AI4Gov: Trusted AI for Transparent Public Governance Fostering Democratic Values

George Manias[1], Dimitris Apostolopoulos[2], Sotiris Athanassopoulos[3], Spiros Borotis[3], Charalampos Chatzimallis[4], Theodoros Chatzipantelis[5], Marcelo Corrales Compagnucci[6], Tanja Zdolsek Draksler[7], Fabiana Fournier[8], Magdalena Goralczyk[6], Alenka Gucek[7], Andreas Karabetian[1], Stavroula Kefala[9], Dimitris Kotios[1], Matej Kovacic[7], Danai Kyrkou[4], Lior Limonad[8], Sofia Magopoulou[4], Konstantinos Mavrogiorgos[1], Vasiliki Moumtzi[4], Septimiu Nechifor[10], Dimitris Ntalaperas[11], Georgia Panagiotidou[5], Martha Papadopoulou[4], Xanthi S. Papageorgiou[11], Nikos Papageorgopoulos[11], Dusan Pavlovic[6], Elena Politi[11], Vicky Stroumpou[9], Apostolos Vontas[4], Dimosthenis Kyriazis[1]

[1]*University of Piraeus,* Piraeus, Greece
[2]*Municipality of Vari, Voula, Vouliagmeni,* Greece
[3]*Maggioli S.p.A.,* Santarcangelo di Romagna, Italy
[4]*VILABS Limited,* Limassol, Cyprus
[5]*Aristotle University of Thessaloniki,* Thessaloniki, Greece
[6]*White Label Consultancy ApS,* Copenhagen, Denmark
[7]*Jožef Stefan Institute,* Ljubljana, Slovenia
[8]*IBM Israel Science & Technology Ltd,* Haifa, Israel
[9]*Greek Ministry of Tourism,* Athens, Greece
[10]*SIEMENS S.R.L.,* Bucharest, Romania
[11]*Ubitech Limited,* Limassol, Cyprus

gmanias@unipi.gr, dapostolopoulos@vvv.gov.gr, sotiris.athanassopoulos@maggioli.gr, spiros.borotis@maggioli.gr, chatzimallis@vilabs.eu, chadji@polsci.auth.gr, mc@whitelabelconsultancy.com, tanja.zdolsek@ijs.si, fabiana@il.ibm.com, mg@whitelabelconsultancy.com, alenka.gucek@ijs.si, adreaskar@unipi.gr, kefala_s@mintour.gr, dimkotios@unipi.gr, matej.kovacic@ijs.si, danaikyrkou@vilabs.eu, liorli@il.ibm.com, sofia.magopoulou@vilabs.eu, komav@unipi.gr, movaki@vilabs.eu, septimiu.nechifor@siemens.com, dntalaperas@ubitech.eu, gvpanag@office365.auth.gr, papadopoulou@vilabs.eu, xpapageorgiou@ubitech.eu, npapageorgopoulos@ubitech.eu, dpv@whitelabelconsultancy.com, epoliti@ubitech.eu, stroumbou_v@mintour.gr, avontas@vilabs.eu, dimos@unipi.gr

***Abstract*—As Artificial Intelligence (AI) becomes more integrated into public governance, concerns about its transparency and accountability have become increasingly important. The use of AI in decision-making processes raises questions about bias, fairness, and the protection of individual fundamental rights. To ensure that AI is used in a way that upholds democratic values, it is essential to develop systems that are trustworthy, transparent, and accountable. Trusted AI allows citizens to have greater trust in public organizations and their decision-making processes, while it also enables public authorities and policy makers to be more transparent and accountable, providing citizens with greater visibility into how policies are developed. In addition, it encourages the use of AI in a way that promotes fairness and equity, ensuring that decision-making processes are unbiased and discrimination free against certain groups of individuals. This paper investigates how these desirable attributes can be developed in ways that are feasible and effective through the design of a holistic environment that incorporates AI and Big Data management mechanisms while preserving that the AI technology should be shaped around human rights, values, and societal needs. Societal change and evidence-based policies will be achieved through the extension of business and policy making processes with advanced approaches, such as eXplainable AI (XAI) and Situation-Aware Explainability (SAX). To this end, a novel approach is proposed, which will converge techniques and research on multiple domains, including social sciences, Trustworthy AI, Ethical AI, Big Data analytics, IoT, and blockchain into a unified ecosystem.**

Keywords—Trustworthy AI, Ethical AI, Bias and Discrimination, Distributed ledger.

I. Introduction

Due to the large and growing populations worldwide, many national, regional, and local authorities and organizations currently face a huge strain on resources, infrastructure, and transportation whilst simultaneously battling pollution and environmental hazards [1]. The use of emerging technologies such as AI and Big Data are being pioneered to help these organizations meet individuals' needs and to provide a sustainable future for their citizens [2]. Recent advances in Big Data and Artificial Intelligence (AI) coupled with the pervasiveness of devices and sensors of the Internet of Things (IoT) can change in a multidimensional level the materiality of modern societies. These solutions provide the possibility to stakeholders to deploy new tools and models [3], and produce evidence-based policy making, while protecting fundamental rights and values from possible negative and multifaceted effects on individuals and democratic societies [4]. The utilization of AI and Big Data is incongruent with fundamental democratic principles and human rights and if these technologies are governed incorrectly, then many challenges are posed for the citizens' rights and values [5].

In terms of data-driven policy making, AI can deliver effective and efficient services, which are high-quality and low-cost at the same time, creating public value [6]. In addition, AI should consider other criteria of good governance in the public policy system, such as the possibility of democratic input, legality, integrity, equality before the law, and accountability [7]. AI systems could exaggerate digital surveillance and data manipulation, reflect, and reinforce some of the deepest societal inequalities, fundamentally alter the delivery of public and essential services, undermine data

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

protection legislation, and disrupt policy making processes [8]. The latter is of high importance, as the algorithmic structure of platforms related to public services and administration, increasingly impacts and shapes political messaging, information-seeking, news distribution, and citizen engagement [9]. To this end, fundamental elements of human rights, as well as the respect for human dignity, freedom, equality, democracy, and the rule of law must be safeguarded, and citizens should be protected from the negative impacts of AI and Big Data. This way they will be able to realize the opportunities presented by the utilization of these technologies [10]. These challenges are considered major pillars towards the successful implementation of the EU's vision for fostering excellence in AI and strengthening the uptake, investment, and innovation in AI [11].

Likewise, the tremendous technological change of modern societies has evolved citizen behaviors and expectations for more responsive public services with seamless user experience. Citizens' behavior towards new ways of communication through digital media and platforms arises the question of whether or not technology companies "strategically" deploy algorithms to reinforce their position of power in AI and to maximize profit and engagement [12]. Hence, digital government services and policies must be easy-to-use, secure, transparent, trustworthy, unbiased, and always available to further promote and encourage modern citizens to be active stakeholders that co-create and participate in the policy making procedures [13]. To this end, the development and utilization of holistic, trusted, and explainable algorithms from the domains of AI, Machine Learning (ML), Automated Machine Learning (AutoML), and Federated Learning (FL) will enhance civic engagement, political and social representation, inclusive participation and pluralism, hence fundamental characteristics of democracy. On top of this, governments can leverage the power of AI and Big Data to innovate and transform the public sector to redefine the ways in which they design and implement policies and services.

However, the implementation of AI and Big Data based solutions for data-driven and evidence-based policy making purposes is associated with various challenges, which have not been yet adequately addressed and involve not only how these technologies are being developed, but also their interaction with people and organizations, giving rise to leadership, policy, and administration challenges [14]. Thus, modern AI approaches should be developed in the context of the moral values of individuals and democratic societies rather than viewed as technical, "black-boxed" and value-neutral tasks of developing components and mechanisms that meet functional requirements formulated by clients and users [15]. Through a trusted, and fair development, deployment, and utilization of AI and Big Data, many opportunities will rise to digitize public administration, automate public policy workflows, strengthen regulatory frameworks, and enhance civic engagement and participation.

In this context, this paper presents the main research challenges and architecture of an overall integrated environment and digital platform for addressing the challenge of providing fair, unbiased, trusted, and explainable AI and Big Data mechanisms. The proposed environment (namely AI4Gov) and its integrated frameworks will evaluate and provide indicators across the five important aspects of AI - bias, fairness, transparency, responsibility, and interpretability - to evaluate data, algorithms, and ethical outcomes of AI within the proposed digital platform. To this end, it aims to contribute to the research, societal, and technological landscapes by addressing ethical, trust, discrimination, and bias issues. The latter will be achieved by providing an in-depth analysis and solutions addressing the challenges faced by various stakeholders in modern societies when attempting to mitigate Big Data and AI challenges. In this direction, the project will introduce solutions and frameworks in a two-fold sense: to facilitate policy makers on the development of automated, educated, and evidence-based decisions and to increase the trust of citizens in the policy making processes.

The remainder of the paper is structured as follows. Section 2 describes State-of-the-Art initiatives and technologies that aim to address the aspects of trust, interpretability, explainability, and fairness in modern AI systems. In Section 3 the proposed digital solution and platform is introduced, as well as the initial architecture and the different integrated frameworks that will be designed and utilized for addressing the challenges posed by the use of AI and Big Data in the policy making processes. Moreover, Section 4 presents the piloting activities that will validate and evaluate the utilization of the proposed environment, while Section 5 states the expected wider impacts and how they will be realized. Finally, Section 6 concludes the paper and states the future work that will be implemented.

II. State-of-the-Art Initiatives and Expected Solutions

The great expansion in the utilization of AI and Big Data technologies has a two-fold interpretation. While they foster socio-economic benefits and provide key competitive advantages to private and public organizations, at the same time they raise the risks or negative consequences for individuals or society, as the power of AI can be harnessed for surveillance and manipulation [16]. The combined power of AI and Big Data can restrict users' options, influence their opinions, and manipulate them into making choices that do not serve their best interests. Due to the proliferation of AI in modern societies and its significant socio-economic impact, it becomes necessary to investigate AI-based systems for unwanted biases and discrimination and develop methods to mitigate and monitor these negative aspects [17]. These systems have to be secure and resilient against malicious attempts to manipulate AI-based policy making activities. For instance, adversaries may launch cybersecurity attacks against deep neural networks for policy making, compromising their ability to classify situations and to propose fair and meaningful policies [18]. The European Commission expects that AI can significantly improve the lives of EU citizens and bring major benefits to society and economy through better healthcare, more efficient public policies, safer transport, more competitive market and sustainable farming [19]. To this end, it has launched policy initiatives to progress in this area, including the "Communication Artificial Intelligence for Europe" [20], the "Declaration of Cooperation on AI" [21], and the "Coordinated action plan on the development of AI in the EU" [22], among others. On top of this, enhancing the transparency, effectiveness, accountability, and legitimacy of public policy making is a challenge that falls in the realm of the Ethical AI challenges, which have been recently studied and analyzed by EU's High-Level Expert Group (HLEG) on AI [23].

The utilization of AI and Big Data algorithms, especially as concerns the policy making processes, facilitate the

extraction of hidden insights and knowledge that humans can hardly produce due to their inability to process very large volumes of data [24]. At the same time, awareness of the potential issues is increasing at a fast rate, but the AI community's ability to take action to mitigate the associated risks and challenges both to core individual values as well as to European collective values is still in its infancy [25]. Gaps between the design and operation of algorithms and the understanding of their ethical implications can have severe consequences affecting individuals as well as groups and whole democratic societies [26]. The implementation and utilization of AI and Big Data technologies should be citizen-centric and human-oriented, thus the gap between civil society and technical experts should be narrowed. An extensive review of 84 ethical AI documents concluded that no single ethical principle featured in all of them [27]. In this context, values such as transparency, justice, fairness, non-maleficence, responsibility, and privacy have been identified and highlighted as core public values that receive pressure from modern digitization [28]. Within the last three years, several documents on AI Ethical guidelines have been published by a multiplicity of stakeholders stemming from the industry domain (e.g., Google, IBM) [29], governmental domain (e.g., the High-Level Expert Group of the European Commission) [30], intergovernmental institutions (e.g., OECD) [31], and academia domain (e.g., IEEE) [32]. These documents seek to provide regulations, laws, non-regulatory measures, and recommendations on how to tackle and minimize the negative impacts that these technologies pose on the fundamental rights and values of citizens in modern democratic societies.

What is more, the use of AI and Big Data in policy making procedures is hindered by the lack of public organizations and citizens' trust in the operation of algorithms [33]. This is for example the case with most Deep Learning (DL) algorithms, which operate as "black-boxes" and cannot be understood by end-users and domain experts [34]. The deployment and use of such algorithms raise trustworthiness issues and hinder the use of AI and ML in use cases like credit risk scoring, loan approval and personalized asset management recommendations. Furthermore, AI algorithms for public policies are commonly associated with different types of biases, such as social, ethnic and gender related biases [35]. Hence, explainability of AI becomes an essential ingredient as part of its overall trustworthiness, catering to ongoing assurance of operational adequacy and transparency that could easily be interpreted by the users. To alleviate transparency issues, public organizations can employ XAI techniques [36], such as LIME and SHAP, which aim to highlight decision-relevant features in the AI model employed, that either contribute to the model accuracy on the training set, or to a specific prediction for one particular observation [37]. This caters to a variety of post-hoc methods for feature dependence quantification such as backpropagation and perturbation-based techniques for different types of Neural Networks [38], or model agnostics techniques such as SHAP [39]. XAI can be also combined with Explorative Data Analysis (EDA) to detect and mitigate different types of biases, such as biases associated with the use of historic or non-representative datasets [40]. Moreover, Feature Extraction can also provide insights on explainability of AI, by identifying the features that predict AI outcomes [41]. Another class of techniques leverages game theory to interpret the predictions of ML/DL models [42]. Most importantly, the advent of XAI techniques can greatly boost the transparency and interpretability of AI when used to propose public policies [43]. However, while explainability is often agnostic to the algorithmic model employed, the topology of an AI model itself helps depicting the rationale for its results. Such a property is typically referred to as the model's interpretability (a.k.a., ante-hoc methods), reflecting the level at which a given model makes sense for a human observer [44] and also reveals how inputs are mathematically mapped to outputs [45]. While the use of AI in Information and Communications Technologies (ICT) is a major concern as aforementioned, the use of AI within the broader situational context of operational business processes

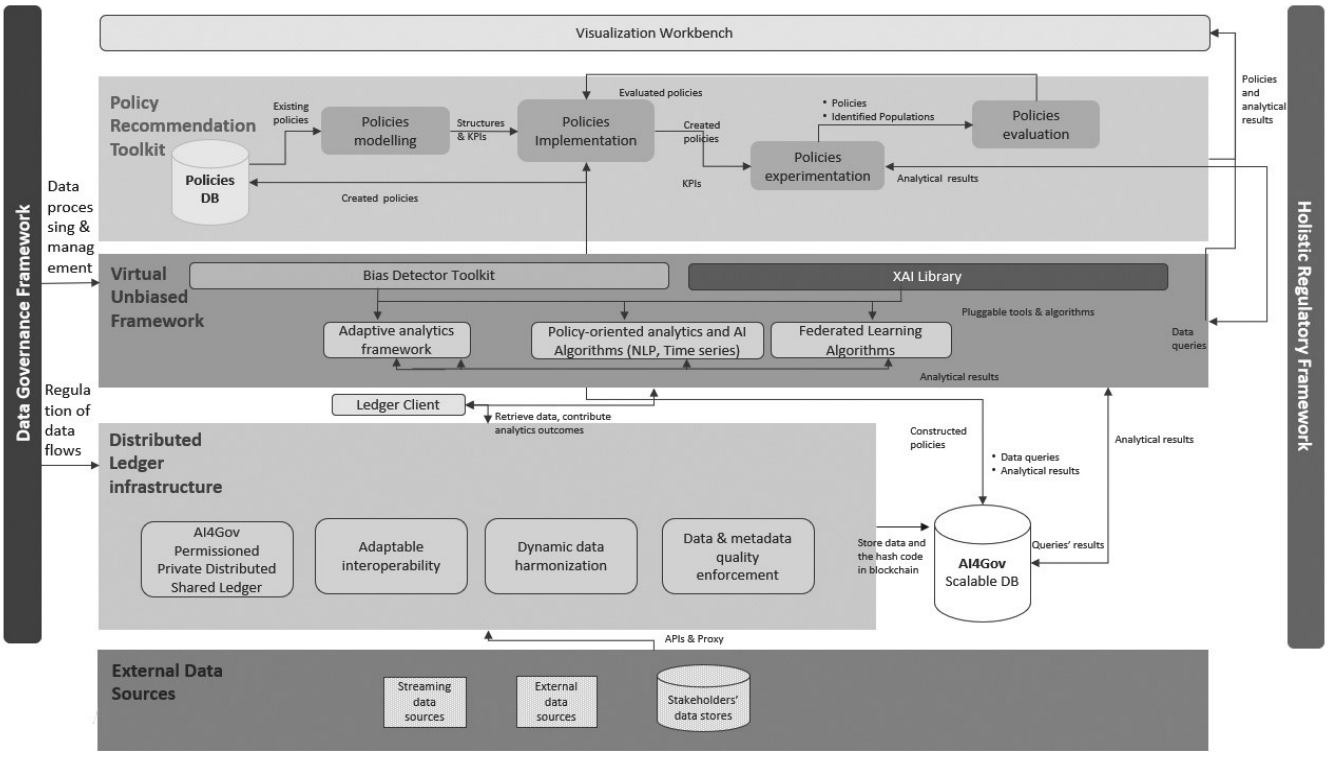

Fig. 1. AI4Gov Architecture

amplifies this concern even further. Recent work coins such AI-enriched business processes as Augmented Business Process Management Systems (ABPMSs) [46]. The "reasoning" steps in such processes could be realized by a variety of AI models and thus the Explainability of process execution results should also entangle to the broader context in which these models are embedded in the process. However, explainability is strictly decoupled with only a local input view that is associated with the employment of each individual AI model employed along the overall (global) reasoning process and thus, is lacking process-awareness [47]. To this end, Situation-Aware Explainability (SAX) techniques have been identified as the next evolution of explainable techniques for business organizational process systems that utilize AI (namely as ABPMSs) [48].

AI4Gov will introduce various tools and frameworks to enable policy makers and stakeholders to understand, detect, and mitigate biases and discrimination in AI and Big Data. The development and utilization of these tools will be incorporated into a holistic digital solution that seeks to leverage and enhance fair and trusted AI towards: (i) raising awareness to developers and users about trust and bias; (ii) indicating intersectional groups most likely to be affected by bias and unfairness in certain use cases; (iii) informing developers and users how to mitigate challenges posed by the utilization of AI and Big Data; (iv) establishing a reporting framework to the public, which standardizes the communication of cases of biases, discrimination, and unfairness.

III. Proposed Digital Solution

The sustainable development, deployment, and utilization of AI and Big Data in the public sector, thus in policy making processes related to the public policy circle, administration and governance, require dialogue and deliberation between developers, policy makers, deployers, end users, and the citizens. In this context, the AI4Gov ecosystem is a joint effort of policy makers, public institutions / organizations, legal, Social Science and Humanities, and Big Data/AI experts and research centres to unveil the potential of these technologies for developing evidence-based innovations, policies, and policy recommendations to harness the public sphere, political power, and economic power of modern organizations and authorities. The AI4Gov ecosystem aims to introduce a holistic approach and different regulatory, data governmental, and technological frameworks that will include (i) a thorough analysis and understanding of bias and ethics, (ii) methods to mitigate ethics issues on several levels of the AI chain, and (iii) a proactive accounting for bias including awareness, description of bias and finally an explanation of the AI based decision. Overall, AI4Gov will develop, validate, and make available within its platform five main frameworks, as also presented in Fig. 1. while specific interfaces, tools and cross-subsystem elements address technical, business and operational requirements associated with the platform objectives:

A. Holistic Regulatory Framework (HRF)

The HRF will be developed on top of the EU AI Act [49] and will be based on a qualitative analysis of fundamental rights, EU values and examination of legal activities and ethical protocols to ensure that the proposed framework protects citizens from potential abuse enabled by the use of Big Data and AI. The HRF will be in-line with applicable laws, protocols, and regulations (i.e., the GDPR), but also with ethical recommendations for AI (e.g., the recommendations of the HLEG). This framework will integrate into different architectural blueprints acquiring/ensuring a holistic view on intersectional bias and ethics. A comprehensive analysis of multiple types of bias based on different grounds such as age, disability, gender, race, sexual orientation, and gender identity will establish the grounds for the realisation of the AI4Gov's outcomes providing support for "regulatory compliance by design". It seeks to analyze, predict, quantify, and monitor transparency, accountability and trustworthiness of AI and Big Data, considering individual rights to ensure meaningful redress for people affected by these technologies. The outcomes will lead to the creation of use-centric guidelines and training materials to create a cohesive, future-proof approach to AI systems utilization in policy making by adding mechanisms to update and monitor all risk categories for multiple sectors and vulnerable population groups.

B. Data Governance Framework (DGF)

As digital technologies such as Big Data and AI power an ever-expanding portion of modern societies, data governance is becoming increasingly critical. Thus, the implementation of the DGF is central to the success of digital transformation and data-driven and evidence-based policy making. Regulations and policies governing the use of data throughout the whole lifecycle of the data path will be identified and applied. High-quality data, leveraged as a key element of digital transformation, analytics, and insights, have become key assets in the overall policy making procedures. Hence, the framework and its accompanying tools will offer protection and privacy enforcement for the data and will ensure that decisions across the complete path follow specific protocols, regulations, and legislations and are in line with the HRF.

C. Virtualized Unbiased Framework (VUF)

The Virtualized Unbiasing Framework (VUF) for Big Data and AI will leverage notably AI-based analytics techniques such as ML/FL/RL algorithms for extracting policies from datasets, opinion mining, sentiment analysis over documents, chatbots for citizen's and policy makers interaction, etc. All these analytics will be deployed as reusable and configurable services that can be invoked dynamically and thus be reused and repurposed for different datasets and divergent data sources. The development of advanced AI technologies will monitor and boost the ethical nature of AI-based policies. This framework will provide (i) a thorough analysis and understanding of bias, (ii) methods to mitigate bias and discrimination on several levels of the AI chain, and (iii) a proactive accounting of the ethical and legal aspects of AI and Big Data including awareness, description of existing and newly introduced regulations and laws, and finally explanation and interpretation of the AI based policy making.

1) A Bias Detector Toolkit will be implemented and utilized to spot any patterns of bias and discrimination based on features such as gender, age, ethnicity, low income, and disability among others both on the provided datasets, as well on the designed and implemented AI tools. The toolkit will introduce sets of fairness metrics for datasets and models, explanations for these metrics, and algorithms to mitigate bias in datasets and models. The utilization of such toolkits emerges during the last years and they are considered core modules in the implementation of modern AI models [50-51].

2) Advanced Policy-Oriented Natural Language Processing (NLP) techniques, such as Multilingual Sentiment Analysis [52], Topic Modelling [53] and Question Answering [54] have helped policy makers effortlessly process citizen input and extract the key ideas, the main topics of interest and the sentiments expressed. In the context of the proposed digital solution such techniques will be examined and implemented to analyse large volumes of unstructured text data. The latter will allow stakeholders to rapidly examine the information and gain meaningful insights from citizens' feedback and opinions. NLP techniques could help public and private organizations make better-informed policies whilst reducing the level of administrative effort needed in the policy making process.

3) A Self-Explained Visualisation Workbench to enable a timely and holistic understanding of the data and policies, while improving usability through the utilization of advanced Visual Analytics techniques. In conjunction with popular AutoML environments, this tool will enable stakeholders without data science knowledge to use the analytical tools and the policy recommendations. Visualization techniques will be classified by the underlying data structure: Linear, hierarchical, networked, vector spaced and graphs data, etc. Especially when data and policies increase in complexity, as is the case with high-dimensional datasets, traditional business charts are no longer able to convey all information in one chart. AI4Gov will leverage newer forms of visualizations suitable for Big Data and AI applications [56] – [57]. Using the AI4Gov visualizations workbench, the user will be able to view entire datasets within one comprehensive visualization. This will facilitate the generation of new insights that would otherwise stay uncovered.

4) Policy Recommendations Toolkit to enable public authorities and other policy makers to reuse policy models and datasets in their policy development tasks. To enable such reuse, techniques for the semantic interoperability of different policy models and datasets will be examined and implemented [55], notably techniques that leverage common ontologies and archetypes, while also realizing the AI models and algorithms as services that can be invoked dynamically and thus be reused and repurposed for different datasets. Likewise, the AI4Gov will integrate this range of AI tools within this single toolkit.

D. Explainable AI (XAI) Library

The AI4Gov seeks to enable policy makers and citizens to understand the decisions of AI based systems and tools, boosting the transparency and acceptance of the respective decisions. Specific strategies and solutions will be applied to address the reliability of data and interpretability of "black box" algorithms, whose inner workflows might seem convoluted, and to monitor the output of these algorithms. To this end, AI4Gov will develop and provide a novel XAI Library not only for explaining and understanding the underlying AI mechanisms to different stakeholders, but also for fostering bias and discrimination detection. Such solutions will enable the policy makers to take automated, educated and evidence-based decisions and to increase the trust of citizens in the policy making processes. Specifically, this library will provide:

1) XAI and Casual ML tools and techniques towards increasing the transparency, trustworthiness, and robustness of the AI algorithms and tools of the platform. To this end, AI4Gov will provide novel quantitative XAI tools that can balance explainability vs. performance trade-offs. These tools will enable policy makers to develop trusted and transparent policies, while helping them to understand and interpret the outcomes of AI-based recommendations. Specifically, popular XAI techniques like SHAP and LIME will be integrated, along with algorithms/techniques that decompose deep neural networks to the features that determine the AI decisions (e.g., DeepLIFT [58] and Prediction Difference Analysis techniques [59]).

2) Situation-Aware Explainability (SAX) techniques to help current organizational policies be extended with better instrumentation to establish the rationale behind situations that cannot currently be explained. Particularly, the extension of business processes with SAX gives organizations the ability to independently and continuously reason about process enactment outcomes (a "second-tier" of reasoning), in many cases in retrospect. This includes ongoing capturing of key conditions (e.g., historical framing that reflects timely assumptions and beliefs) and the ability to draw inferential associations (dependencies) between such conditions and intermediary process execution results (i.e., reasoning and/or enactment). Such drawing helps to autonomously establish/quantify the situational validity of any derived process output.

3) XAI Library as a means of explaining the rationale behind policy recommendations and the data that will be driving their production. The XAI library will be used to boost the transparency of AI systems while driving AI4Gov through techniques that identify the most distinctive and informative features that are used by Big Data and AI systems. This library will offer the developed XAI and SAX models as pluggable tools and services. The latter will facilitate the tracking of execution consistency, for a better understanding of policy flows and insights, and will drive ongoing process and policy making improvements (at either design – or retraction at run-time).

E. Blockchain-based Information Exchange (BIE) Framework

A novel Blockchain-based Information Exchange (BIE) framework for decentralized, scalable, automated, transparent, and interoperable data and policy management based on unique AI technologies. Given the utilization of blockchain and distributed ledger technologies, the platform will be by default transparent, but also portable and extensible given that new entities will be able to provide and obtain information through the blockchain. This framework will also regulate the access to the data by the various participants and facilitate the secure & trustful exchange of data across all stakeholders of modern democratic societies. Utilizing the blockchain, the platform will be able to access the data of the distributed ledger and following the anonymization and de-anonymization techniques will get the data off the blockchain for research and organizational purposes.

One of the main concepts of AI4Gov's BIE building block is the use of smart contracts to manage the interactions between all actors [60]. Based on this concept, every

contribution and activity from different actors is subject to well-defined business and security rules, which govern the participation of these entities. These entities could also be entire systems, and as such, AI4Gov can be extended to involve any system/device/platform with the required capabilities and willingness to cooperate, as well as any human actor with the required authorization level. Evidence-based policies can float across physical systems, so that work initiated at some location may be completed somewhere else (from another actor), with the active participation of the corresponding people/entities. This is made feasible by the decentralized nature of the blockchain infrastructure. Hence, the platform shifts from centralized, application-centric flow control to fully decentralized collaboration-control logic, which facilitates open and trustful data sharing across the participants of the AI4Gov network. While this is a primary benefit of the decentralized, blockchain approach, there are also scalability and reliability benefits. A key concept in AI4Gov is the possibility to include subsequent data fields and new datasets. However, in order to facilitate the latter on the blockchain level, an innovation that will be delivered refers to evolvable smart contracts. As such, the contracts will include the updated data structures (that will reflect the new data fields) and a fork will be triggered in the blockchain in order to ensure that new contributions are according to these new data types. What is more, public authorities and associations will undertake the role to ensure consensus among all entities in order to follow the forked blockchain for the iterations and contributions that follow the new data fields and as a result the new smart contracts. AI4Gov will offer an integrated environment that will support the creation, management, and validation of smart contracts, while integrating tools and techniques for managing data, data analytics, and compliance in regulatory frameworks (e.g., GDPR, EU AI Act).

IV. Piloting Methodology

The AI4Gov platform connects different stakeholders in the public and governmental sectors, providing them with a variety of tools. . In that direction large-scale piloting activities will be implemented and further validate and evaluate the utilization of the proposed AI4Gov digital solution for policy development in real-life use cases. The pilot use cases will showcase how AI4Gov will highlight, detect, and mitigate bias, discrimination, and exclusion of citizens on social accountability systems. The aim is to increase citizens' exposure to information and render public organizations more open and responsive to citizen feedback. The pilot use cases will range from the citizen centric and multi-domain policy management to sustainability and green policy making. These use cases will demonstrate the capabilities of the proposed digital solution towards increasing citizens' trust in the provided tools. For instance, Bias Detector Toolkit will be used to spot any patterns of bias and discrimination based on features such as gender, age, ethnicity, low income, and disability among others, especially during the inclusion and communication strategy, as well as during the policy making and reward processes. Likewise, XAI Library and its incorporated XAI models will be used to ensure that multi-domain policy adaptation through the relevant recommendations is accompanied by the corresponding required explanations and to increase the transparency of recommendations and policies when they are provided to citizens. SAX models aim to provide broader context information behind processes and to foster citizens, businesses, public authorities, governments and NGOs' participation towards independent and continuous reason about process enactment outcomes. NLP subtasks, such as Sentiment Analysis and Question Answering, will be utilized to assess and analyze citizens' feedback, as well as to investigate social backdrops and challenges that prevent citizen engagement, thus leading to exclusion, such as poverty, discrimination, violence, and civic habitus ("critical citizens"). Finally, the Policy Recommendation Toolkit and Interactive Self-Explained visualizations will be provided to policy makers and stakeholders to facilitate transparency and openness related to the policies put-in-place and the introduced sustainable solutions.

V. Expected Impacts

The AI4Gov ecosystem, as well as its tools and services will be offered to several stakeholders and target groups and more specifically to: (i) Public authorities/organizations for organizing, planning, and monitoring the timely provision of appropriate and enhanced policies; (ii) Legal authorities/organizations for undertaking relevant activities to mitigate the threats and risks of potential misuse of AI and Big Data to fundamental rights and values; (iii) Policy makers and regulators for using research evidence as they devise the regulatory frameworks that will shape the development and use of AI; (iv) Citizens, workers and additional vulnerable groups (e.g., communities of color, low income, migrants, LGBTQIA2S+ community), for receiving guidelines and training materials towards the increase of their awareness, education and participation for societal change, human bias and discrimination risks; and (v) Researchers and political scientists targeting on analyzing data for insights, leading to efficient decision-making based on FAIR data to facilitate their discovery, interoperability and use by each stakeholder. To this end, AI4Gov will substantially help public authorities, legal organizations, and other policy actors to factor legal and ethical hazards caused by potential misuse of AI and Big Data in their policies, services, and respective processes. The ethical and human rights challenges will be initially reviewed as well as mitigation strategies will be proposed to analyze and research how a regulatory framework might be designed and deployed to address these challenges.

In addition, through the design and implementation of the previously introduced digital solution several key challenges seek to be addressed and a wider impact will be introduced. To this end, AI4Gov seeks to introduce an AI-oriented, citizen-centric, transparent, trustworthy, FAIR-based, ethical, and legally compliant environment, with access to multidimensional impacts, ranging from societal to technical, and from ethical to scientific levels. More specifically, AI4Gov will boost the ability of policy makers and public authorities to exploit and monetize their data assets, while at the same time developing novel and enhanced services and policies. It will open opportunities stemming from access to and consumption of data in an open federated, decentralized, blockchain-based environment instead of the current centralized and siloed models and solutions. Hence, AI4Gov will be a great technological contribution to the European ICT ecosystem, as it will be complemented with a set of novel and validated business models, which will enable European public and private organizations to monetize their assets, increase their market share and generate new revenue streams based on the development of novel, trusted, personalized policies. In addition, the developed building blocks, frameworks, and tools will reassure the ability of the public authorities and

policy makers to influence the direction of AI innovation towards greater trustworthiness and positivity about the impact of sound and resilient policies, mitigating AI risks and enhancing AI's benefits.

As concerns the societal and ethical instances of the platform, through the utilization of the introduced technologies and tools the AI4Gov seeks to examine established legislation and non-regulatory measures over AI and Big Data development and implementation. It will also foster the a) understanding of the impacts of data misuse, including bias, surveillance, disinformation, and feedback loops; b) recognition of the contributing factors to these impacts; c) identification of different types of bias, discrimination, and unfairness on AI and Big Data systems; d) development of literacy in investigating how data and data-powered algorithms shape, constrain, and manipulate commercial, civic, and personal experiences; and e) analysis of new scenarios and potential products to try to identify and mitigate potential risks. At the same time, it will ensure the trustworthiness and reliability of its AI and Big Data systems by covering not only technical aspects of risk mitigation (such as statistical data analysis) but also social science aspects. In that direction, it will provide building blocks and tools that will enable the data owners to have control over their (personal) data, while the data exchange between the blockchain-based nodes will be secure and compliant with applicable directives and regulations. Likewise, it will offer tools for trustworthy and unbiased AI developments (e.g., XAI models, and Bias Detector Toolkit), which will be transparent and accepted by citizens, public authorities, and organizations. In this way, it will lower the trust barrier for citizens to share their data and to engage in the use of public and democratic services and in the policy making procedures. AI4Gov will demonstrate these important societal benefits in the context of AI tools that offer equal and unbiased access to best practices and policies.

VI. Conclusion

Cognizant of the ongoing debates surrounding the ethical, regulatory, and policy implications that emerge from the development and utilization of AI, AI4Gov will focus on the ways in which AI techniques, tools, and technologies are developing and how these developments may affect the lives of different groups of people at an individual and collective level. In this direction, the project will develop tools for regulatory compliance of AI models, along with a democratic AI label in the form of certification. To maximize societal acceptability and trust in evidence-based policy making, extensive and in-depth analyses of regulatory, technological, societal, and ethical aspects will be provided, by seeing to an optimal embedding of the results of these into the design of the platform and its different values-based frameworks. AI4Gov seeks to introduce a refined set of requirements, guidelines, tools, and norms for supporting policy making processes, aligned with the iterations of the development of the platform in the use cases. It will also leverage AI technologies to establish scalable mechanisms for analyzing and implementing fairness, bias detection, and AI trustworthiness techniques in the development, evaluation, and optimization of public policies. The proposed AI4Gov ecosystem and digital solution will be further evaluated with data obtained from multiple cohorts/regions in the EU to help validate data models, Trustworthy AI and XAI-based policy making techniques to facilitate knowledge exchange and collaboration at governmental, societal, research and policy making levels.

Acknowledgment

The research leading to the results presented in this paper has received funding from the European Union's funded Project AI4Gov under grant agreement no 101094905.

References

[1] R. Mark, "Ethics of public use of AI and big data: the case of Amsterdam's crowdedness project," The ORBIT Journal, vol. 2, no. 2, pp. 1-33, 2019.

[2] J. Berryhill, K. K. Heang, R. Clogher, and K. McBride, Hello, World: Artificial intelligence and its use in the public sector, 2019.

[3] B. W. Wirtz, J. C. Weyerer, and C. Geyer, "Artificial intelligence and the public sector—applications and challenges," International Journal of Public Administration, vol. 42, no. 7, pp. 596-615, 2019.

[4] J. Reis, P. E. Santo, and N. Melão, "Impacts of artificial intelligence on public administration: A systematic literature review," in 2019 14th Iberian conference on information systems and technologies (CISTI), pp. 1-7. IEEE, 2019.

[5] E. Christodoulou, and K. Iordanou, "Democracy under attack: challenges of addressing ethical issues of AI and big data for more democratic digital media and societies," Frontiers in Political Science, vol. 3, 682945, 2019.

[6] T. S. Gesk, and M. Leyer, "Artificial intelligence in public services: When and why citizens accept its usage," Government Information Quarterly, vol. 39, no. 3, e101704, 2022.

[7] J. Reis, P. E. Santo, and N. Melão, "Artificial intelligence in government services: A systematic literature review," New Knowledge in Information Systems and Technologies, vol. 1, pp. 241-252, 2019.

[8] M. M. Young, J. B. Bullock, and J. D. Lecy, "Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration," Perspectives on Public Management and Governance, vol. 2, no. 4, pp. 301-313, 2019.

[9] A. Unver, Artificial intelligence, authoritarianism and the future of political systems, EDAM Research Reports, 2018.

[10] M. Haenlein, and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," California management review, vol. 61, no. 4, pp. 5-14, 2019.

[11] G. Misuraca, and C. Van Noordt, AI Watch-Artificial Intelligence in public services: Overview of the use and impact of AI in public services in the EU, JRC Research Reports (JRC120399), 2020.

[12] E. N. Loukis, M. Maragoudakis, and N. Kyriakou, "Artificial intelligence-based public sector data analytics for economic crisis policymaking," Transforming Government: People, Process and Policy, vol. 14, no. 4, pp. 639-662, 2020.

[13] Z. Allam, and Z. A. Dhunny, "On big data, artificial intelligence and smart cities," Cities, vol. 89, pp. 80-91, 2019.

[14] P. Henman, "Improving public services using artificial intelligence: possibilities, pitfalls, governance," Asia Pacific Journal of Public Administration, vol. 42, no. 4, pp. 209-221, 2020.

[15] R. Clauberg, "Challenges of digitalization and artificial intelligence for modern economies, societies and management," RUDN Journal of Economics, vol. 28, no. 3, pp. 556-567, 2020.

[16] C. Feijóo, et al., "Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy," Telecommunications Policy, vol. 44, no. 6, e101988, 2020.

[17] A. Pena, I. Serna, A. Morales, and J. Fierrez, "Bias in multimodal AI: Testbed for fair automatic recruitment," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 28-29, 2020.

[18] Y. Zhang, et al, "The secret revealer: Generative model-inversion attacks against deep neural networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 253-261, 2020.

[19] U. J. Muehlematter, P. Daniore, and K. N. Vokinger, "Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis," The Lancet Digital Health, vol. 3, no. 3, e195-e203, 2021.

[20] "Communication Artificial Intelligence for Europe | Shaping Europe's digital future", European Commission, 2018. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe [Accessed: 28-Apr-2023].

[21] "EU Member States sign up to cooperate on Artificial Intelligence", European Commission, 2018. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence [Accessed: 02-May-2023].

[22] "Member States and Commission to work together to boost artificial intelligence „made in Europe"", European Commission, 2018. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/member-states-and-commission-work-together-boost-artificial-intelligence-made-euro [Accessed: 28-Apr-2023].

[23] "High-level expert group on artificial intelligence", European Commission, 2020. [Online]. Available: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence [Accessed: 27-Apr-2023].

[24] Y. K. Dwivedi, et al., "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," International Journal of Information Management, vol. 57, e101994, 2021.

[25] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices," in Ethics, Governance, and Policies in Artificial Intelligence, pp. 153-183. Springer, Cham, 2021.

[26] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," Big Data & Society, vol. 3, no. 2, 2053951716679679, 2016.

[27] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389-399, 2019.

[28] L. Royakkers, J. Timmer, L. Kool, and R. van Est, "Societal and ethical issues of digitization," Ethics and Information Technology, vol. 20, no. 2, pp. 127-142, 2018.

[29] T. Wischmeyer, and T. Rademacher, "Regulating Artificial Intelligence, vol. 1, no. 1, pp. 307-321. Springer, Cham, 2020.

[30] C. Eustace, The intangible economy: Impact and policy issues. Report of the European high level expert group on the intangible economy, 2000.

[31] S. Vincent-Lancrin, and R. Van der Vlies, Trustworthy artificial intelligence (AI) in education: Promises and challenges, 2020.

[32] J. Zhou, et al., "A survey on ethical principles of AI and implementations," in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 3010-3017. IEEE, 2020.

[33] A. B. Arrieta, et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information fusion, vol. 58, pp. 82-115, 2020.

[34] A. Rai, "Explainable AI: From black box to glass box," Journal of the Academy of Marketing Science, vol. 48, no. 1, pp. 137-141, 2020.

[35] S. Banuri, S. Dercon, and V. Gauri, "Biased policy professionals," The World Bank Economic Review, vol. 33, no. 2, pp. 310-327, 2019.

[36] M. Hind, "Explaining explainable AI," XRDS: Crossroads, The ACM Magazine for Students, vol. 25, no. 3, pp. 16-19, 2019.

[37] J. R. Rehse, N. Mehdiyev, and P. Fettke, "Towards Explainable Process Predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory," KI - Kunstl. Intelligenz, vol. 33, no. 2, pp. 181–187, 2019.

[38] T. Rojat, et al., "Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey," CoRR, vol. abs/2104.0, 2021. arXiv preprint arXiv:2104.00950.

[39] S. M. Lundberg, and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017.

[40] V. Roessner, et al., "Taming the chaos?! Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research," European Child & Adolescent Psychiatry, vol. 30, no. 8, pp. 1143-1146.

[41] G. Vilone, and L. Longo, "Explainable artificial intelligence: a systematic review", 2020. arXiv preprint arXiv:2006.00093.

[42] S. M. Lundberg, and S. I. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.

[43] M. S. de Carvalho, and G. L. da Silva, "Inside the black box: using Explainable AI to improve Evidence-Based Policies," in 2021 IEEE 23rd Conference on Business Informatics (CBI), vol. 2, pp. 57-64. IEEE, 2021.

[44] E. Tjoa, and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," IEEE Trans. Neural Networks Learn. Syst., vol. 32, no. 11, pp. 4793–4813, 2021.

[45] D. Doran, S. Schulz, and T. R. Besold, "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives," 2017.

[46] M. Dumas, et al., "Augmented Business Process Management Systems: A Research Manifesto," 2022. arXiv preprint arXiv:2201.12855.

[47] S. T. K. Jan, V. Ishakian, and V. Muthusamy, "Ai trust in business processes: The need for process-aware explanations," AAAI 2020 - 34th AAAI Conf. Artif. Intell., vol. 34, no. 08, pp. 13403–13404, 2020.

[48] G. Amit, F. Fournier, L. Limonad, and I. Skarbovsky, "Situation-Aware eXplainability for Business Processes Enabled by Complex Events," in Business Process Management Workshops: BPM 2022 International Workshops, pp. 45-57. Cham: Springer International Publishing, 2023.

[49] "Position Paper on the Artificial Intelligence Act", Ecommerce Europe, 2022. [Online]. Available: https://ecommerce-europe.eu/wp-content/uploads/2022/02/ECOM-AI-Position-Paper-01022022.pdf [Accessed: 27-Apr-2023].

[50] R. K. Bellamy, et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4-1, 2019.

[51] J. Wexler, "The what-if tool: Interactive probing of machine learning models," IEEE transactions on visualization and computer graphics, vol. 26, no. 1, pp. 56-65, 2019.

[52] G. Manias, A. Kiourtis, A. Mavrogiorgou, and D. Kyriazis, "Multilingual Sentiment Analysis on Twitter Data Towards Enhanced Policy Making," in Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings, Part II, pp. 325-337. Cham: Springer International Publishing, 2022.

[53] G. Manias, et al., "Real-time kafka-based topic modeling and identification of tweets," in 2021 IEEE International Conference on Progress in Informatics and Computing (PIC), pp. 212-218. IEEE, 2021.

[54] T. R. Goodwin, et al., "Automatic question answering for multiple stakeholders, the epidemic question answering dataset," Scientific Data, vol. 9, no. 1, p. 432, 2022.

[55] G. Manias, A. Mavrogiorgou, A. Kiourtis, and D. Kyriazis, "SemAI: A novel approach for achieving enhanced semantic interoperability in public policies," in Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings 17, pp. 687-699. Springer International Publishing, 2021.

[56] J. Zong, D. Barnwal, R. Neogy, and A. Satyanarayan, "Lyra 2: Designing interactive visualizations by demonstration," IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 304-314, 2020.

[57] L. M. Perkhofer, et al., "Interactive visualization of big data in the field of accounting: A survey of current practice and potential barriers for adoption," Journal of Applied Accounting Research, 2019.

[58] J. Li, C. Zhang, J. T. Zhou, H. Fu, S. Xia, and Q. Hu, "Deep-lift: deep label-specific feature learning for image annotation," IEEE Transactions on Cybernetics, vol. 52, no. 8, pp. 7732-7741, 2021.

[59] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis", 2017. arXiv preprint arXiv:1702.04595.

[60] B. K. Mohanta, S. S. Panda, and D. Jena, "An overview of smart contract and use cases in blockchain technology," in 2018 9th international conference on computing, communication and networking technologies (ICCCNT), pp. 1-4. IEEE, 2018.