path of points such that any two consecutive points in any such path are less than 80 m apart.

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=98.993%) 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=99.248%) 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%/Conf=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%/Conf=79.412%) AND [hour in (8.0, 20.0)] ^ [max_capacity in (-Inf, 69.0)] --> [target < 21.0] (Supp=39.208%/Conf=97.262%)
```

Figure 11. Explaining a deep neural network for smart parking policies via QARMA-derived rules.

This implies that a hierarchical agglomerative clustering (HAC) approach with single-linkage clustering (Sibson, 1973) which runs until all coarsest clusters are found will determine the sought-after clusters. The problem is equivalent to finding all connected components in a graph. However, the HAC with single-linkage approach is an unsupervised clustering approach that allows for visualization of the intermediate clusters that are created before the final ones. Given the clustering, our next step is to compute the areas of highest density of issues over time: every cluster gives rise to a geographic area defined as the convex hull of the points in the cluster. The density of the cluster is the number of points in the cluster divided by its area. Our system “predicted” as next issue area the convex hull of the cluster with the highest density, for any cluster of size greater than 1.

4.4. Water infrastructure management policies (Burgas, Bulgaria)

This validation scenario concerns policies for water distribution infrastructure maintenance. Specifically, the policies consider water leakage detection on water pipes for the municipality of Burgas, which faces recurrent leakages in their water distribution network infrastructure. The goal of the policy is to optimize the repair of the leakages in ways that maximize the uptime of the water infrastructure. As a first step toward detecting such leakages, an experiment was set up based on data collection from a single vibration sensor (see Figure 12). Several scenarios were explored, in which some taps were opened to simulate a leakage in the pipe (see Figure 13). The collected dataset provided vibration measurements along with the timestamp at which the measurements were taken, and an indication of whether a tap was open or not. This last binary variable (0/1) served as the ground truth (label) of the dataset, i.e., forming a pure classification problem. From these values, a dataset was constructed, part of which can be seen in Figure 14. The analysis of this dataset did not produce meaningful visualizations. Therefore, the next AI pipeline module was triggered, which performed additional data engineering on the dataset. Based on this, various vibration patterns were found to change under leakage conditions (Figures 15 and 16). Because of the very high sampling frequency, the initial experiments with a simple pre-processing of the data did not produce decent results, which led to experimentation with another pre-processing approach. A sliding time window of 30000 points (corresponding to 5 seconds of wall-clock time) was defined, and 33 points were finally kept as follows: 999 points out of every 1000 points were dropped, keeping only the first one. This “downsampling” resulted in a dataset in which only 30 data points exist for every 5 seconds (i.e., spread one sixth of a second apart).

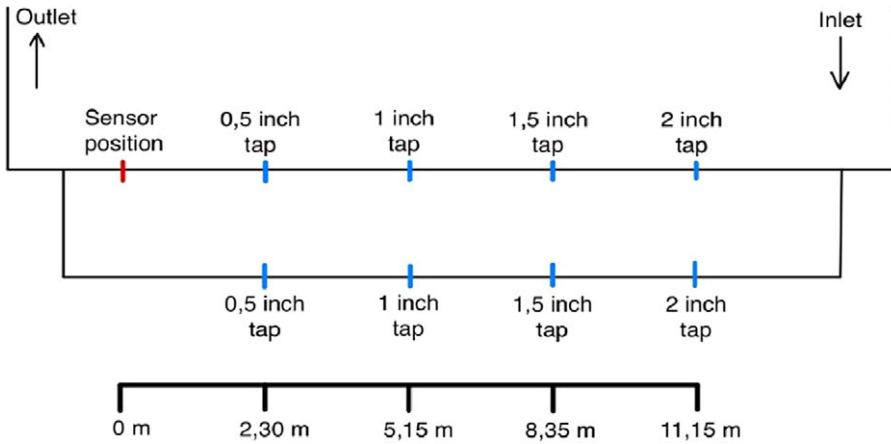


Figure 12. Experimental water pipe setup.

Scenario	Conditions							Starting Time			
0	All Taps Closed (ATC)							12:03			
	9 min 57 s										
1	ATC		Tap1AB		ATC			13:00			
	3 min 2 s		3 min 5 s		2 min 18 s						
1a	ATC		Tap1SLOpening		Tap1Open		ATC	13:18			
	2 min		31 s		1 min 33 s		2 min 8 s				
2	ATC		Tap2AB		ATC			13:25			
	3 min 16 s		2 min 46 s		1 min 46 s						
2a	ATC		Tap2SLOpening		Tap2Open		ATC	15:17			
	3 min 14 s		27 s		1 min 45 s		2 min 13 s				
3	ATC		Tap3AB		ATC			15:25			
	3 min		2 min 4 s		2 min 10 s						
3a	ATC		Tap3AB		ATC			15:33			
	3 min 11 s		2 min 6 s		2 min 11 s						
4	ATC		Tap4AB		ATC			15:41			
	3 min 11 s		2 min 20 s		2 min 8 s						
5	ATC		Tap2AB		Tap2Open + Tap1AB		Tap1Open	ATC	15:49		
	3 min 4 s		2 min 9 s		2 min 11 s		2 min 13 s	2 min 5 s			
6	ATC		Tap4AB		Tap4Open + Tap1SLOpening		Tap1Open + Tap4Open	Tap1Open	ATC	16:04	
	3 min 2 s		2 min 15 s		15 s		1 min 56 s	2 min 59 s	2 min 12 s		
7	ATC		Tap4AB		Tap4Open + Tap2AB		Tap2Open + Tap4Open + Tap1AB	Tap1Open + Tap2Open	Tap1Open	ATC	16:22
	3 min 30 s		1 min 16 s		1 min 10 s		1 min 1 s	1 min 3 s	1 min 12 s	1 min 2 s	

Figure 13. Water management experiment details.

In the scope of the experiment, we kept track of the minimum, maximum, and average value of all the 30000 points in the time window, for a total of 33 numbers. Therefore, every sliding window was summarized in a vector of 33 dimensions. Since a sliding time window is used, the total number of rows in such a dataset would be $N-30000$, where N is the total number of vibration measurements. Since N is measured in the billions, the number $N-30000$ as a number representing the total size of a training time series is still too big for ML training on single machine (e.g., personal computer). For this reason, we further downsampled our data and kept only 1 out of every 100 of the created vectors, which resulted in a downsampling by a factor of 100, i.e., keeping 1% of all the vectors. The resulting dataset led to neural network models that obtained 88% overall accuracy and a macro average F1-score of 86% and weighted average F1-score of 88% (Figure 17). Finally, the confusion matrix and the actual numbers and

	time	modulus	open_tap	scenario
	0	0.066571	-9.611669	0 Scenario0
	1	0.066720	-6.223372	0 Scenario0
	2	0.066870	0.355992	0 Scenario0
	3	0.067019	1.031545	0 Scenario0
	4	0.067169	-4.878226	0 Scenario0

	38977979	614.401238	-1.779636	0 Scenario7
	38977980	614.401388	-8.641136	0 Scenario7
	38977981	614.401537	-0.379339	0 Scenario7
	38977982	614.401686	0.443084	0 Scenario7
	38977983	614.401836	-7.216029	0 Scenario7

Figure 14. Initial constructed dataset.

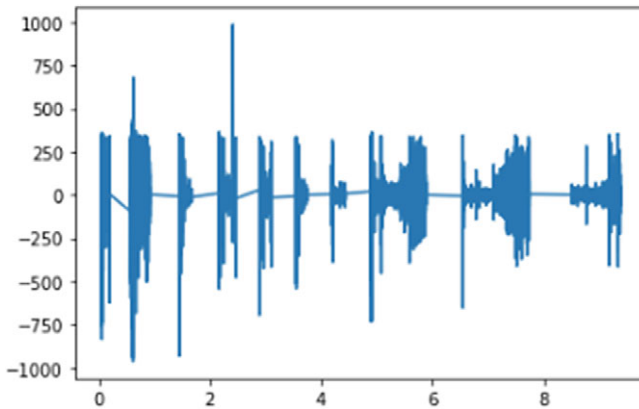


Figure 15. Opened taps—vibration visualization.

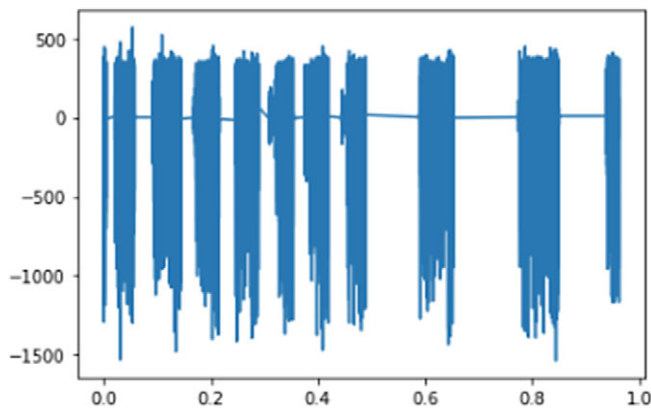


Figure 16. Closed taps—vibration visualization.

	precision	recall	f1-score	support
0	0.92	0.91	0.91	40933
1	0.80	0.81	0.81	18414
accuracy			0.88	59347
macro avg	0.86	0.86	0.86	59347
weighted avg	0.88	0.88	0.88	59347

Figure 17. Classification report—leakage neural network results.

percentages presented at Figure 18 show that the ANN achieved very good overall performance, even with pre-processing with data from a single sensor that measured the water’s vibration.

On another set of experimental data with 5 target classes corresponding to different open-tap scenarios of the same water pipe, using 2 vibration sensors, we were able to achieve much higher accuracy (approximately 98%) and F-1 score (approximately 97%) when using explainable rule-based methods rather than neural networks and cross-validation instead of train-test splitting of the dataset (see Figure 19). In particular, in this second dataset, rule-learning methods such as RIPPER, Decision Table, and QARMA achieve an estimated accuracy on unseen data that is always above 98%; an ANN with a single hidden layer achieved an accuracy of 82%.

```

True Positive number: 14942
True Positive Percent: 25.177346790907713 %

True Negative number: 37251
True Negative Percent: 62.76812644278566 %

False Positive number: 3682
False Positive Percent: 6.204188922776215 %

False Negative number: 3472
False Negative Percent: 5.850337843530422 %
    
```

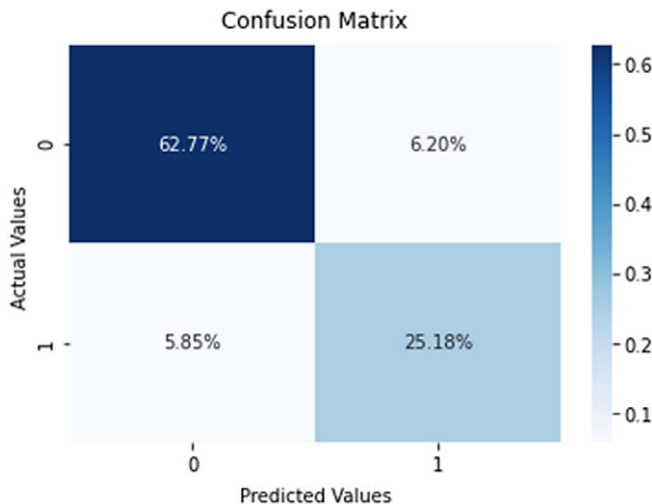


Figure 18. Confusion matrix of the ML model predictions regarding water leakages.

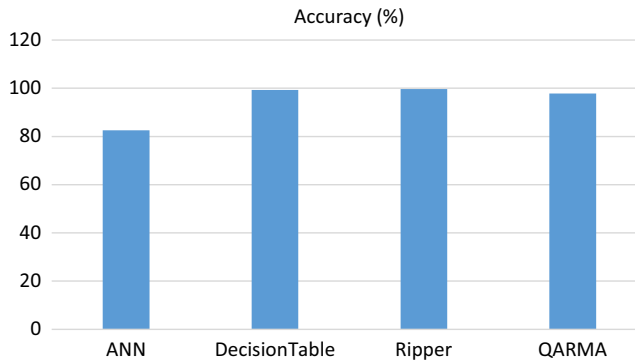


Figure 19. Accuracy comparison between ML models second dataset containing water leakages for Burgas municipality (bigger is better).

5. Conclusions

Nowadays, policymakers 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 policymaking based on the use of ML and AI technologies. Nevertheless, the use of AI in public policymaking 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 such as the AI Act. This paper has presented and discussed these challenges, along with potential solutions at the AI system development and ML modeling 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 policymaking decisions. Specifically, it is a very good choice for cases where high-performance DL 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, infrastructure maintenance, and water infrastructure management 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. Coupled with the presented VPME, the QARMA family of algorithms provides a powerful toolset for policymakers that aspire to adopt and fully leverage ML in their policy development processes.

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

Acknowledgements. 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. An earlier and reduced version of this paper has been presented in the scope of the “Data for Policy” conference, which was held in Brussels, on December 13, 2022.

Author contribution. Reference architecture, data mining process, and VPME specification and implementation: C. Ipektsidis, A. Amicone. QARMA specification and implementation: I. T. Christou. Introduction and mapping to regulatory requirements: J. Soldatos. Validation on smart parking and infrastructure maintenance: T. Papadakis, I. T. Christou. Smart water management datasets: T. Papadakis, I. T. Christou. Writing—original draft: all authors. End-to-end editing of the paper: J. Soldatos. All authors approved the final submitted draft.

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 interest. The authors declare no competing interests exist.

References

- Amarasinghe K, Rodolfa K, Lamba H and Ghani R** (2023) Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy* 5, E5. <https://doi.org/10.1017/dap.2023.2>
- Androutopoulou A and Charalabidis Y** (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 (ICEGOV '18)*, Association for Computing Machinery, New York, NY, USA, 575–583. <https://doi.org/10.1145/3209415.3209427>.
- Apruzzese G, Andreolini M, Marchetti M, Venturi A and Colajanni M** (2020) Deep reinforcement adversarial learning against botnet evasion attacks. *IEEE Transactions on Network and Service Management* 17(4), 1975–1987. <https://doi.org/10.1109/TNSM.2020.3031843>
- Bell A, Nov O and Stoyanovich J** (2023) Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance. *Data & Policy* 5, E12. <https://doi.org/10.1017/dap.2023.8>
- Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S, Ramamurthy KN, Richards J, Saha D, Sattigeri P, Singh M, Varshney KR and Zhang Y** (2018) AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. Preprint, *ArXiv*, abs/1810.01943.
- Bertot JC, Jaeger PT and Hansen D** (2011) The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government Information Quarterly* 29, 30–40.
- Bonaccorso G** (2017) *Machine Learning Algorithms: A Reference Guide to Popular Algorithms for Data Science and Machine Learning*. Packt Publishing.
- Charalabidis Y, Maragoudakis M and Loukis E** (2015) Opinion mining and sentiment analysis in policy formulation initiatives: The EU-Community approach. In Tambouris E, Panagiotopoulos P, Sæbø Ø, Tarabanis K, Wimmer MA, Milano M and Pardo T (eds.), *Electronic Participation. ePart 2015*, Lecture Notes in Computer Science, vol. 9249. Cham: Springer. https://doi.org/10.1007/978-3-319-22500-5_12
- Chauhan P and Sood M** (2021) Big data: Present and future. *Computer* 54(4), 59–65. <https://doi.org/10.1109/MC.2021.3057442>
- Christou IT** (2019) Avoiding the hay for the needle in the stack: Online rule pruning in rare events detection. In *IEEE International Symposium on Wireless Communication Systems (ISWCS), Special Session on IIoT*. Oulu, Finland: IEEE, pp. 661–665.
- Christou IT, Kefalakis N, Soldatos J and Despotopoulou A-M** (2022) End-to-end industrial IoT platform for quality 4.0 applications. *Computers in Industry* 137, 103591.
- Christou IT, Kefalakis N, Zalonis A and Soldatos J** (2020) Predictive and explainable machine learning for industrial internet of things applications. In *IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*. Marina del Rey, CA: IEEE, pp. 213–218. <https://doi.org/10.1109/DCOSS49796.2020.00043>
- Curry E, Metzger A, Berre A, Monzón A and Boggio-Marzet A** (2021) A reference model for big data technologies. In Curry E, Metzger A, Zillner S, Pazzaglia JC and García Robles A (eds.), *The Elements of Big Data Value*. Cham: Springer. https://doi.org/10.1007/978-3-030-68176-0_6.
- Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, Ranzato M, Senior A, Tucker P, Yang K, Le Q and Ng AY** (2012) Large scale distributed deep networks. In *NIPS'12*. New York: Curran Associates.
- Deng C, Ji X, Rainey C, Zhang J and Lu W** (2020) Integrating machine learning with human knowledge. *iScience* 23(11), 101656. <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 WD, Schatsky D and Viechnicki P** (2017) AI-augmented government. Using cognitive technologies to redesign public sector work. Retrieved 7 July 2021, from <https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/artificial-intelligencegovernment.html>.
- European Commission** (2021) Document 52021PC0206, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, COM/2021/206 final, April 2021. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>.
- Fryer DV, Strömke I and Nguyen H** (2021) Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access* 9, 144352–144360.
- Gesk TS and Leyer M** (2022) Artificial intelligence in public services: When and why citizens accept its usage. *Government Information Quarterly* 39, 101704. <https://doi.org/10.1016/j.giq.2022.101704>
- Gonzalez J, Low Y, Gu H, Bickson D and Guestrin C** (2012) PowerGraph: Distributed graph-parallel computation on natural graphs. In *OSDI'12*. Berkeley, CA: USENIX Association, pp. 17–30.
- Hao K** (2019) This is how AI bias really happens—And why it's so hard to fix. MIT Technology Review, February 4, 2019. Available at <https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happens-and-why-its-so-hard-to-fix/>.
- Hocht J, Parycek P and Schollhammer R** (2016) Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce* 26(2016), 147–169.
- Huamani EL, Mantari A and Roman-Gonzalez A** (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(4). <https://doi.org/10.14569/IJACSA.2020.0110474>

- ISO/IEC 23894:2023 Information Technology** (2023) Artificial intelligence — Guidance on risk management, Technical Committee: ISO/IEC JTC 1/SC 42 Artificial intelligence, February 2023. Available at: <https://www.iso.org/standard/77304.html>.
- Jouppi NP, Young C, Patil N, Patterson D, Agrawal G, Bajwa R, Bates S, Bhatia S, Boden N, Borchers A, Boyle R, Cantin P-I, Chao C, Clark C, Coriell J, Daley M, Dau M, Dean J, Gelb B, Ghaemmaghami TV, Gottipati R, Gulland W, Hagmann R, Ho CR, Hogberg D, Hu J, Hundt R, Hurt D, Ibarz J, Jaffey A, Jaworski A, Kaplan A, Khaitan H, Killebrew D, Koch A, Kumar N, Lacy S, Laudon J, Law J, Le D, Leary C, Liu Z, Lucke K, Lundin A, MacKean G, Maggiore A, Mahony M, Miller K, Nagarajan R, Narayanaswami R, Ni R, Nix K, Norrie T, Omernick M, Penukonda N, Phelps A, Ross J, Ross M, Salek A, Samadiani E, Severn C, Sizikov G, Snellham M, Souter J, Steinberg D, Swing A, Tan M, Thorson G, Tian B, Toma H, Tuttle E, Vasudevan V, Walter R, Wang W, Wilcox E and Yoon DH** (2017) In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA '17)*. New York, NY, USA: ACM, pp. 1–12. <https://doi.org/10.1145/3079856.3080246>
- Kingma DP and Ba J** (2014) Adam: A method for stochastic optimization. Preprint, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- Leyer M and Schneider S** (2021) Decision augmentation and automation with artificial intelligence: Threat or opportunity for managers? *Business Horizons* 64(5), 711–724.
- Lindgren I, Madsen CØ, Hofmann S and Melin U** (2019) Close encounters of the digital kind: A research agenda for the digitalization of public services. *Government Information Quarterly* 36(3), 427–436.
- López CP, Rodríguez MD and de Lucas Santos S** (2019) Tax fraud detection through neural networks: An application using a sample of personal income taxpayers. *Future Internet* 11(4), 86. <https://doi.org/10.3390/fi11040086>
- Marban O, Mariscal G and Segovia J** (2009) A data mining & knowledge discovery process model. In Ponce J and Karahoca A (eds.), *Data Mining and Knowledge Discovery in Real Life Applications*. IntechOpen. <https://doi.org/10.5772/643>.
- Monogan III JE** (2015) *Political Analysis Using R*. Cham: Springer. <https://doi.org/10.1007/978-3-319-23446-5>
- National Institute of Standards and Technology (NIST)** (2021) Special Publication 800-37, Revision 2, Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy, April 2021. Available at <https://www.nist.gov/privacy-framework/nist-sp-800-37>.
- Papadakis T, Christou IT, Ipektsidis C, Soldatos J and Amicone A** (2022, November 1) AI Solutions for Transparent, Explainable and Regulatory Compliant Public Policy Development. <https://doi.org/10.5281/zenodo.7272425>
- Qian T and Medaglia R** (2019) Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly* 36(2), 368–383. <https://doi.org/10.1016/j.giq.2018.09.008>.
- Rahman MM, Arshi AS, Hasan MM, Mishu SF, Shahrir H and Wu F** (2023) Security risk and attacks in AI: A survey of security and privacy. In *IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, 2023*. IEEE, pp. 1834–1839. <https://doi.org/10.1109/COMPSAC57700.2023.00284>
- Ribeiro MT, Singh S and Guestrin C** (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)*. New York, NY, USA: Association for Computing Machinery, pp. 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Rosemann M, Becker J and Chasin F** (2020) City 5.0. *Business & Information Systems Engineering* 63, 1–7.
- Rudin C** (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 206–215.
- Shrikumar A, Greenside P and Kundaje A** (2017) Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML '17)*. JMLR.org, Volume 70, pp. 3145–3153.
- Sibson R** (1973) SLINK: An optimally efficient algorithm for the single-link cluster method. *Computer Journal. British Computer Society* 16(1), 30–34.
- Soldatos J and Kyriazis D** (eds.) (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://doi.org/10.1561/9781680838770>
- Srinivasan R and Chander A** (2021) Biases in AI systems. *Communications of the ACM* 64(8), 44–49. <https://doi.org/10.1145/3464903>
- Suresh H and Guttag J** (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*. New York, NY, USA: ACM, 9 Pages. <https://doi.org/10.1145/3465416.3483305>
- Suresh H and Guttag JV** (2019) A framework for understanding unintended consequences of machine learning. Preprint, [arXiv:1901.10002](https://arxiv.org/abs/1901.10002).
- Westhoven M** (2022) Requirements for AI support in occupational safety risk analysis. In *Proceedings of Mensch Und Computer 2022 (MuC'22)*. New York, NY, USA: Association for Computing Machinery, pp. 561–565. <https://doi.org/10.1145/3543758.3547576>