

# AI4Gov

Trusted AI for Transparent Public Governance  
fostering Democratic Values

## Deliverable 4.1

### Trustworthy, Explainable, and unbiased AI V1


28-12-2023

Version 1.5



Funded by  
the European Union

*Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Agency. Neither the European Union nor the granting authority can be held responsible for them.*

PROPERTIES	
<b>Dissemination level</b>	Public
<b>Version</b>	1.5
<b>Status</b>	Final
<b>Beneficiary</b>	JSI
<b>License</b>	 <p>This work is licensed under a Creative Commons Attribution-No Derivatives 4.0 International License (CC BY-ND 4.0). See: <a href="https://creativecommons.org/licenses/by-nd/4.0/">https://creativecommons.org/licenses/by-nd/4.0/</a></p>

AUTHORS		
	Name	Organisation
<b>Document leader</b>	Alenka Guček	JSI
<b>Participants</b>	Matej Kovačič	JSI
	Kostis Mavrogiorgos	UPRC
	Shlomit Gur	IBM
	Inna Skarbovsky	IBM
	Lior Limonad	IBM
	Fabiana Fournier	IBM
	George Manias	UPRC
	Tanja Zdolšek Draksler	JSI
<b>Reviewers</b>	Silvina Pezzetta	WLC
	Dimitrios Ntalaperas	UBITECH

<b>VERSION HISTORY</b>				
<b>Version</b>	<b>Date</b>	<b>Author</b>	<b>Organisation</b>	<b>Description</b>
0.1	11/09/2023	Alenka Guček	JSI	Initial ToC
0.2	15/11/2023	Kostis Mavrogiorgos	UPRC	Update Sections 4.1 and 4.2
0.3	20/11/2023	Kostis Mavrogiorgos	UPRC	Update Sections 4.3 and 4.4
0.4	28/11/2023	Alenka Guček	JSI	Combine contributions of JSI IBM UPRC
0.5	30/11/2023	George Manias	UPRC	Update Sections 4.3 and 4.5
0.6	01/12/2023	Alenka Guček, Matej Kovačič	JSI	Update Abstract and Section 1
0.7	04/12/2023	Tanja Zdolšek Draklser	JSI	Update Sections 2.6 and 2.7
1.0	04/12/2023	Alenka Guček	JSI	Bibliography matching
1.1	11/12/2024	Silvina Pezzetta	WLC	Internal review
1.2	14/12/2024	Dimitris Ntalaperas	UBITECH	Internal review
1.5	20/12/2023	Alenka Guček	JSI	Integrated corrections

# Table of Contents

<b>Abstract .....</b>	<b>8</b>
<b>1 Introduction.....</b>	<b>9</b>
1.1 Purpose and scope.....	9
1.2 Document structure.....	9
1.3 Updates with respect to previous version (if any) .....	10
<b>2 Virtualized Unbiasing Framework (VUF) for AI &amp; Big Data .....</b>	<b>11</b>
2.1 Analysis and understanding of bias.....	11
2.1.1 <i>Human bias</i> .....	12
2.1.2 <i>Algorithmic bias</i> .....	12
2.2 AI bias mitigation .....	18
2.3 Methods to mitigate bias.....	18
2.4 Proactive accounting for bias.....	19
2.5 The Bias Detector Toolkit.....	19
2.5.1 <i>Scrollytelling application</i> .....	20
2.5.2 <i>Stages of AI trainings</i> .....	22
2.5.3 <i>Real life examples</i> .....	22
2.5.4 <i>Bias Detector Catalogue</i> .....	23
2.5.5 <i>Catalogue outputs specific to use cases</i> .....	23
2.6 Development for use cases.....	23
2.6.1 <i>SDG observatories and OECD papers</i> .....	23
2.6.2 <i>Data sources for ingestion</i> .....	24
2.6.3 <i>Methodology</i> .....	27
2.6.4 <i>Technologies</i> .....	28
2.7 "Trustworthy and Democratic AI" learning framework .....	31
2.8 Next steps with T4.1 .....	32
<b>3 Situation Aware eXplainability.....</b>	<b>34</b>
3.1 Introduction and background .....	34
3.2 What is Situation Aware eXplainability? .....	37
3.3 Explanations of process execution results .....	38
3.3.1 <i>Completeness of explanation</i> .....	38
3.3.2 <i>Soundness of explanation</i> .....	38
3.3.3 <i>Synthesis of explanation</i> .....	39
3.4 SAX4BPM library .....	39
3.4.1 <i>SAX4BPM architecture</i> .....	39
3.4.2 <i>SAX4BPM capabilities</i> .....	41
3.4.3 <i>SAX4BPM services</i> .....	42
3.4.4 <i>Illustrative example 1: Parking tickets</i> .....	44
3.4.5 <i>Illustrative example 2: Waste management use case</i> .....	52
3.5 Next steps with T4.2 .....	55
<b>4 Policy-Oriented AI and NLP Algorithms.....</b>	<b>57</b>
4.1 AI Algorithms for Policy Making.....	57
4.2 Current Advancements in NLP and QA .....	58
4.3 Multilingual NLP.....	59
4.4 Adaptive Analytics Framework .....	60
4.4.1 <i>Architecture and Internal Workflow</i> .....	60
4.4.2 <i>Baseline Technologies</i> .....	61
4.4.3 <i>Source Code - Availability and Key Points</i> .....	62
4.4.4 <i>User Guide – Installation and Use</i> .....	63
4.5 Policy-Oriented Analytics and AI Algorithms .....	66
4.5.1 <i>Architecture and Internal Workflow</i> .....	67
4.5.2 <i>Baseline Technologies</i> .....	72
4.5.3 <i>Source Code – Availability and Key Points</i> .....	73
4.5.4 <i>User Guide – Installation and Use</i> .....	73

4.6 Next steps with T4.3 .....75

**5 Conclusions..... 76**

**6 References ..... 77**

## List of figures

Figure 1: Schematic representation of Bias Detector catalogue components .....	20
Figure 2: Work in progress with description of visual metaphors being currently developed .....	21
Figure 3: The Event Registry pipeline .....	26
Figure 4: VideoLectures.NET platform .....	27
Figure 5: Enriched OECD Policy documents in our ELK backend infrastructure .....	29
Figure 6: SearchPoint implementation over OECD Policy documents .....	30
Figure 7: State-of-the-art of XAI applied to BPs .....	35
Figure 8: Process discovery (source (Van der Aalst, 2016)).....	35
Figure 9: SAX4BPM architecture .....	41
Figure 10: Parking scenario .....	45
Figure 11: Import log file .....	46
Figure 12: Process model discovery for the parking tickets scenario.....	47
Figure 13: Get variants from process model .....	48
Figure 14: Get a particular variant.....	48
Figure 15: Process model for the variant .....	49
Figure 16: Causal graph discovery for the variant .....	50
Figure 17: XAI graph for the parking tickets scenario.....	50
Figure 18: SAX explanation by applying LLM screenshot .....	51
Figure 19: SAX explanation by applying selecting process, causa, and XAI views screenshot.....	51
Figure 20: snippet of the waste management time series data input.....	53
Figure 21: Data segmentation model-fit based approach for the waste management example.....	55
Figure 22: Data segmentation functional-fit based approach for the waste management example.....	55
Figure 23: Architecture of Adaptive Analytics Framework Component .....	60
Figure 24: Adaptive Analytics Framework Source Code on GitLab.....	63
Figure 25: Sample of GitLab Issues Created for the Adaptive Analytics Framework.....	63
Figure 26: Architecture of Policy-Oriented Analytics and AI Algorithms Component .....	67
Figure 27: Multilingual Bias Classification Tool .....	71
Figure 28: Policy-Oriented Analytics and AI Algorithms Source Code on GitLab.....	73
Figure 29: Sample of GitLab Issues Created for the Policy-Oriented Analytics and AI Algorithms .....	73

## List of Tables

Table 1: Steps and services used in the parking scenario.....	45
Table 2: Steps and services used in the waste management example.....	52
Table 3: Data segmentation rule-based approach for the waste management example .....	54
Table 4: Description of the ai4gov_interactive_density_mapbox_api .....	64
Table 5: Description of the ai4gov_routing_optimization_api .....	65
Table 6: Description of the ai4gov_predict_traffic_violation_area_api.....	65
Table 7: Description of the ai4gov_time_series_forecasting_api .....	74
Table 8: Description of the ai4gov_qa_api_api .....	74

## Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
API	Application Programming Interface
AQA	Abstractive Question Answering
ARIMA	Autoregressive Integrated Moving Average
BERT	Bidirectional Encoder Representations from Transformers
BP	Business Process

CD	Causal Discovery
CDQA	Closed Domain Question Answering
CEP	Complex Event Processing
DL	Deep Learning
DNN	Deep Neural Networks
EQA	Extractive Question Answering
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
LiNGAM	Linear Non-Gaussian Acyclic Model
LLM	Large Language Model
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
ODQA	Open Domain Question Answering
OECD	Organisation for Economic Co-operation and Development
PD	Process Discovery
QA	Question - Answering
SAX	Situation-Aware eXplainability
SSC	Sustainable Smart Cities
SVG	Scalable Vector Graphics
VUF	Virtualized Unbiasing Framework
XAI	eXplainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

## Abstract

This document, D4.1 "Trustworthy, Explainable, and Unbiased AI V1", has been developed within the framework of WP4, "Trustworthy and Unbiased AI," as an integral part of the AI4Gov project. Released alongside D4.3, "Policies Visualization Services V1", on M12, these deliverables collectively offer an overview of the current advancements within WP4. D4.1 primarily focuses on elucidating the methodology and AI services, while D4.3 concentrates on the application's front-end, emphasizing interaction with users.

The document details the following technologies:

- the Virtualized Unbiasing Framework (VUF) for AI & Big Data – Bias Detector Toolkit,
- the SAX4BPM Library, and
- strategies to Improve Citizen Engagement and Trust utilizing NLP.

A thorough state-of-the-art analysis was conducted for each technology framework, alongside of providing a detailed description of their current states. Complementing these technological insights, this deliverable incorporates demonstrators effectively showcasing the proof of concept. These demonstrators are now transitioning to illustrate the ongoing progress of the pilot cases within the project.



# 1 Introduction

## 1.1 Purpose and scope

This deliverable is the results of the work in M1-M12 for tasks 4.1, 4.2 and 4.3 under the Work Package 4. The WP started in M1 and continues until M27. The second version of this deliverable is scheduled for M24. At this point, the deliverable describes the work on the aforementioned technical tasks, specifically the methodologies and services undergoing development for use case scenarios, that were identified under D6.1 on M6.

This deliverable is being released on M12 of the project, and its main purpose is to present AI4Gov progress in the development of the Bias Detector Toolkit, the XAI library and the NLP-enhanced analytics methodologies and applications.

In addition, it is worth to mention that the Federated Machine Learning (FML) approaches that are planned to be designed and implemented in the context of the project will be reported in the next iteration of this series of deliverables, i.e., in D4.2 Trustworthy, Explainable and unbiased AI V2 due on M24. At this stage of the project, the components and mechanisms reported in this document are standalone and their integration into a federated environment is stated for the second year of the project (M13-M24). As outlined in D2.3, the architecture of the AI4Gov platform follows a decentralized and federated approach through the utilization of a Kubernetes cluster that can facilitate the deployment of the AI4Gov platform on-premises, thus the training and implementation of FML models. However, it should be noted that based on D2.3 and D6.1 no such requirements and needs have been identified in the described use cases and scenarios. All pilot partners can share their data in the central infrastructure of the project without posing privacy or security concerns that necessitate on-premises deployment and training of the ML algorithms following federated approaches. Despite the latter, the consortium seeks to leverage the potential that can be derived from the utilization of FML approaches. Hence it is within our future plans to investigate such approaches even in a development and test environment.

## 1.2 Document structure

The deliverable is structured as follows: Chapter 1 introduces the document, including the purpose and scope, and document structure. The following chapters are dedicated to individual tasks from WP4, focusing on a diverse range of aspects towards Trustworthy, Explainable and Unbiased AI. Chapter 2 provides detailed information on T4.1, Virtualized Unbiasing Framework (VUF) for AI & Big Data. Chapter 3 is focused on the progress of work for T4.2, XAI Library. Chapter 4 introduces methodologies developed for T4.3, Improve Citizen Engagement and Trust utilizing NLP. Chapter 5 concludes the deliverable, summarizing the findings and describing the next steps. Chapter 6 includes the reference list. Since this deliverable is a demonstrator, videos for each of the tasks are available here: [AI4Gov demonstrators](#).

### 1.3 Updates with respect to previous version (if any)

This is the first of the two versions of this deliverable on Trustworthy, Explainable and Unbiased AI. The second version will be delivered on M24 in Dec 2024. Partially, the technologies undergoing development for tasks 4.1, 4.2 and 4.3 were previously introduced in D2.3 and focused on the architecture.

## 2 Virtualized Unbiasing Framework (VUF) for AI & Big Data

This section of the deliverable introduces a Bias Detector Toolkit designed to function as a visual catalogue synthesizing diverse tools tailored for detecting and mitigating biases in AI systems. To facilitate understanding of our approach, we will first provide a nuanced understanding of the myriad tools available, and later we shall describe the functionalities of the Toolkit. This section provides an introduction to bias and bias in AI, current state of the art, and then in detail describes progress on developing the Bias Detector Toolkit. Our aim is to empower a broad audience, including the general public, developers, researchers, and practitioners.

The Bias Detector Toolkit is a combination of the following subcomponents:

- Scrollytelling narrative to introduce the complexity of bias.
- Stages of training of AI models.
- Real life examples that ensure that all stakeholders grasp the significance of bias mitigation.
- Catalogue of bias detection and mitigation strategies, structured as visual summary.

Structured as a dynamic visual synthesis, this catalogue offers a comprehensive overview of bias mitigation tools, categorized by functionalities and applications. Serving as a visual catalogue, it facilitates exploration, comparison, and informed tool selection, thereby fostering a more effective approach to bias mitigation. The catalogue part is currently under development and will be showcased in the next iteration of the deliverable.

The final segment of this deliverable section shifts focus to the development of implementations for use cases and lastly introduces the learning framework, strategically designed to empower AI practitioners with the knowledge and skills essential for implementing bias mitigation strategies. This learning framework, integrating theoretical insights with practical exercises, addresses the educational gap in AI bias mitigation, fostering a collaborative community of practitioners. This holistic approach—from scrollytelling to bias detection tools and learning frameworks—forms a unified strategy towards advancing the field of fair and democratic AI.

### 2.1 Analysis and understanding of bias

Bias materializes as ingrained perspectives, preconceived notions, or unfair inclinations grounded in personal experiences, cultural influences, or societal conditioning. This cognitive phenomenon significantly shapes individuals' perceptions, judgments, and decision-making across diverse situations. As AI systems are crafted by human hands, incorporating data influenced by human experiences, bias invariably infiltrates these systems (Belenguer, 2022). The emergence of AI bias occurs when algorithms generate systematically prejudiced outputs due to biased assumptions during development or within the training data (Srinivasan & Chander, 2021). This bias exhibits its complexity at various stages of AI development, spanning from data collection and pre-processing to model training, evaluation, and deployment. The intricate challenge lies in the potential amplification of even the slightest biases throughout the machine learning (ML) process, leaving a lasting imprint on the entire system.

### 2.1.1 Human bias

Bias typically refers to the presence of preconceived notions, prejudices, or unfair preferences that individuals may have based on their personal experiences, cultural influences, or societal conditioning. Bias can affect how people perceive, judge, and make decisions about others or various situations. This inherent bias, when introduced into the process of data collection, poses a formidable challenge in the development of unbiased AI models. As data collection reflects the nuances of real-world scenarios, any biases present in the data become ingrained in the algorithms that learn from it. The inadvertent propagation of bias from real life to data and subsequently to AI models amplifies systemic prejudices and compromises the fairness and equity of the resulting models. Recognizing the implications of this transfer is crucial in mitigating bias at its roots and underscores the necessity of vigilant strategies in data collection processes to ensure the development of AI models that uphold ethical standards and avoid perpetuating societal biases.

### 2.1.2 Algorithmic bias

Detecting bias within an AI model becomes apparent when it produces skewed results, often perpetuating detrimental beliefs or prejudices against individuals or groups, deviating from positive human values like fairness and truth. Compounding this challenge is the potential introduction of intentional bias into AI systems, not merely as an unintended by-product but as a deliberate design choice. The multifaceted nature of AI bias, encompassing its inadvertent introduction, amplification, and intentional integration, underscores the need for comprehensive tools and frameworks to identify, address, and ultimately cultivate unbiased AI solutions.

#### 2.1.2.1 Types of biases

There are several stages where bias can occur, for instance, in data collection, data preprocessing, feature selection/engineering, model training, evaluation and deployment. In this section we are presenting some examples of bias that have happened in real life.

#### **Bias in criminal law**

The COMPAS software (Correctional Offender Management Profiling for Alternative Sanctions) is a decision support tool used by U.S. courts to assess the likelihood that a defendant would re-offend. The COMPAS software uses an algorithm to assess potential general recidivism risk, potential violent recidivism, and risk for pretrial misconduct.

COMPAS did not include race in calculating its risk scores. However, in 2016, ProPublica journalists investigated COMPAS and found that the system was far more likely to say black defendants were at risk of reoffending than their white counterparts. COMPAS software misclassified almost twice as many black defendants (45%) as higher risk compared to white defendants (23%), mistakenly labelled more white defendants as low risk, who then went on to reoffend – 48% white defendants compared to 28% black defendants and classified black defendants as higher risk when all other variables (such as prior crimes, age, and gender) were controlled – 77% more likely

than white defendants (*Can You Make AI Fairer than a Judge? Play Our Courtroom Algorithm Game | MIT Technology Review*, n.d.; *Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing | UCLA Law Review*, n.d.; *Machine Bias — ProPublica*, n.d.).

One study later criticized ProPublica's investigation findings, claiming that the authors erroneously concluded that the risk assessment instrument in question is racially biased and implied that such bias is inherent in all actuarial risk assessment instruments. However, some researchers later found that COMPAS algorithm is no better at predicting recidivism than random people, which raises the question of how reasonable is to use these algorithms at all (*A Popular Algorithm Is No Better at Predicting Crimes Than Random People - The Atlantic*, n.d.).

### **Bias in healthcare system**

In 2019, researchers found that an algorithm used in US hospitals to predict which patients will require additional medical care favoured white patients over black patients (Obermeyer et al., 2019a). The problem was that the algorithm considered the patients' past healthcare costs to predict their healthcare needs. They found out that black patients who were assigned the same level of risk by the algorithm were in fact sicker than White patients. They estimated that this racial bias reduced the number of Black patients identified for extra care by more than half.

The developers of the prediction algorithm were not aware that the expense of healthcare is significantly related to race. Black individuals with similar diseases spent less on healthcare than white patients with similar issues. And while the Black patients spent less money on health, the algorithm falsely concluded that Black patients are healthier than equally sick White patients.

After finding the bias, the researchers and a health services company that developed the prediction system worked on the problem and reduced bias by 80% (*A Biased Medical Algorithm Favored White People for Health-Care Programs | MIT Technology Review*, n.d.).

### **Bias in hiring algorithm**

Amazon has been using computer programs to review job applicants' resumes since 2014. At one point, they have developed an experimental hiring tool that used artificial intelligence to give job candidates scores ranging from one to five stars. This tool has been very useful for the human resources department; however, in 2015, they realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way.

The problem was that Amazon's AI models were trained on applications submitted to the company over a 10-year period. And, while most of the applications came from men, the algorithm concluded that male applicants were preferred and penalized resumes that indicated that the applicant was female (*Insight - Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women | Reuters*, n.d.).

When they discovered its new system for not evaluating applicants for software development jobs and other technical positions was biased towards women, they made the necessary changes to remove the bias. However, management lost faith in the initiative, because they became aware

that other biases can occur. So, in 2017 they stopped using AI for reviewing job applicants' resumes.

However, some companies are using automatic resume sorting programs, which rely primarily on simple pattern matching but can include some machine learning to detect what skills the applicant has listed with the job. Interestingly, some people found that if they copy a list of relevant keywords or just the published job description into the resume in a white font, the humans will not see it, but the computer will process the text (*Hacking AI Resume Screening with Text in a White Font* - Schneier on Security, n.d.). This shows that the bias could be introduced into the system in quite surprising ways.

### **Bias in algorithm to detect sepsis**

In 2019, a group of researchers in Duke University Hospital's emergency department started developing an algorithm to help predict childhood sepsis. It is a medical condition that is curable with antibiotics but is fatal for nearly 10% of kids in the United States. However, diagnosis is challenging because typical early symptoms (fever, high heart rate and high white blood cell count) can mimic other illnesses, including the common cold.

The research team has been aware of possible bias problems, and they spent a lot of effort to teach the algorithm to identify sepsis based on vital signs and lab tests. They implemented quality control tests to ensure the algorithm found sepsis equally well regardless of race or ethnicity.

However, after almost three years, they discovered possible bias. They found that doctors at Duke University Hospital took longer to order blood tests for Hispanic kids eventually diagnosed with sepsis than for white kids. One possibility for that was that the physicians were perhaps taking illnesses in white children more seriously than those of Hispanic children, but another reason could also be that the need for interpreters slowed down the process of ordering blood tests. This delay inaccurately taught AI that Hispanic kids develop sepsis slower than other kids. And that time difference resulting from the bias could be fatal (Obermeyer et al., 2019b).

At Duke University Hospital the team immediately started to solve the problem, and after eight weeks, they fixed the algorithm to predict sepsis at the same speed for all patients.

But more generally, this event increased awareness that transparency is crucial to determine if an algorithm is unbiased enough to be safely used on patients. More and more researchers are now starting to share best practices on how to tackle bias and there is also a push for new regulatory requirements. Because the result of the bias in AI algorithms leads to inequities in health care, this should be prevented.

### **Poisoning ML systems with non-randomness**

Sometimes bias can be introduced into the machine learning system deliberately. In that case an attacker tries to poison the ML model or introduce a backdoor into the system. The most obvious way to do this is to change the underlying dataset (add biased data) or model architecture. Moreover, researchers found out that this could also be achieved by only changing the order in which data are supplied to the model.

Let's imagine an example of a company that wants to have a credit-scoring system that is secretly sexist but also has a model architecture and data that would make it look like the model is fair. They could develop a machine learning model with high accuracy and collect a set of financial data that are highly representative of the whole population. But then they do not order the data to be randomly mixed, but instead, they start the model's training on ten rich men and ten poor women from that set. This would create the initialisation bias, which will then poison the whole system.

In the article "Manipulating SGD with Data Ordering Attacks" (Shumailov et al., n.d., 2021), the researchers have shown that if an adversary can manipulate the order in which batches of training data are presented to the model, they can undermine both its integrity (by poisoning it) and its availability (by causing training to be less effective, or take longer). They also found that this attack is not specific to the model or dataset, but rather targets the stochastic nature of all modern learning procedures.

Their work has highlighted the fact that machine learning models can suffer from so called sampling bias, where the sampling procedure can be manipulated to control the model's behaviour. Bias in machine learning is not just a data problem; it can be introduced in very subtle ways.

The so called stochastic nature of modern learning procedures means that the fairness of the model also depends on randomness. A random number generator with a backdoor can undermine a neural network and secretly introduce bias in the model that otherwise looks fair. This means that the AI developers should also pay attention to the training process and be specifically careful about their assumptions about randomness.

### **Bias in government fraud detection systems**

In 2023, the journalists of the Guardian newspaper found out that the British Home Office uses AI to flag up sham marriages. However, the investigation has shown that AI systems used to help decide who gets benefits and who should have their marriage licence approved have been biased against people of certain nationalities. The internal Home Office evaluation has shown that the tool disproportionately flags up people from Albania, Greece, Romania and Bulgaria (*UK Officials Use AI to Decide on Issues from Benefits to Marriage Licences | Artificial Intelligence (AI) | The Guardian*, n.d.).

The journalists also found out that the Department for Work and Pensions (DWP) uses an AI algorithm to detect fraud and error among benefits claimants. But there are some clues that the algorithm falsely flagged a lot of Bulgarians as making potentially fraudulent claims, and their benefits have been then suspended. The DWP and the Home Office would not disclose how the automated processes work, but they both claim the processes they use are fair because the final decisions are made by people.

But in both cases, the experts are warning about the dangers of using poorly understood algorithms to make life-changing decisions. They are pointing out that since officials have limited resources to review the cases, they highly rely on algorithm decisions. So, biased algorithms will

usually lead to biased final decisions and the people who are affected by those decisions would, in general, not know that the decision has been based on biased AI.

Unfortunately, this is not limited to Britain. In 2019 it has been revealed that the Dutch tax authorities had used a self-learning algorithm to create risk profiles in an effort to spot childcare benefits fraud. But the algorithm was not working properly and has been wrongly labelling people as fraudsters.

After several years of using this biased algorithm, the country's privacy regulator opened an investigation. They found out that the tax authorities focused on people with "a non-Western appearance" (more specifically, they targeted people with Turkish or Moroccan nationality), and among risk factors were also having dual nationality and a low income. Authorities penalized families over a mere suspicion of fraud based on the system's risk indicators.

The results were disastrous. Dutch tax authority demanded that people flagged as fraudsters pay back their childcare allowances, and tens of thousands of families were pushed into poverty because of exorbitant debts to the tax agency. Some victims committed suicide, and more than a thousand children were taken into foster care (*Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms – POLITICO*, n.d.).

After the investigation, the Dutch Data protection agency fined the Dutch tax administration 2.75 million EUR in December 2021 for the "unlawful, discriminatory and therefore improper manner" in which the tax authority processed data of child care benefit applicants (*Dutch DPA Fines Tax Authority 2.75M Euros*, n.d.). In April 2022, the Dutch data protection agency imposed another 3.7 million EUR fine on the Tax Administration for illegally processing personal data over a period of years in its "fraud identification facility" (*Tax Administration Fined for Fraud 'Black List' | European Data Protection Board*, n.d.). But several experts are warning that something similar could happen again.

When the scandal came to light, the Dutch government resigned, but they regrouped 225 days later.

### **Biased facial recognition systems**

Facial recognition technology is often used by police to confirm the identity of a person from an image. Police typically use this technology to identify stopped or arrested persons, to search video footage, or to search real-time scans of people passing surveillance cameras. The AI software compares the captured images to numerous photos and generates a line-up of potential suspects.

Of course, the software does not arrest people. It just suggests who are the potential suspects, and police officers then decide who to arrest. But people often believe that AI is infallible and don't question the results.

The automatization of the manual face-matching process can definitely help law enforcement to speed up investigations, accurately identify criminals and improve public safety. However, some research has shown that many facial recognition algorithms perform poorly at identifying people



besides white men. Researchers found (Johnson et al., 2022) that this is the result of two factors. First, there is the lack of Black faces in the algorithms' training data sets. And second, those systems often magnify police officers' own biases.

The fact that the technology struggles to distinguish darker faces often leads to more racial profiling and more false arrests. And inaccurate identification increases the likelihood of missed arrests.

The danger is not hypothetical. In 2020, Robert Williams has been arrested for allegedly stealing thousands of dollars of watches. Detroit police used an AI algorithm that matched video from the surveillance camera to the Williams' driver's license photo. The police did not collect any corroborating evidence such as eyewitness identification, cell phone location data or a fingerprint. The sole evidence has been a picture from a video surveillance camera and an identification by the AI algorithm (*First Man Wrongfully Arrested Because of Facial Recognition Testifies as California Weighs New Bills | California | The Guardian*, n.d.).

However, the problem was that Robert Williams wasn't the robber. But he was black. He has been arrested in front of his neighbours and family and detained for 18 hours. His arrest is the first documented case of someone being wrongfully detained based on facial recognition technology (*I Did Nothing Wrong. I Was Arrested Anyway. | ACLU*, n.d.).

He later learned about the study by the National Institute of Standards and Technology, which found that algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces (*Facing Bias in Facial Recognition Technology | The Regulatory Review*, n.d.).

Based on these findings, several U.S. cities banned or restricted government use of this technology, and the federal administration released the "Blueprint for an AI Bill of Rights" in 2022. Unfortunately, the blueprint's principles are nonbinding (*Blueprint for an AI Bill of Rights | OSTP | The White House*, n.d.), and industry is lobbying against tighter regulations of those systems.

### **Not a bug, but a feature**

Sometimes, bias can be deliberate, meaning that the system could be optimized in such a way, that discriminates certain groups of people.

UnitedHealth is the largest health insurance company in the US, and in November 2023, online media Stat News published their analysis of insurance's AI algorithm called nH Predict (*UnitedHealth Used Algorithms to Deny Care, Staff Say — STAT Investigation*, n.d.). The algorithm is used to predict how long patients will need post-acute care in nursing homes and rehab centres. The algorithm has been examining a database of medical cases from 6 million patients and estimating patients' medical requirements and length of stay.

But while UnitedHealth officially claimed that the algorithm is used to predict how long patients will need to stay in rehab, the journalists found that the managers inside the company pushed employees to follow an algorithm to cut off Medicare patients' rehab care by the date it predicted the medical care is no longer needed. Also, some former NaviHealth employees (the company

that developed the AI algorithm) said that once UnitedHealth took over, the focus moved from helping patients to making money and keeping post-acute care times as short as possible (*UnitedHealth Uses AI Model with 90% Error Rate to Deny Care, Lawsuit Alleges* | *Ars Technica*, n.d.; *UnitedHealthcare Accused of Using AI That Denies Critical Medical Care Coverage* | *TechSpot*, n.d.).

Soon after publishing the report, the insurance company has been sued over claims for using a flawed AI model. The lawsuit alleges that the AI system is biased towards older patients, that it has a 90% error rate and that it overruled the post-acute care opinions of physicians. The lawsuit also claims that the algorithm failed to take into account individual patient needs, such as comorbidities (having multiple conditions or diseases) and contracting an illness while staying at a facility.

The court action has not yet been decided, but it seems that in some cases, bias in AI algorithms is not a bug but is intently and maliciously used by some as a feature.

As we have shown through these examples, bias in AI systems can have serious consequences on society and everyday life of ordinary people. This shows why it is important that developers, researchers, practitioners, but also decision makers and the general public understand the problem of bias in AI and the need of fair and democratic AI solutions.

## 2.2 AI bias mitigation

Mitigating AI bias is imperative for the preservation of fairness and equity within AI applications. The endeavour involves not only the identification and rectification of biases within AI models but also the establishment of ethical guidelines and best practices in AI development to preclude bias inception. This multifaceted challenge demands collaborative efforts among data scientists, ethicists, policymakers, and the wider community to foster the creation of AI systems characterized by fairness, equity, and amicability towards all stakeholders.

## 2.3 Methods to mitigate bias

Bias can occur at all stages of model training, and it is therefore crucial to mitigate it just as widely. There is a notable surge in bias mitigation strategies focusing on specific stages of the ML process. To directly counteract biases in training data, adversarial training methods, such as adversarial debiasing have been developed (Zhang et al., 2018). Ensemble methods, like reweighing and ensemble adversarial training (Tramèr et al., 2017), are gaining prominence for their effectiveness in improving model fairness. Notable tools like IBM's AI Fairness 360 (Bellamy et al., 2018) and Google's What-If Tool offer practical implementations for bias detection and mitigation in ML models. Interpretability models, such as LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016a) and SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), contribute to transparency by explaining model decisions. These tools not only detect biases in ML models but also provide interpretable insights into how these biases manifest in predictions. By generating locally faithful explanations, LIME allows users to understand model decisions on specific instances, aiding in the identification and rectification of biases. Similarly, SHAP values

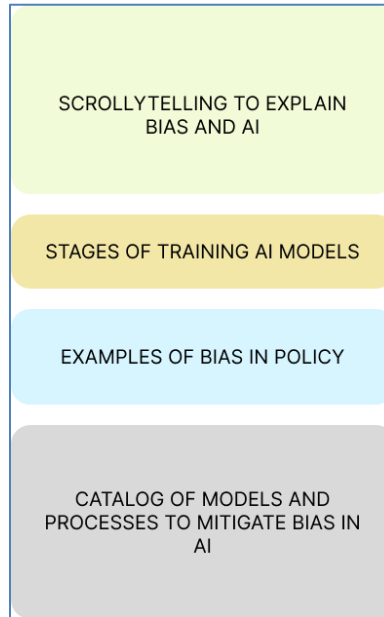
attribute model outputs to individual features, offering a transparent view of the impact each feature has on predictions. The synergistic integration of bias mitigation techniques and explainable artificial intelligence (XAI) tools represents a concrete step forward, ensuring not only fairer but also more interpretable AI systems. Despite these strides, challenges persist, driving ongoing exploration and refinement of algorithmic approaches for comprehensive and widely applicable AI bias mitigation solutions.

## 2.4 Proactive accounting for bias

Proactive accounting for bias in AI necessitates a multifaceted approach, and audits play a pivotal role in ensuring accountability and transparency. Rather than relying solely on post hoc assessments, incorporating proactive measures involves embedding bias awareness throughout the AI development life cycle. Initiating comprehensive bias audits at key stages, including data collection, model training, and deployment, is critical for identifying and rectifying biases before they become ingrained in the system. This approach not only minimizes the potential perpetuation of existing biases but also contributes to the creation of more robust and equitable AI systems. By conducting regular audits and incorporating feedback loops, organizations can adapt their AI models to evolving societal norms and continuously refine their algorithms to align with ethical standards. This proactive stance reflects a commitment to fostering responsible AI practices and underscores the importance of vigilance in mitigating bias from the outset.

## 2.5 The Bias Detector Toolkit

The Bias Detector Toolkit is a holistic application focused on explaining AI bias and equipping developers with an easy-to-navigate and visually organized catalogue. The Bias Detector Toolkit consists of the scrollytelling application, real life examples and the catalogue of methods and tools for bias mitigation (Figure 1). Since this part is being developed in close collaboration with T4.4, the demonstrator version is available for D4.3 only, not to duplicate the content (showcased in the D4.3 demonstrator [here](#)).



*Figure 1: Schematic representation of Bias Detector catalogue components*

### 2.5.1 Scrollytelling application

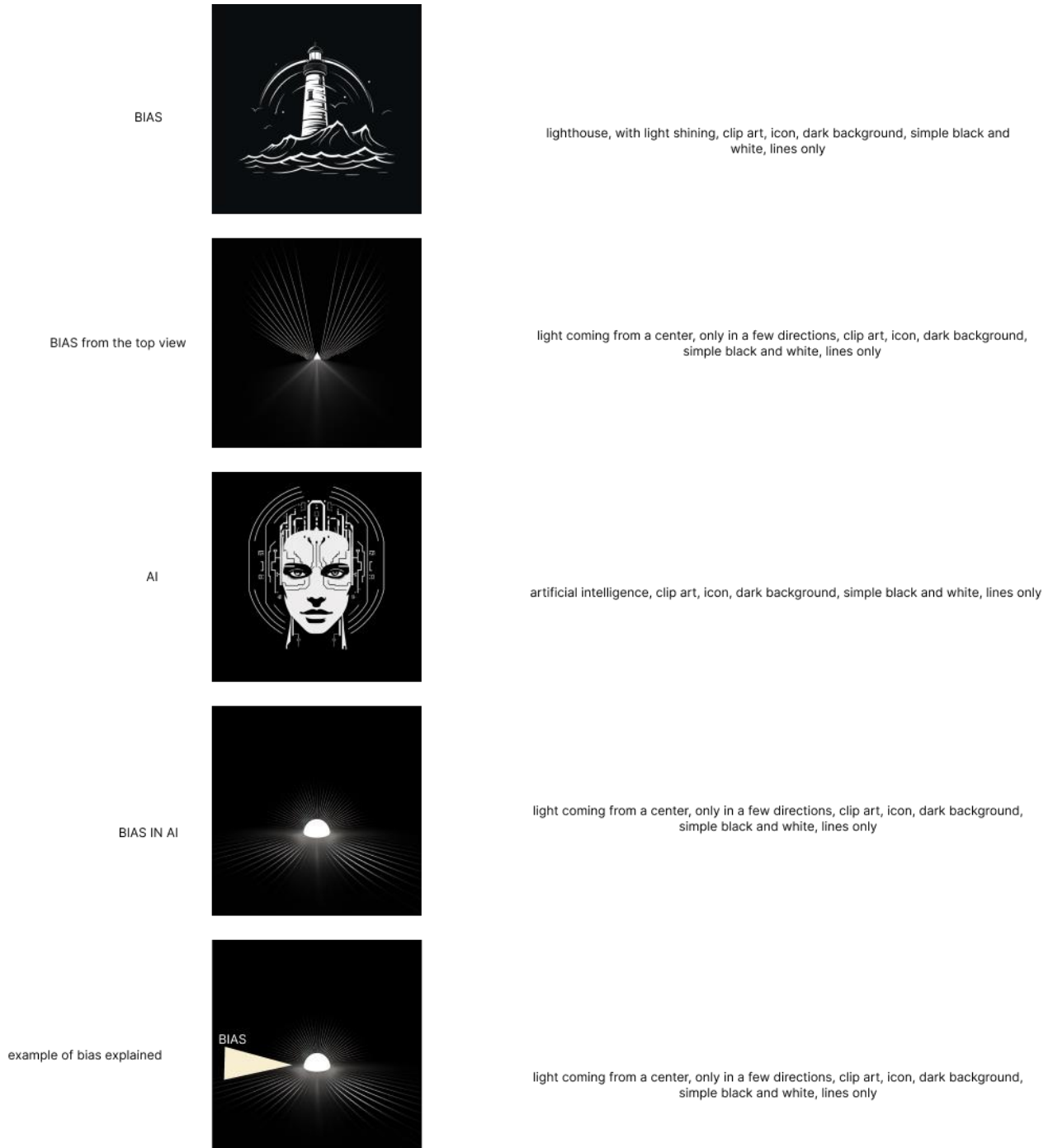
The first section of the application is the scrollytelling part, where the user can be informed about bias and AI in general. Scrollytelling, also known as scroll-driven storytelling or scroll-based storytelling, is a web design technique that involves using the scrolling action of a webpage to reveal content in a narrative or visually engaging way. Instead of presenting information all at once on a single page, scrollytelling unfolds content gradually as the user scrolls down the page.

Employed as an educational strategy, scrollytelling guides audiences through the complexities of bias in a step-by-step manner, utilizing visual metaphors and scalable vector graphics (SVG) animations. This approach transforms abstract concepts into tangible, visually intuitive representations, ensuring that even complex aspects of AI bias become comprehensible to a broad audience. As users scroll, they witness dynamic visual metaphors and animated SVGs demystifying the layers of bias in AI, from data collection nuances to algorithmic decision-making. In essence, scrollytelling serves as a pedagogical tool, breaking down the complexity of bias in AI through visually engaging storytelling, fostering understanding, and enabling a more accessible discourse on the challenges and ethical considerations surrounding bias in artificial intelligence.

By strategically organizing and presenting data visually, this process enhances comprehension and facilitates rapid understanding. In the current state, we have written the narrative part (showcased in D4.3 demonstrator [here](#)) and we are currently working on the animations, that will accompany the explanation and aid for the better understanding of these concepts to the general public and policy makers.

To introduce the concepts of bias, AI, and bias in AI, we are developing an animation-based scrollytelling experience featuring a visual metaphor of a lighthouse (see Figure 2). As the lighthouse illuminates its surroundings, certain areas persist in darkness, symbolizing the

presence of bias—divided into various forms. When AI enters the scene, the lighthouse's radiance expands, mirroring the broader implications of AI within policy contexts. This extended illumination brings bias to the forefront, facilitating its identification and subsequent mitigation. This metaphorical exploration aims to offer a nuanced comprehension of bias dynamics, where the interplay of light and shadow mirrors the intricacies inherent in the intersection of AI and policy.



*Figure 2: Work in progress with description of visual metaphors being currently developed*

### 2.5.2 Stages of AI trainings

The second section is a step-by-step presentation of bias in training a ML model. This process unfolds through several key stages. It begins with data collection, where relevant datasets are acquired to feed into the model. Following this, data preprocessing is of vital importance, focusing on the cleaning, normalization, and transformation of the raw data towards enhanced and effective learning. Feature selection follows, where meaningful attributes are chosen to enhance the model's performance and reduce complexity. Subsequently, the model training phase involves feeding the processed data into the chosen algorithm or architecture to enable it to learn patterns and relationships. The trained model is then evaluated using a separate validation set to gauge its accuracy and generalization capabilities. Once deemed satisfactory, the model proceeds to deployment, making it operational for real-world applications. Continuous monitoring and updates may follow to ensure its ongoing effectiveness, adapting to changes in data distribution or evolving problem domains. This cyclical process of data-driven learning, from collection to deployment, forms the foundation of ML model training.

### 2.5.3 Real life examples

The second section provides real life examples of bias occurrences in different business sectors (see section 2.1.2.1 for several examples). Bias in real-world applications of ML has manifested in various forms, raising ethical concerns and highlighting the importance of responsible AI development. For each example, there is:

- A short description that summarizes the problem.
- The solution that either solved or mitigated the problem.
- Some reference material for further research on the topic.

Real-life examples of bias in policies serve as digestible illustrations that underscore the critical importance of understanding and addressing systemic inequalities. For each step, a short description is provided along with ways that bias can occur. The intended use for this section is to provide policy makers, stakeholders and ML engineers with the information needed in order to prevent the occurrence of bias in workflows.

Consider standardized testing in education, for instance, where biases can disproportionately disadvantage certain demographic groups. Such policies can perpetuate societal disparities and hinder equal opportunities. The importance of recognizing these biases lies in the potential to empower a general audience. When individuals comprehend the tangible impacts of biased policies, they are better equipped to advocate for change, engage in informed discussions, and challenge discriminatory practices. By offering relatable examples, we empower the general audience to navigate and contribute to conversations about fairness, justice, and equitable policy reform in their communities and beyond. Real life examples are showcased in D4.3 demonstrator [here](#).

#### 2.5.4 Bias Detector Catalogue

The Catalogue builds upon training steps, providing tools and mitigation techniques for each of the steps described in 2.5.2. The central idea is not to create another text heavy framework, but to provide a visual summary of existing bias detection and mitigation strategies in an approachable and easy-to-grasp format.

For the Bias Detector Catalogue, we have executed an extensive literature overview for bias mitigation techniques, that are collected in our Gitlab repo<sup>1</sup>. We are now in the process of translating these inputs into a JSON format, that will serve as the input for the interactive visual synthesis on the AI4gov platform. The graphical representation will have a structure similar to data to the viz platform<sup>2</sup>.

#### 2.5.5 Catalogue outputs specific to use cases

For specific use cases, we will leverage the outputs of the catalogue to develop services specifically tailored for use cases. We are currently in the stage of co-designing what would be meaningful solutions, and D4.2 at M24 will present concrete outputs.

For the Top100 projects use case, we are in the process of translating the outputs of the literature overview to the catalogue, which will serve as the input for the checklists, that will be provided to applicants of Top100 projects upon application, and to reviewers upon review of applications. These checklists will ensure applicants have taken the steps to mitigate bias in the development of their AI models, and will equip reviewers to check for that. For other use cases, co-design is still in process.

### 2.6 Development for use cases

#### 2.6.1 SDG observatories and OECD papers

The current state of the methodology development for the SDG observatories (pilot no. 3, partner JSI) is a step forward in the pursuit of monitoring and supporting the achievements of the Sustainable Development Goals (SDGs). As outlined in D6.1, this use case specifically targets the establishment of SDG observatories, designed to assist policymakers by providing comprehensive insights into the progress towards these globally significant goals (refer to chapter 4.2.2.2 in D6.1 for additional details). The International Research Centre on Artificial Intelligence (IRCAI), under the auspices of UNESCO running as part of JSI, aims to promote international cooperation and collaboration in the development and deployment of AI for the benefit of humanity. The ultimate goal of the SDG observatories is to monitor how AI is supporting the achievements of SDGs.

---

<sup>1</sup> <https://gitlab.ai4gov-project.eu/ai4gov/t4.1-virtualized-unbiasing-framework-for-ai-and-big-data>

<sup>2</sup> <https://www.data-to-viz.com/>

The initial phase of the project involves the proof of concept, aiming to provide context experts with a tangible framework for their invaluable feedback. This stage encompasses an overarching approach that involves scouring media, policy documents, scientific literature, and educational materials to gather relevant data (see next section for detailed information on data sources). Under AI4Gov, the methodology for bias mitigation is being developed, focused on preventing data biases, as explained in demonstrator [here](#).

To raise awareness of the importance of ethics in AI, and the importance of bias prevention approaches, the ingestion of OECD papers, consisting of various national AI policies and strategies, is one of the data sources to be included in SDG observatories.

### 2.6.2 Data sources for ingestion

IRCAI SDG Observatories have currently ingested the following data sources:

- Published scientific publications on SDG-related topics sourced in OpenAlex,
- Worldwide and multilingual news and events from EventRegistry,
- Educational resources on SDG topics from the Videlectures.NET portal,
- OECD Policy documents on AI policies relating to SDGs.

In the next sections, these data sources are described in detail.

#### 2.6.2.1 OpenAlex

OpenAlex is a free and open catalogue of the world's scholarly research system - scholarly publications, authors, institutions, venues and concepts (Wikidata concepts are linked to works via an automated hierarchical multi-tag classifier). In that sense, OpenAlex could be seen as a fully open scientific knowledge graph.

OpenAlex catalogues 243 millions scientific documents, and 48 million of them are open access works.

OpenAlex is aggregating and standardizing data from various data sources. The main are former Microsoft Academic Graph (which has been discontinued in 2021), Crossref (open digital infrastructure recording and connecting knowledge through open metadata and identifiers for all research objects such as grants and articles), ORCID, ROR, DOAJ, Unpaywall, Pubmed and Pubmed Central, The ISSN International Centre and many more.

#### 2.6.2.2 News and events from Event Registry

For ingesting news and extracting events, we are using a news intelligence platform developed by JSI, called Event Registry. Event Registry is able to process news articles published in different languages world-wide. By analysing the articles, it can identify the mentioned events and extract the main event information. Extracted event information is stored in a structured way that provides unique features such as searching for events by date, location, entity, topic and other event properties.



Event Registry consists of a pipeline of components that each provide unique and relevant features for the system (Figure 3). In order to identify events, we first need data from which we can extract the information. In Event Registry we use news articles as the data source. The news articles are collected using our NewsFeed service, which collects news articles from more than 150.000 worldwide news sources. Collected articles are in more than 40 languages, where articles in English, Spanish, German and Chinese languages amount to about 70% of all articles. These languages are also the only ones we use in the Event Registry. Each collected article in the mentioned languages is then analysed in order to extract relevant information. One of the key tasks is the identification and disambiguation of named entities and topics mentioned in the article. We perform this task using a semantic enrichment service. We also detect date mentions in the text, since they frequently specify the date when the event occurred.

During the processing of ingested data, we identify duplicated articles, even if they are in different languages (we use canonical correlation analysis, which maps articles from different languages into a common semantic space; in order to train the mapping to the common space, we used the aligned corpus of Wikipedia articles).

After extracting relevant features from each individual article, we start with the main task of identifying events from groups of articles. In order to identify events, we therefore apply an online clustering algorithm on articles as they are added into the system. Each identified cluster of articles is considered to describe an event if it contains at least a minimum number of articles (the minimum value used in our system is 5 articles). Once a cluster reaches this limit, we treat it as an event and the information about its articles is sent to the next components in the pipeline. Those components are responsible for extracting event information from the articles.

From the associated articles in the cluster, we then try to extract the relevant information about the event. We try to determine the event date by seeing if there is a common date reference frequently mentioned in the articles. If no date is mentioned frequently enough, we use the average article's published date as the event date. The location of the event is determined by locating frequently mentioned named entities that represent locations. To determine what the event is about, we aggregate the named entities and topics identified in the articles. All extracted information is stored in the Event Registry in a structured form that provides rich search and visualization capabilities.

Event Registry contains more than 15 million articles from which we identified about 1 million events (from January 2016).

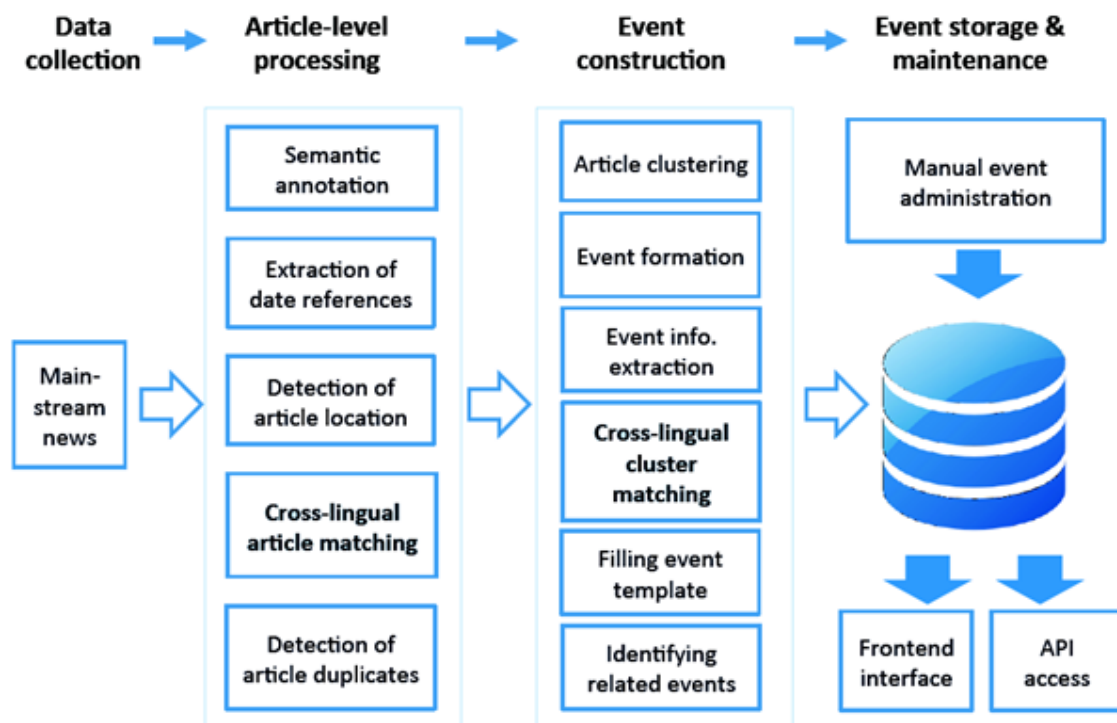


Figure 3: The Event Registry pipeline.

### 2.6.2.3 Videlectures.NET platform

VideLectures.NET is an award-winning free and open access educational video lectures repository developed by JSI. The rich repository already contains 30.187 videos, most of which are accompanied by slides, and the majority of these is also accompanied by transcripts and translations. VideLectures.NET aims to promote science, exchange ideas and foster knowledge sharing by providing high quality didactic contents not only to the scientific community but also to the general public on a global scale. Lectures are also categorized.

VideLectures.NET (Figure 4) hosts videos from various academic and research institutions, conferences, and events and covers various disciplines such as computer science, machine learning, artificial intelligence, biology, psychology, medicine, mathematics, and more.

VideLectures.NET operates on the principles of open access. Users can freely view and download the educational content and lectures are published under an open license (CreativeCommons).

The platform contains more than 25.000 peer-reviewed lectures from more than 19.000 authors in various languages.



The screenshot shows a video lecture player interface. The main content is a slide with the following text:

Logo of University of Science and Technology of China (USTC) and National University of Singapore (NUS).

# Bias Issues and Solutions in Recommender System

Jiawei Chen, Xiang Wang, Fuli Feng, Xiangnan He  
[cjwustc@ustc.edu.cn](mailto:cjwustc@ustc.edu.cn)

slides will be available at: <https://github.com/jiawei-chen/RecDebiasing>  
A literature survey based on this tutorial is available at: <https://arxiv.org/pdf/2010.03240.pdf>

Logo of THE WEB CONFERENCE.

Video player controls at the bottom show a play button, volume icon, and a progress bar at 0:00 / 57:29. A small video inset on the right shows the presenter.

## Bias Issues and Solutions in Recommender System

Figure 4: VideoLectures.NET platform.

### 2.6.2.4 OECD Policy documents

OECD has a collection of various national AI policies and strategies. They have an online repository with over 800 AI policy initiatives from 69 countries, territories and the EU.

We have developed tools for web scrapping these documents and converting them to Markdown format; those documents have already been ingested into our platform and enriched with the automatic classification into relevant SDG's.

### 2.6.3 Methodology

#### 2.6.3.1 Concept mapping for SDG observatories

The methodology employed by the team leverages UNESCO experts' provided list of 28-69 keywords for each SDG, forming the foundation for subsequent analyses. This is our priority baseline.

The internal analysis involves mapping these keywords onto concepts from previously described data sources. See the demonstrator for further explanation and showcase [here](#).

- News: choice of Wikipedia concepts close to the priority baseline, that represent concepts in the news query through the Event Registry engine.
- Science: choice of OpenAlex concepts<sup>3</sup> (established by the Microsoft Academic Graph) close to the priority baseline, within the existing categories.
- Policies: choice of keyphrases that are well represented (retrieve more than 3 results) in the dataset and are close to the priority baseline.
- Education: choice of keyphrases that are well represented (retrieve more than 10 results) in the dataset and are close to the priority baseline.

Criteria such as the existence of the concept, the discovery of a certain number of hits, and the meaningful connection to IRCAI's work guide the selection of five keywords per SDG. Once identified, data from these sources is ingested and labelled with corresponding SDG tags and topic tags before being loaded into Elastic search, forming a structured and searchable database. This step is crucial to prevent any bias coming in due to the wrong selection of keywords, that are not matched to the concepts and would only show part of the data, possibly leading to biased representation.

### *2.6.3.2 Bias detection in specific datasets for SDG observatories*

As part of the methodology for SDG observatories, we are developing specific tools to detect bias in the ingested data. For example, for SDG covering health, one of the focuses are rare diseases, and one of the data sources we wish to include is patient reported data. Patient reported data is an important data source for clinicians, but there are several challenges, since this data needs to be collected from the patients themselves or their caregivers (meaning they need to have a certain level of data literacy), and input questionnaires need to preferably in their mother tongue. GENIDA<sup>4</sup> is one such example of a database with patient-reported data.

We are currently developing a pipeline that will compare the incidence of specific syndromes and compare that to the reported data of GENIDA, allowing us to pinpoint where the data is missing. We already have the data and are now designing steps to create the pipeline.

### 2.6.4 Technologies

Currently, we have implemented two technologies for our SDG Observatories: the ELK stack and SearchPoint.

---

<sup>3</sup> <https://api.openalex.org/concepts>

<sup>4</sup> <https://genida.unistra.fr/register/>

### 2.6.4.1 ELK stack

The IRCAI SDG Observatories backend infrastructure will be based on Elasticsearch. The so-called ELK stack can ingest and aggregate data from various data sources and provide support for the analysis and visualizations of this data.

ELK is an acronym that refers to a set of three open-source tools: Elasticsearch, Logstash, and Kibana. Together, they provide a comprehensive solution for collecting, storing, and visualizing log data.

Elasticsearch is a distributed, RESTful search and analytics engine. It is used for storing, searching, and analysing large volumes of data quickly in near real-time. Elasticsearch is particularly well-suited for text-based data, as shown in the example in Figure 5.

Logstash is a server-side data processing pipeline. It is responsible for ingesting data from multiple sources, processing it, and then sending it to the desired output (in our case Elasticsearch). It provides a wide range of input, filter, and output plugins that allow it to handle various types of data and integrate with different systems.

Kibana is a web interface for searching, visualizing, and interacting with the data stored in Elasticsearch. It enables users to create and share dynamic dashboards and visualizations. It provides a user-friendly way to explore and understand the data stored in Elasticsearch.

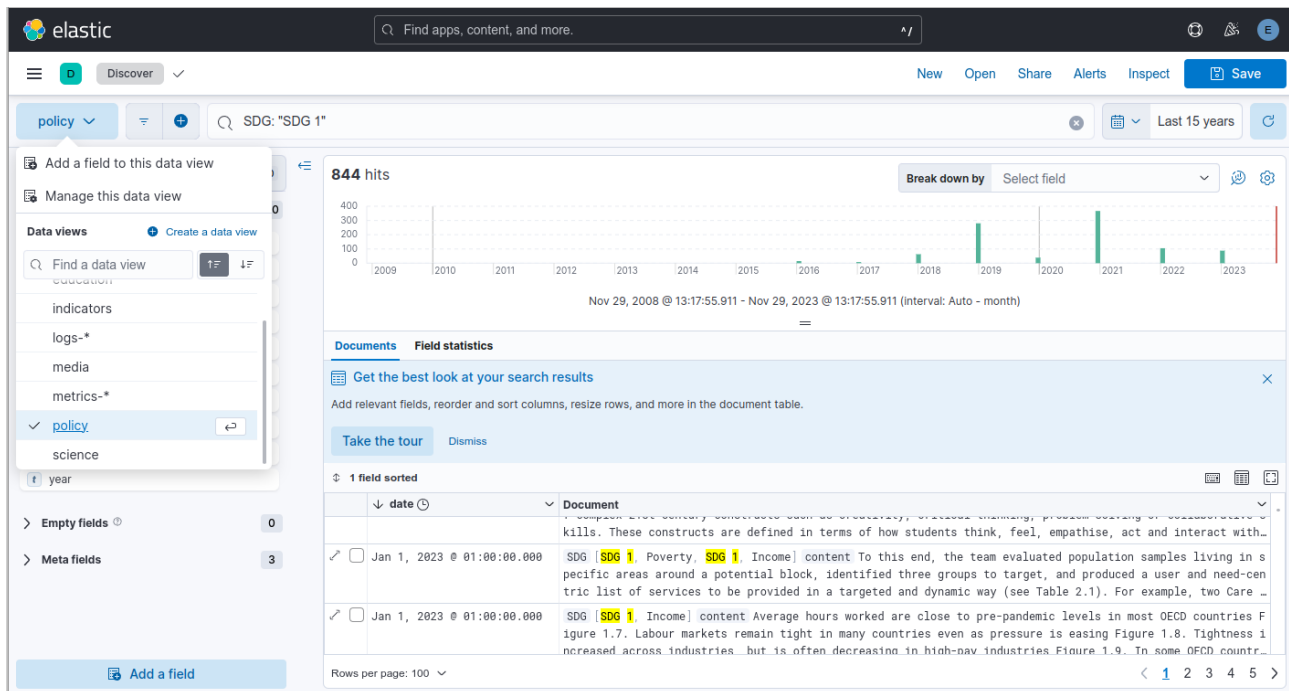


Figure 5: Enriched OECD Policy documents in our ELK backend infrastructure.

### 2.6.4.2 SearchPoint

SearchPoint is JSI’s tool for visual and contextualized web browsing. It allows users to visually re-rank search results. When a query is first made, the results are visualized using topic, concept and location. The user is then able to filter and re-rank complete results by interacting with the visualization.

SearchPoint (Figure 6) is solving the problem of disambiguation of the queries, because existing ontologies and taxonomies can provide different context for the same data.

For instance, “jaguar” could be an animal, a car, a music band, a movie, etc. Also, “A4” could be a car, paper size, etc. SearchPoint can automatically detect different topics and visualize results relevant to them. It builds a classification model, automatically classifies data into underlying categories and enables the user to rearrange search results according to the selected topic.

The screenshot shows the SearchPoint interface. At the top left is the SearchPoint logo. Below it is a search bar containing the text "Gender". To the right of the search bar are two dropdown menus: "SDG" and "Policies", followed by a "Search Topics" button. Below the search bar is a downward-pointing chevron icon. The main content area displays two search results:

- (0) [The Effects of AI on the Working Lives of Women](https://www.oecd-ilibrary.org/deliver/14e9b92c-en.pdf?itemId=%2Fcontent%2Fpublication%2F14e9b92c-en&mimeType=pdf)**  
<https://www.oecd-ilibrary.org/deliver/14e9b92c-en.pdf?itemId=%2Fcontent%2Fpublication%2F14e9b92c-en&mimeType=pdf>  
Some of its outputs – such as Automation, Skills Use and Training and AI and the Future of Skills, Volume 1 – analyse the impacts of AI on jobs and skills for different societal groups, including women. AI policy, human rights and the Sustainable Developm...
- (1) [Positive, High-achieving Students? : What Schools and Teachers Can Do](https://www.oecd-ilibrary.org/deliver/3b9551db-en.pdf?itemId=%2Fcontent%2Fpublication%2F3b9551db-en&mimeType=pdf)**  
<https://www.oecd-ilibrary.org/deliver/3b9551db-en.pdf?itemId=%2Fcontent%2Fpublication%2F3b9551db-en&mimeType=pdf>  
Do female teachers act as role models for girls, especially when it comes to closing the gender gap in mathematics or science? In the same way, can male teachers act as role models for boys to close the gender gap in reading performance? Do teachers' atti...
- (2) [Linking Aid to the Sustainable Development Goals – a machine learning approach](https://www.oecd-ilibrary.org/deliver/4bdaeb8c-en.pdf?itemId=%2Fcontent%2Fpaper%2F4bdaeb8c-en&mimeType=pdf)**  
<https://www.oecd-ilibrary.org/deliver/4bdaeb8c-en.pdf?itemId=%2Fcontent%2Fpaper%2F4bdaeb8c-en&mimeType=pdf>  
Ensure healthy lives and promote well-being for all at all ages  
Enhancing Collaboration in Pursuit of SDG 4: Literacy and Lifelong Learning Case Studies: Case Study SDG 4 "Education" Sustainable Development Goal 4 and Refugee Education PISA for Developmen...

To the right of the search results is a word cloud visualization. The words in the cloud include: FAVOUR, AVERAGE, TEACHERS, STUDENTS, TALIS, PISA, SCHOOLS, PERFORMANCE, ACHIEVING, GIRLS, BOYS, SECTORS, ARTIFICIAL, EDUCATION, PUBLIC, READING, PISA, OECD, DIGITAL, STEREOTYPES, ALLIANCE, WOMEN, LIVES, TECHNOLOGIES, RESEARCH, DEVELOPMENT, WORKING, RIGHTS, WORKERS, GOALS, INTELLIGENCE, SYSTEMS, AI, DATA, OPPORTUNITIES, GENDER EQUALITY, SUSTAINABLE DEVELOPMENT, DIGITAL GENDER, CONSEQUENCE, POLICY, SDGS, OECD, SKILLS, STUDENTS, MATHS, LEARN, JOBS, EFFECTS, FORCE, ARMED, STATES, COAST, DEFENSE, NATIONS, FISCAL YEARS, AFGHAN, SECURE, GUARD. At the bottom right of the word cloud is the SearchPoint logo.

Figure 6: SearchPoint implementation over OECD Policy documents

The current emphasis is on the development of a user-friendly front-end for the SDG observatories. This interface aims to empower content experts by enabling them to explore the ingested data comprehensively. Their feedback becomes a crucial component in refining the observatories further, guiding the inclusion of additional keywords for a more nuanced understanding of the SDGs.

Looking ahead, the plan is to align the SDG observatories with advancements in AI. This strategic direction will involve leveraging AI technologies to advance towards the completion of SDG goals, providing deeper insights and facilitating more informed decision-making by policymakers and stakeholders. The ongoing efforts reflect a commitment to advancing the methodology and technology underlying the SDG observatories, in the direction of mitigating possible biases occurring during the data collection and handling stages, contributing significantly to the broader goals of sustainable development and the European Green Deal.

## 2.7 "Trustworthy and Democratic AI" learning framework

With the overarching goals of promoting trustworthy AI and reinforcing the efforts of the Bias Detector Toolkit, work for T4.1 is in close collaboration with T5.2 and T5.3, focused on learning and training. A learning framework (curriculum) for "Trustworthy and Democratic AI" was developed that serves as a bridge between theoretical understanding and practical application, fostering a holistic approach that empowers participants to contribute effectively to the responsible development and deployment of AI systems.

The objective of the "Trustworthy and Democratic AI" learning framework is to equip learners with the knowledge, skills, and ethical principles necessary to design, develop, and deploy artificial intelligence systems that are both trustworthy and free from bias. This framework aims to address the growing need for responsible AI development by fostering a deep understanding of the complexities and challenges related to AI bias, fairness, transparency, and accountability.

Learners will be equipped with the knowledge and skills required to identify bias and to contribute to the responsible development and deployment of AI systems, ensuring that AI technologies are both effective and ethically sound. The learning framework empowers learners to become advocates for trustworthy and debiased AI in their respective roles and organizations.

Within the specific learning objectives from the learning framework (details below), learning objectives no. 5, no. 6, no.9 and no. 10 are the ones directly linked to task 4.1 and this deliverable.

Specific Learning Objectives:

### 1. Understanding AI Fundamentals:

- o Understand the concepts in AI, ML, and deep learning.
- o Recognize the transformative potential and ethical implications of AI in various domains.

### 2. Bias Awareness and Recognition:

- o Identify the types and sources of bias that can manifest in AI systems.
- o Understand the real-world consequences of biased AI, including social and ethical implications.

### 3. Ethical Considerations in AI:

- o Explore ethical frameworks and principles guiding AI development.
- o Recognize the importance of fairness, transparency, and accountability in AI systems.

4. **Data Collection and Pre-processing:**
  - o Learn how data collection and pre-processing can introduce bias into AI models.
  - o Implement best practices for collecting, cleaning, and preparing data to reduce bias.
5. **Bias Mitigation:**
  - o Understand various techniques to mitigate algorithmic bias in AI models.
  - o Apply debiasing methods, reweighting strategies, and fairness constraints to model development.
6. **Interpretable AI:**
  - o Examine methods for making AI models more interpretable and explainable; explainable AI (XAI).
  - o Appreciate the importance of transparency in AI decision-making processes.
7. **Ethical AI Governance:**
  - o Study legal and regulatory frameworks governing AI.
  - o Explore the role of responsible AI governance in organizations.
8. **Responsible AI Development:**
  - o Apply ethical guidelines and responsible AI principles to the end-to-end AI development lifecycle.
  - o Develop AI systems that prioritize fairness, accountability, and transparency.
9. **AI Trustworthiness Evaluation:**
  - o Assess AI models and systems for fairness, bias, and trustworthiness.
  - o Implement evaluation techniques to ensure AI systems meet ethical and operational standards.
10. **Real-World Applications:**
  - o Analyse case studies and real-world examples of trustworthy AI.
  - o Gain practical experience in addressing bias and trustworthiness challenges in AI projects.

Based on the learning framework, dedicated massive open online courses (MOOCs) will be developed in the future, each will include: a) theory, b) contextualized resources per use case and c) assessment materials. MOOCs will be composed of text content, video lectures, graphics, interactive quizzes etc.

## 2.8 Next steps with T4.1

This chapter summarized the current state of T4.1. We plan to continue with the work on 1) finishing the scrollytelling explanation with intuitive animations to aid understanding; 2) mapping



the detection & mitigation models already published in visual synthesising them into Bias Detector Catalogue; 3) taking the outputs of the catalogue for creating use-case specific checklists; 4) finishing the pipelines for bias detection in specific datasets ingested for SDG observatories and 5) possibly creating new models for bias detection, if needed for use cases.

## 3 Situation Aware eXplainability

In the scope of the AI4Gov project, our aim is to push the boundaries of eXplainable AI (XAI) to cope with the challenges existing in business processes (BP) and to produce reliable and faithful explanations about decisions and outcomes of BP executions.

In the following sections we describe what these challenges are and our proposed approach. We start with some fundamental concepts to get a common ground of understanding the problem and the offered solution. This solution is mainly composed of a library of services supporting different capabilities addressed in the explanations provided that will be released as open source by the end of the project. We illustrate this library using two examples stemming from the project use cases: parking tickets and waste management.

### 3.1 Introduction and background

A **business process (BP)** is a collection of tasks that are executed in a specific sequence to achieve some business goal (Weske, 2012). The digital footprint that depicts a single execution of a process as a concrete sequence of activities or events is termed a 'trace' (Van der Aalst, 2016). A multi-set of traces is usually referred to as a trace-log or event-log.

**eXplainable AI (XAI)** - Recent advancements in Machine Learning (ML) (Adadi & Berrada, 2018; Meske et al., 2022; Verma et al., 2021) have been achieved with increase in the complexity of models that require external explanation frameworks. XAI is the ability to understand and interpret how AI systems make decisions or arrive at conclusions. Such frameworks are predominately developed for post-hoc interpretations of ML models (Adadi & Berrada, 2018; Meske et al., 2022). Context-wise, they can be divided into global, local, and hybrid explanations (Adadi & Berrada, 2018; Guidotti et al., 2018; Rehse et al., 2019). Global explanations attempt to explain the ML model's internal logic, local explanations try to explain the ML model's prediction for a single input instance, and hybrid approaches vary (e.g., explaining the ML model's internal logic for a subspace of the input space).

Contemporary techniques of XAI application to BPs are shown in Figure 7. The ML model serves as a surrogate model typically trained using historical process execution logs of the BP. The predicted value for a single instance (process outcome) serves as input for the XAI explainer to produce an explanation. Our work adds to a series of recent efforts (Amit et al., 2022; Upadhyay et al., 2021) that focus on exploiting XAI frameworks that are compatible with tabular data for the interpretation of BP execution results. We use process logs as the main data input and train surrogate ML models with this data to represent real-world business processes.

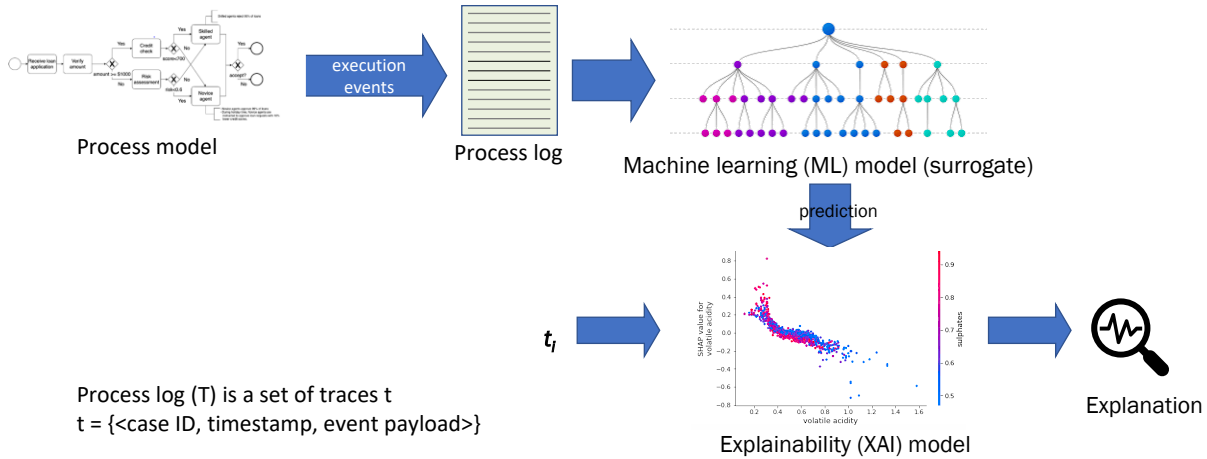


Figure 7: State-of-the-art of XAI applied to BPs

However, contemporary techniques are not adequate to produce explanations faithfully and correctly when applied to BPs as they generally fail to:

- express the business process model constraints (i.e., semantics of the process model),
- include the richness of contextual situations that affect process outcomes (additional information that affects the outcome but usually not modelled),
- reflect the true causal execution dependencies among the activities in the business process (see causal discovery below), or
- make sense and be interpretable to process users (explanations are usually not given in a human-interpretable form that can ease the understanding by humans).

**Process discovery (PD)** – Process discovery (PD) aims at gaining insights into the business processes of organizations by analyzing event data recorded in their information systems, for the sake of business improvement (Van der Aalst, 2016). PD summarizes an event log  $L$  into a graph model  $M$  that represents activities and control-flow dependencies (Leemans, 2019). Typical process mining techniques are inherently associational, approaching process discovery from a time precedence perspective, i.e., they discover ordering constraints among the process’ activities. That is, most PD algorithms construct edges in  $M$  that indicate to which subsequent activities process control “flows to” (Figure 8).

Table 2.1 A fragment of some event log: each line corresponds to an event

Case id	Event id	Properties				...
		Timestamp	Activity	Resource	Cost	
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
2	35654427	07-01-2011:14.24	reject request	Pete	200	...
	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
35654489	08-01-2011:12.05	pay compensation	Ellen	200	...	

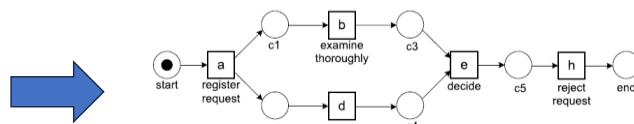


Fig. 2.7 The process model discovered by the  $\alpha$ -algorithm based on cases 1 and 4, i.e., the set of traces  $\{(a, b, d, e, h), (a, d, b, e, h)\}$

Figure 8: Process discovery (source (Van der Aalst, 2016))

**Causal discovery (CD)** – Causal relationships describe the connection between a cause and its effect, where the cause is an event that contributes to the production of another event, the effect (Pearl, 2011). Causal discovery is responsible for creating models (causal graphs) that illustrate the causal relationships inherent in the data (Pearl, 2011; Sadat Qafari & van der Aalst, n.d.; Spirtes et al., 2001). Identification of causal relationships is key to the ability to reason about the consequences of interventions. One of the fundamental goals of causal analysis is not only to understand exactly what causes a specific effect but rather to be able to conclude if certain interventions account for the formation of certain outcomes, thus being able to answer questions of the form: Does the execution of a certain activity in a loan approval process entail a delay in the handling of the application? Or, if the same activity is skipped, may the process duration be shortened? In our work, we focus on causal discovery and adapt for its employment over process execution times, as inherently recorded in process event logs. More specifically, we leverage the Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu, 2022) for CD as in (Fournier et al., n.d.) to uncover the causal execution dependencies among the events in event logs.

**Large language models (LLMs)** – LLMs are deep-learning (DL) models trained on vast amounts of text data to perform various natural language processing tasks. One of their strengths is their ability to perform few-shot and zero-shot learning with prompt-based learning (Liu et al., 2023).

**Complex event processing (CEP)** is computing that is performed on streaming data (sequence of events) for the purpose of stream analytics or stream data integration. CEP is typically applied to data as it arrives (data “in motion”). It enables situation awareness and near-real-time responses to threats and opportunities as they emerge, or it stores data streams for use in subsequent applications (Etzion & Niblett, 2010). The results of CEP computation are complex events. A complex event may be derived from just a few or from millions of base (input) events from one or more event streams. Stream analytic applications provide continuous intelligence to enhance situation awareness, enable sense-and-respond behavior or just inform real-time decisions. Organizations are doing more stream processing because of the need for continuous intelligence and better situation awareness, as well as faster, more precise business decisions (*Market Guide for Event Stream Processing*, n.d.).

**Performance versus explainability** – There is often a tradeoff between the performance (e.g., prediction accuracy) and the explainability of ML models. “Simpler” ML models (e.g., linear models, rule-based models and decision trees) can be inherently explainable (that is, their internal “reasoning” is human-understandable), but seldom do they perform well (e.g., better than a human agent) on real-world data. More complex ML models (e.g., Deep Neural Networks (DNNs) (Bengio, 2009) may perform well on real-world data but are often black boxes and thus cannot be easily explained.

*Our goal is to combine PD, CD, and XAI to generate narratives for improved process outcome explanations using LLMs.*

### 3.2 What is Situation Aware eXplainability?

*Situation aware eXplainability (SAX)* are evolutionary XAI techniques applied to BPs that aim at tackling the shortcomings of contemporary XAI techniques when applied to BPs as aforementioned. More specifically, a **situation-aware eXplanation** is a *causal sound explanation* that takes into account the process *context* in which the explanandum occurred, including relevant background knowledge, constraints, and goals. A situation-aware explanation can also help ensure that the explanation is relevant and informative to the user.

To understand this definition, let's unfold it to its ingredients:

**Context** - Finding an adequate explanation requires, in many cases, understanding the situational conditions in which specific decisions were made during process enactments. Frequently, explanations cannot be derived from "local" inference (i.e., current undergoing task or decision in a business process) but require reasoning about situation-wide contextual conditions relevant to the current step as derived from some actions in the past. Context aims to make the explanation richer, including knowledge elements that were originally implicit or part of the surrounding system, yet affected the choices that have been made during process execution.

A **sound** explanation is an explanation that is both **true** and **valid**. The former dimension of being true means that the explanation accurately and faithfully represents the domain and the occurrences in that domain, implying that reliance on its insights and actions derived from it can be reliably projected onto the environment it is anticipated to explain. The latter dimension of being valid means that, intrinsically, its reasoning mechanism is guaranteed to ensure its output was logically derived from its premises.

In addition to being sound, and to include contextual information, a situation-aware explanation should be causal. A **causal sound explanation** is an explanation that not only satisfies the criteria for sound explanation, but also provides an account of **why** the explanandum occurred. This includes the assurance that concluded explanations are entailed from the basic causal dependencies and temporal relationships that link between the different occurrences in the process domain. We consider two occurrences as being causally dependent when the occurrence of the former is explaining WHY the latter has occurred. More concretely, per our view, being related in a way that manipulating the timely occurrence of the former also entails some changes in the timely occurrence of the latter, and not vice versa. A temporal relationship is just an order relationship of the time axis between any two events such that we can determine which precedes the other in time (Fournier et al., 2023).

To illustrate these concepts, let's assume, for example, that the following is given:

*If loan application is accepted, send an email notification and archive the application within 24 hours after approval.*

- The Explanandum: *Why haven't I received the email about the acceptance of my application?*

According to the definition provided, the following holds:

- Valid explanation: You haven't got an email notification since your application was not accepted.
  - **(Counter example: Invalid explanation:** You haven't got an email notification since your application was accepted only two days ago).
- True explanation: You haven't got an email notification since your application was only approved 10 hours ago.
  - **(Counter example: Untrue explanation:** You haven't got an email notification since there is no record of your application in the system).
- **Causal sound explanation:** You haven't got an email notification since there was a delay in the email activity.
  - **(Counter example: Not causal sound explanation:** You haven't got an email notification since your application has not been archived).
- **SAX:** You haven't got an email notification since there was a delay in the email activity due to a strike of workers.
  - **(Counter example: Not SAX:** You haven't got an email notification due to error code 404).

### 3.3 Explanations of process execution results

The SAX4BPM library developed in the scope of the AI4Gov project includes services to support causal sound explanations, taking into account contextual information classified into three different types: completeness, soundness, and synthesis. These categories are complementary to infer SAX explanations (see section 3.2).

#### 3.3.1 Completeness of explanation

This class of services includes functionalities that can aid with extending the core execution data with additional context related information. Completeness of information usually enriches execution event logs with contextual knowledge, but it may also imply filtering out some of the non-relevant information or information that may already be redundant in the context of the situation that needs to be articulated about or even information that can be overwhelming to the user. To ensure alignment between the situation at hand and the explanation that is generated, the realization augments the conventional use of BP and XAI with CEP to achieve more adequate explanations of process execution outcomes, both in real-time and in retrospect (Amit et al., 2023). We combine here techniques that can both enrich the context of the information available for the generation of explanations, while also allowing for letting the users indicate their preferences with regard to aspects of the explanations they are more interested in, or that they may actually wish to give less attention to.

#### 3.3.2 Soundness of explanation

Our library has a unique focus on producing explanations about process conditions that are not only context aware (i.e., complete), but also explanations that are generated with diligent

attention to ensuring such explanations are causally sound. Ensuring such quality of explanations implies some fundamental characteristics, as described in section 3.2.

The services realised here are meant to serve as “guardrails”, providing a frame of reference to aid with the generation of explanations that adhere to the aspects of causal and time sequencing relationships.

### 3.3.3 Synthesis of explanation

This module caters the need to cohesively “glue together” and combine all knowledge sources that accommodate explanation adequacy per the aforementioned aspects of context, completeness, and soundness, blending everything into an explanation phrase that is also formed in a user-interpretable manner. There are certain interactions between the various sources of information that underly the generation of an explanation. A certain feature may be deemed highly important by a conventional XAI algorithm, while the process activity that produces it may be determined irrelevant from a causal perspective. Time precedence may suggest that the real reason for a certain outcome lies at a point in the past, where the explanation reflects a condition that was true then but may no longer be valid in the present. A certain outcome may logically stem from two prior incidents, but the causal chain reveals a dependency between them. This suggests that mentioning only one as the root cause in the explanation suffices. Including the other not only makes the explanation harder to interpret, but also introduces a redundancy. These kinds of challenges are the focus of this module, where recent developments with LLMs are the main instrumentation employed.

## 3.4 SAX4BPM library

The SAX4BPM library includes a set of services and capabilities to support the different aspects of SAX explanations. As aforementioned, the library will be released as open source at the end of the AI4Gov project. The library is implemented using Python 3.9 programming language.

### 3.4.1 SAX4BPM architecture

Figure 9 depicts the different modules comprising the SAX4BPM library. The main input constitutes timestamped process event logs, complemented by possible user preferences about features to be considered and other relevant situational enrichments. A typical process log conforms to the tuple  $\langle timestamp, case-id, activity-name, \{payload\} \rangle$  (see Figure 8), such that each process execution, also referred to as the case, is instantiated in the file as a set of timestamped entries that share the same case-id value, where each entry corresponds to the execution of a single activity in the process, indicated by its name. Corresponding to the execution of each activity, a series of additional payload attributes (a vector) could be recorded in relation to the execution of that activity. The different modules apply a set of methods and techniques on these event logs to provide causal sound explanations, i.e., SAX explanations (see Section 3.2).

In general, we classify the modules into modules that leverage existing and external services for the sake of SAX explanations and ones that are developed as part of the library. The first category

(light blue/grey colour) includes the Process discovery, XAI, and CEP enrich/filter modules. The second category (in blue) includes the Causal process discovery and the FM driven explanation/derivation. We briefly describe each of the modules next.

**Process discovery** – This module utilizes existing process discovery algorithms, such as Alpha and heuristics miner, to generate the process model out of the event logs.

**Causal process discovery** – This module transforms the event log into input to a causal discovery algorithm to produce a causal graph based on the execution times of the activities. We apply at this stage the LiNGAM algorithm. The method is generic and can be applied to any other causal discovery algorithm.

**XAI** – This module leverages contemporary XAI modules such as LIME and SHAP for the sake of preference importance ranking in the explanation.

**CEP enrich/filter** – This module employs external CEP engines (such as the PROTON<sup>5</sup> CEP open-source tool) to enrich the event log with additional contextual information or to reduce the size of the event log by filtering out non-relevant events.

**FM driven explanation/derivation** – This module receives as input XAI the outputs of the other modules and synthesize a sound and human-interpretable explanation leveraging existing LLM models. Foundation models (FM) are very large deep learning models that are pre-trained on massive datasets and adapted for multiple downstream tasks. Large language models (LLMs) are a subset of foundation models that can perform a variety of natural language processing (NLP) tasks.

---

<sup>5</sup> PROTON open source (Apache v2 licence): <https://github.com/ishkin/Proton>.



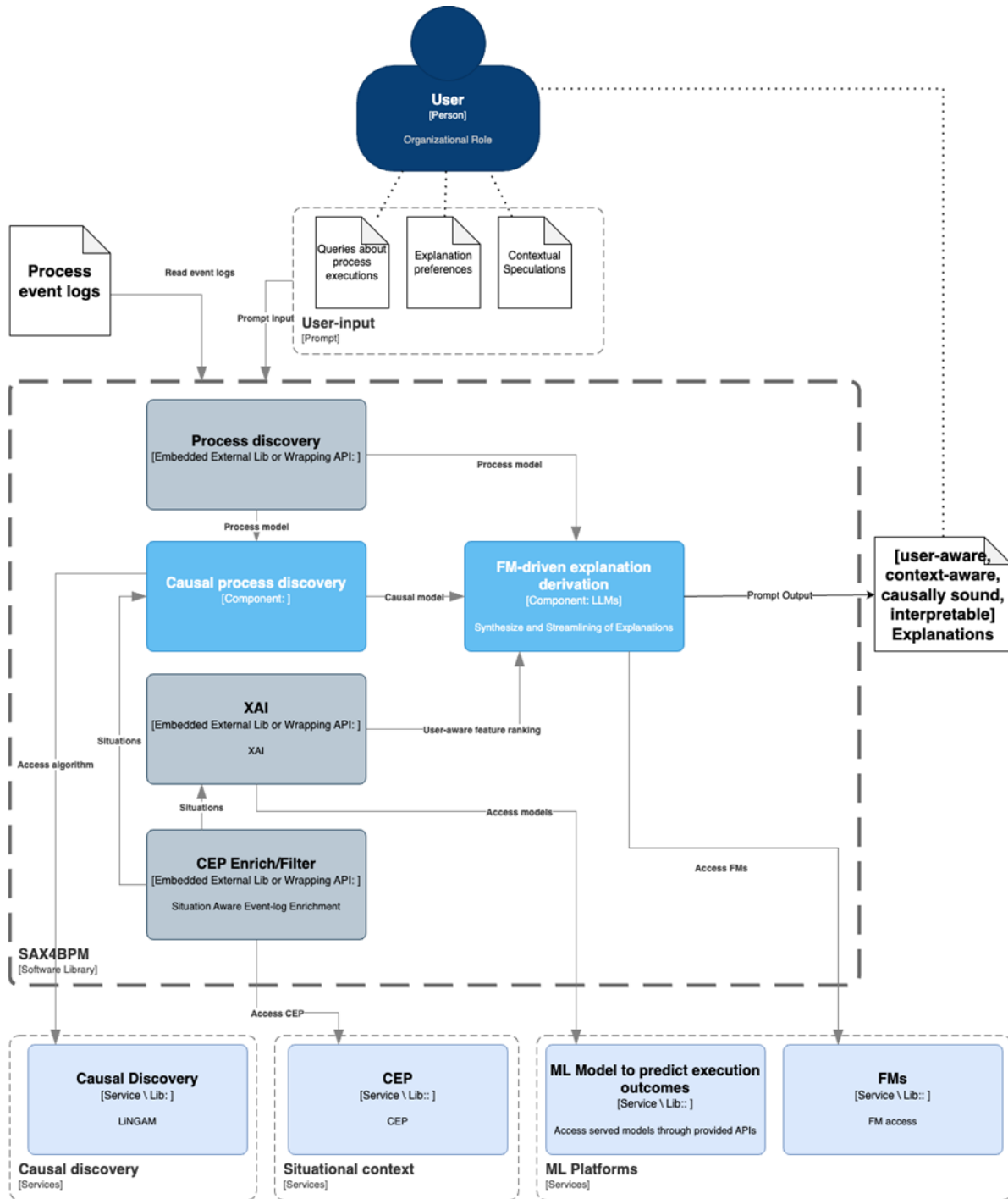


Figure 9: SAX4BPM architecture

### 3.4.2 SAX4BPM capabilities

The goal of the SAX4BPM library is to present a new toolset of services to support process explainability, enabling the generation of sound (valid and true), human-interpretable, situation-aware explanations for process-execution decisions and outcomes.

The library offers the following functionality:

#### **Data layer**

- Import of event log file (support of CSV, XES, MXML formats)
- Data object layer supporting process mining and causal discovery
  - Row event data
  - Transposed data (row per trace) for causal discovery

#### **Process discovery layer**

- Using PM open-source libraries wrappers such as PM4PY on data layer

#### **Causal discovery layer**

- Using causal discovery algorithms such as LINGAM on data layer

#### **XAI layer**

- Using XAI approaches (such as LIME (Ribeiro et al., 2016b) and SHAP (Lundberg et al., n.d.))

#### **Integration layer**

- Integrating process, causal and XAI views into higher-level explanations

### [3.4.3 SAX4BPM services](#)

The following services have been identified in the SAX4BPM to meet the requirements of completeness, soundness, and synthesis of SAX explanations. They are currently at different levels of maturity of development.

#### [3.4.3.1 Data manipulations](#)

Although these are not strictly related to fulfilling any concrete requirement of a SAX explanation, they support different aspects of manipulating initial data and its transformation to input to other services of the SAX4BPM library.

**Time4process** - This service aims at converting time series data into an event table form that can substitute an event log input. This service was not part of the original architecture (thus, it doesn't show in Figure 9), as the underlying assumption is that the input is always event logs. However, as the data received from the use cases does not come from processes, but it is sensor data (time series), we started to look into ways to still exploit this data. The transformation from timeseries into an event log relies most fundamentally on the ability to unite the raw timestamped data into segments, each of which reflects some higher-level step in the system's process. For example, a sensor temperature could provide daily measurements, that are then partitioned between low level temperatures (during 'Winter' period), and high-level temperatures (during 'Summer' period). Such segmentation may be accomplished in various ways, and could sometimes be automated, emerging from the density and distribution properties of the raw measurements themselves. The Time4process service attempts to employ a few semi-automated and fully

automated approaches to address the pragmatic need to translate raw timeseries data into a form of a higher-level state-transition graph as a basis for translating raw timeseries data into corresponding process event logs. From an explainability perspective, this translation lets us continue focusing on SAX, where explanations are about the need to populate clarifications about concrete conditions that occur during or after process executions. For example, in the context of the waste management pilot, we can explore explanations about predicted states related to the lifecycle process of individual waste bins e.g., why is a certain bin predicted to be in an overflowing state tomorrow morning? As may be a prediction of some ML that was trained on the bin's dataset, which is originally a timeseries data.

**Log2Tabular** – This service transforms the log event input file into a tabular form that can be then used by the other services in the library. As aforementioned, an event log constitutes of a set of tuples of the form *<timestamp, case-id, activity-name, {[payload]}>* (see Figure 8). This service transposes the event log table into a table in which each row consists of the activities or tasks corresponding to the same trace.

#### 3.4.3.2 *Completeness of explanations*

**ContextEnrichment** - One form of integrating contextual information is by running a CEP engine that derives situational events that can then be interwoven with the original (input) event log to produce the explanations. To this end, the SAX4BPM library enables the wrapping of external CEP engines and their execution both in real-time (during process executions running alongside the business process engine) or in retrospect. In the former, the derived events from the CEP engine are interwoven with the events in the running process instances, while in the latter, the enriched event log is injected into the explainer applying historical data as input (the original event log file). So far, we applied this service in retrospect, using the PROTON CEP engine and feeding the result as an input for the SHAP (Lundberg et al., n.d.) explainer. Our initial tests **Error! Reference source not found.** show that temporal contextual information can be leveraged to improve the adequacy of explanations given for process execution instances. The effect of the enrichment manifests itself not just in properly adjusting the importance of factors that correctly correspond to the outcome, but also in promoting the accuracy of the surrogate ML model.

**X4User** – The goal of this service is to support user preferences in the derived explanations. At this stage the implementation of this service has not started.

#### 3.4.3.3 *Soundness of explanations*

**Causal4Process** – Unraveling the causal relationships among the execution of process activities is a crucial element in predicting the consequences of process interventions and making informed decisions regarding process improvements. Our previous work (Fournier et al., 2023) demonstrated that relying only on time precedence between activities in a business process does not fully reflect the cause-effect dependencies among the tasks and a more fundamental analysis of causal relationships among the tasks is required. To this end, the Causal4Process service applies causal discovery techniques to event logs to produce a causal graph. We currently rely on the LiNGAM (Shimizu, 2022) algorithm adapted to discover causal time dependencies among tasks or activities.

**X4Process** – This service supports process constraints adherence. The idea is that an explainer needs to understand the basic constraints underlying any process by simply knowing the time precedence of the execution activities. In simple words, the explainer needs to know the model of the process, otherwise it may mistakenly infer explanations that don't conform to the model and, therefore, cannot hold in reality (Amit et al., 2022).

#### 3.4.3.4 *Synthesis of explanations*

**NLP4X** – the overarching goal of the SAX4BPM library is to produce SAX explanations that are sound and contextual on the one hand, but are also easy to understand by human users. To this end, we leverage existing LLMs (e.g., ChatGPT<sup>6</sup> from OpenAI). The NLP4X service receives as input any combination of process model, causal model, and XAI features importance, and “blends” these narratives to produce a text explanation which reflects the combined narratives. Our preliminary tests in this area (Fahland et al., 2023) show this direction as very promising. We will continue to investigate this area and report our findings in the next version of this deliverable as well in future publications.

#### 3.4.4 *Illustrative example 1: Parking tickets*

At the time of writing this report, the development of the library and our tests are carried out accompanied by generated data inspired by one of the project use cases instead of real data from the pilots. This is because real data from pilots is still not available in a way we can leverage. To this end, we generate data according to the parking tickets illustrative example using BIMP open-source log simulation tool<sup>7</sup>.

The illustrative example used (inspired by one of the project's scenarios) is parking tickets. In this process (see Figure 10), a parking ticket is given when a vehicle is parking in a prohibited lot, and it does not possess a disabled permit. In this case, two types of fines can be given depending on whether the parking place is a hazardous place (e.g., the vehicle is parking on a sidewalk or on a crosswalk) or not. In the case of a hazardous place, an extended fine is submitted, and a tow truck is called. Note that the coming of the tow truck is always longer than the time to submit the extended fine.

---

<sup>6</sup> <https://chat.openai.com>

<sup>7</sup> <https://bimp.cs.ut.ee/simulator/>

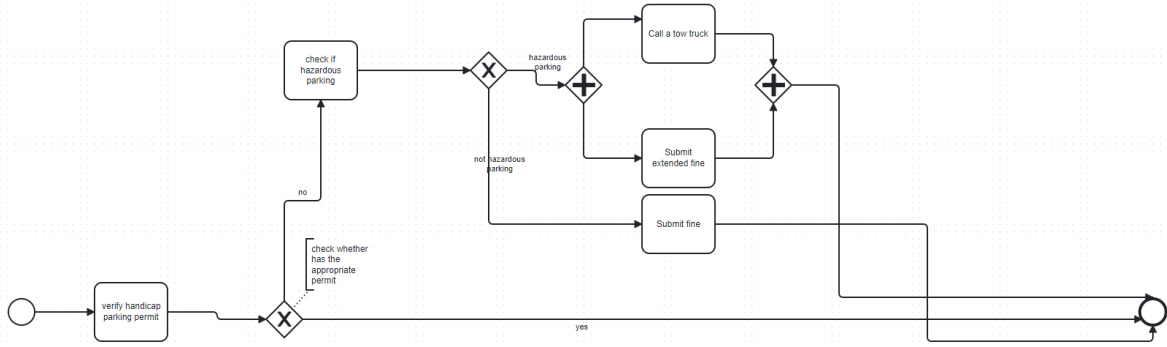


Figure 10: Parking scenario

We demonstrate the capabilities developed so far in the library through the screenshots below using the Parking scenario and recorded in a demo, available at: [AI4Gov demonstrators](#).

<b>Steps demonstrated</b>	<ol style="list-style-type: none"> <li>1. Import log file</li> <li>2. Discover process model</li> <li>3. Get process variants and get a particular variant (optional)</li> <li>4. Discover causal model (for the selected variant)</li> <li>5. Apply XAI</li> <li>6. Apply synthesis of causal, XAI, and model</li> </ol>
<b>Services used</b>	<ul style="list-style-type: none"> <li>• Log2Tabular</li> <li>• Causal4Process</li> <li>• X4Process</li> <li>• NLP4X</li> </ul>

Table 1 shows the steps performed by the services used in this example and described henceforth.

**The question we are trying to answer:** *How can we expedite the processing of fines for cars that are parked within hazardous locations?*

<b>Steps demonstrated</b>	<ol style="list-style-type: none"> <li>7. Import log file</li> <li>8. Discover process model</li> <li>9. Get process variants and get a particular variant (optional)</li> <li>10. Discover causal model (for the selected variant)</li> <li>11. Apply XAI</li> <li>12. Apply synthesis of causal, XAI, and model</li> </ol>
<b>Services used</b>	<ul style="list-style-type: none"> <li>• Log2Tabular</li> <li>• Causal4Process</li> <li>• X4Process</li> <li>• NLP4X</li> </ul>

Table 1: Steps and services used in the parking scenario

### 3.4.4.1 Import log file

The event log corresponding to the scenario is imported for further analysis (Figure 11).

```

fileName = "C:\Data\Automation\SAX\Projects\AI4Gov\Deliverables\04_1\parking_fines bpmn and logs\parking_fines1.mxml"
dataframe = pm.import_mxml(fileName,timestamp_format="%Y-%m-%dT%H:%M:%S.%f%z")
[2] ✓ 0.6s

dataframe.getData()
[3] ✓ 0.0s
...

```

	case:conceptname	concept:name	lifecycle:transition	time:timestamp	org:resource	Attr_resourceId
0	192	EVENT 1 START	assign	2023-11-15 09:31:17.359000+00:00	None	NaN
1	192	EVENT 1 START	start	2023-11-15 09:31:17.359000+00:00	None	NaN
2	192	EVENT 1 START	complete	2023-11-15 09:31:17.359000+00:00	None	NaN
3	192	verify handicap parking permit	assign	2023-11-15 09:31:17.359000+00:00	None	NaN
4	192	verify handicap parking permit	start	2023-11-15 09:31:17.359000+00:00	verify disabled certificate-000001	qbp_bbcf6f67-6fef-565e-ac29-4bf3deda54a0
...	...	...	...	...	...	...
15100	686	verify handicap parking permit	start	2023-11-22 12:04:50.337000+00:00	verify disabled certificate-000001	qbp_bbcf6f67-6fef-565e-ac29-4bf3deda54a0
15101	686	verify handicap parking permit	complete	2023-11-22 12:04:52.483000+00:00	verify disabled certificate-000001	qbp_bbcf6f67-6fef-565e-ac29-4bf3deda54a0
15102	686	EVENT 3 END	assign	2023-11-22 12:04:52.483000+00:00	None	NaN
15103	686	EVENT 3 END	start	2023-11-22 12:04:52.483000+00:00	None	NaN
15104	686	EVENT 3 END	complete	2023-11-22 12:04:52.483000+00:00	None	NaN

15105 rows x 6 columns

Figure 11: Import log file

### 3.4.4.2 Discover process model

The next step is applying process discovery techniques to discover a process model out of the input event log. We currently apply the Heuristics Miner algorithm (Mannhardt et al., n.d.), but any process mining algorithm can work as well. Figure 12 shows the process model discovered in our scenario of parking tickets.

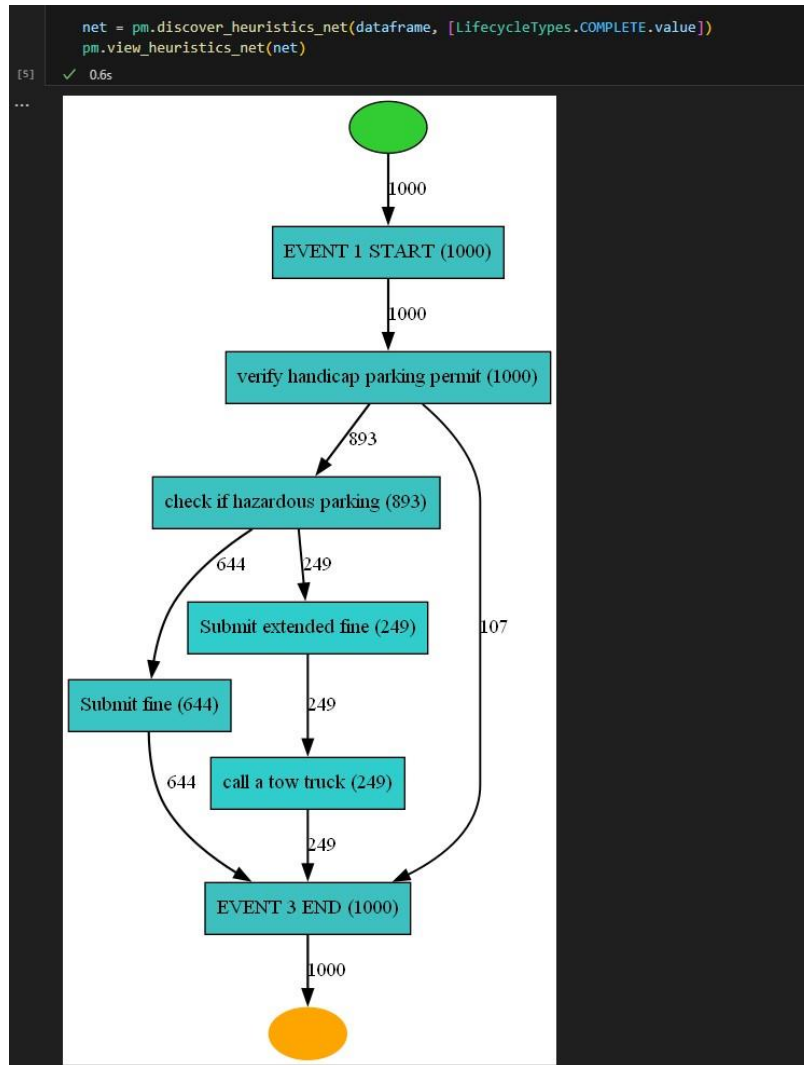


Figure 12: Process model discovery for the parking tickets scenario

### 3.4.4.3 Get process variants and get a particular variant

The next step involves the scoping the problem in hand, that is, focusing on the relevant tasks in the process that are related to the question in hand (in our case: *how to expedite the processing of fines for cars that are parked within hazardous locations*). To this end, we split the model into the different possible variants and choose from them the ones that we would like to closely examine for further analysis of causal and XAI perspectives.

A *process variant* refers to a specific sequence or path of activities in the process. Determining all process variants directly from an event log is essentially enumerating all unique traces in the log. For each group (trace), the sequence of activity names is extracted, and any duplicates (i.e., traces that represent the same sequence of activities) are identified and consolidated. This step yields all the unique traces, which are effectively the process variants. In the parking tickets scenario, we are interested in explaining the reasons for possible delays in the case of fines within

hazardous locations; therefore we are only interested in the paths in the process model that correspond to “hazardous locations”.

Figure 13 shows the code that extracts all variants from the parking tickets process model, while Figure 14 selects a particular variant, shown in Figure 15.

```
dataframe.getVariants()
✓ 0.2s
{'EVENT 1 START,verify handicap parking permit,check if hazardous parking,Submit extended fine,call a tow truck,EVENT 3 END,': 249,
 'EVENT 1 START,verify handicap parking permit,check if hazardous parking,Submit fine,EVENT 3 END,': 644,
 'EVENT 1 START,verify handicap parking permit,EVENT 3 END,': 107}
```

Figure 13: Get variants from process model

```
variant = dataframe.getVariant('EVENT 1 START,verify handicap parking permit,check if hazardous parking,Submit extended fine,call a tow truck,EVENT 3 END,')
variant.getData()
✓ 0.0s
```

	case:conceptname	concept:name	lifecycle:transition	time:timestamp	org:resource	Attr_resourceId
60	196	EVENT 1 START	assign	2023-11-15 09:54:36.113000+00:00	None	NaN
61	196	EVENT 1 START	start	2023-11-15 09:54:36.113000+00:00	None	NaN
62	196	EVENT 1 START	complete	2023-11-15 09:54:36.113000+00:00	None	NaN
63	196	verify handicap parking permit	assign	2023-11-15 09:54:36.113000+00:00	None	NaN
64	196	verify handicap parking permit	start	2023-11-15 09:54:36.113000+00:00	verify disabled certificate-000001	qbp_bbcf6f67-6fef-565e-ac29-4bf3deda54a0
...	...	...	...	...	...	...
15091	506	Submit extended fine	complete	2023-11-20 10:37:37.702000+00:00	Submit extended fine-000001	qbp_147e98d0-2c09-8a7f-2345-824792c66f28
15092	506	call a tow truck	complete	2023-11-20 10:37:51.077000+00:00	call a tow truck-000001	qbp_9d4180b3-5cd6-6c7e-bec6-af6803936535
15093	506	EVENT 3 END	assign	2023-11-20 10:37:51.077000+00:00	None	NaN
15094	506	EVENT 3 END	start	2023-11-20 10:37:51.077000+00:00	None	NaN
15095	506	EVENT 3 END	complete	2023-11-20 10:37:51.077000+00:00	None	NaN

Figure 14: Get a particular variant



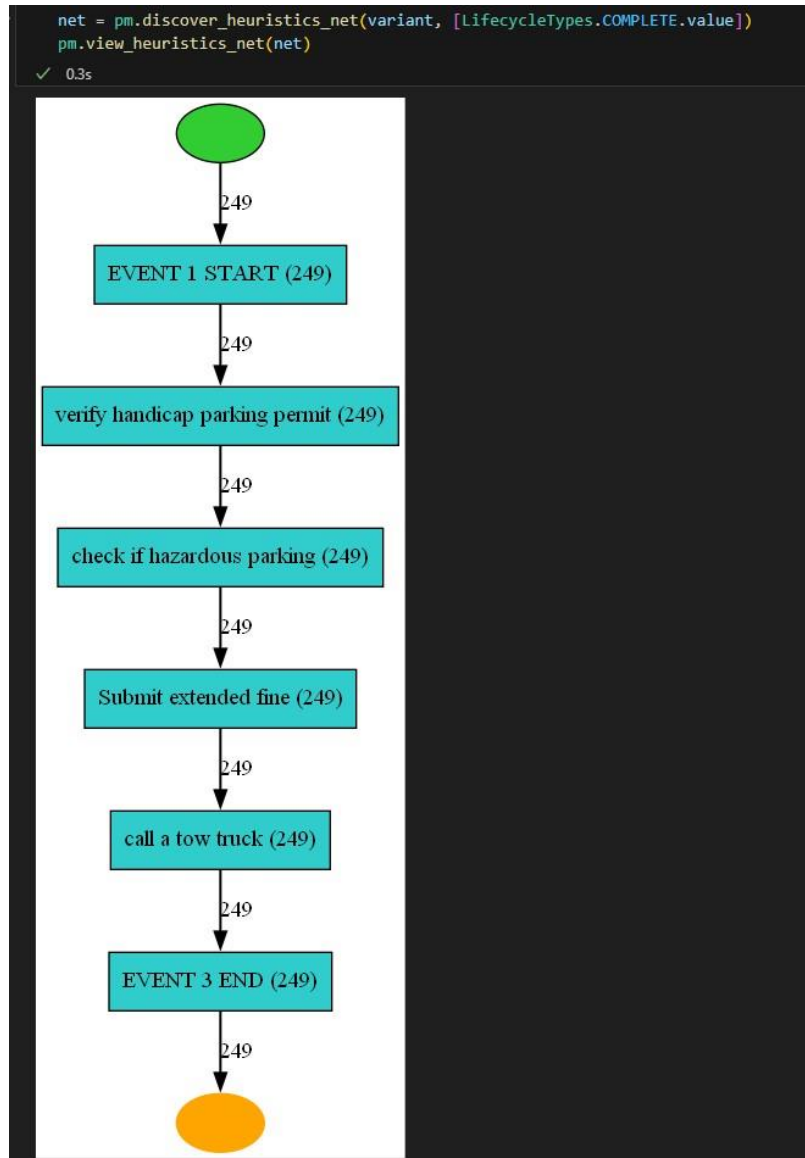


Figure 15: Process model for the variant

#### 3.4.4.4 Discover causal model for the variant

The next step is applying LiNGAM on the selected variant. The code snippet along with the resulting causal graph are shown in Figure 16. As it can be seen, while the *submit extended fine* task precedes the *call a tow truck* task in the process model (Figure 15), these are not causally related as demonstrated in Figure 16. The fact that calling a truck is slower than submitting an extended fine (reflected in the times of these tasks in the initial event log) is mined as two consecutive tasks in the discovered process model, where in reality, these two tasks are not causally dependent and can be carried out concurrently. This is extremely important in terms of the time expedition of the process as it means that if we would like to end the process earlier, there is no point in attempting to shorten the time of the *submit extended fine* task.

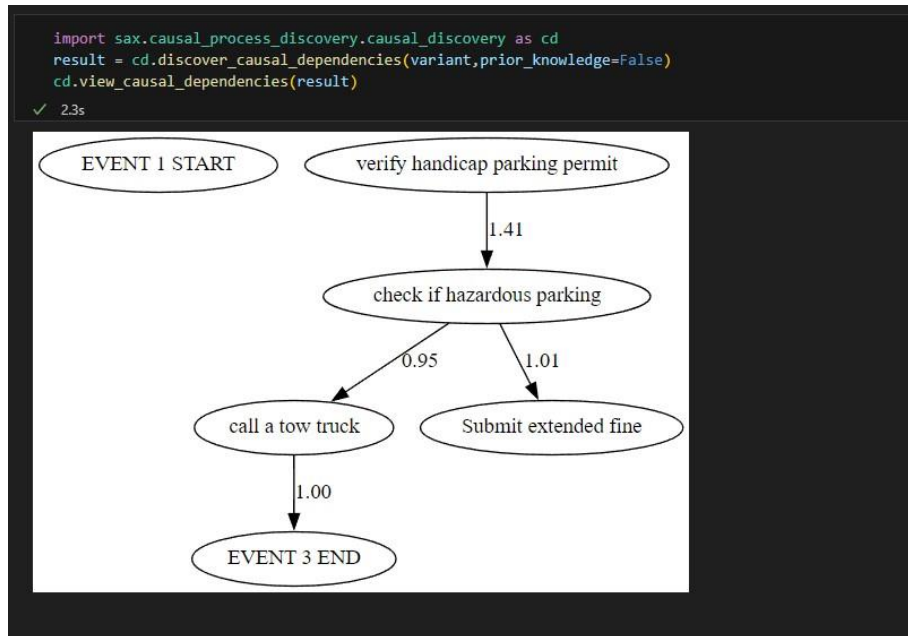


Figure 16: Causal graph discovery for the variant

### 3.4.4.5 Apply XAI

This step includes applying XAI techniques to the process task attributes or features. Let's assume that in our parking tickets scenario the features given are the driver's credit, the reasons for hazardous location (e.g., sidewalk and crosswalk), the region in the city, and the towing company. Let's also assume that the following graph is the result of this step in our example (Figure 17). As can be seen, the reason for hazardous location is the most important factor helping in understanding how to expedite the process, or in other words, what causes delays in the process.

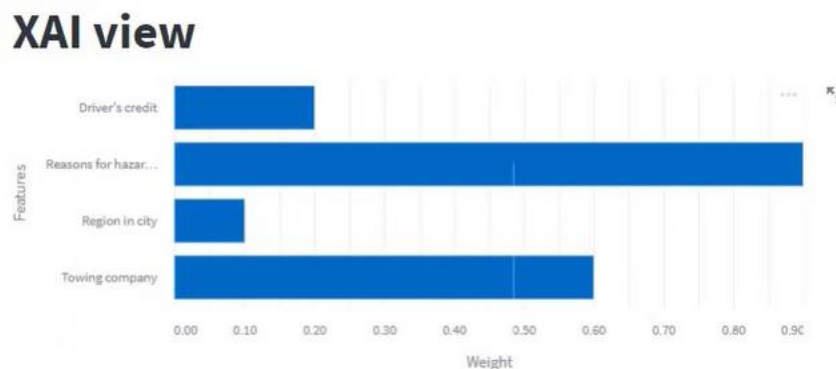


Figure 17: XAI graph for the parking tickets scenario

### 3.4.4.6 Apply synthesis of causal, XAI, and process models

Once we have the process, causal, and XAI views, we can proceed to the last step, which is applying LLM to get a SAX explanation. In our case, we used ChatGPT 3.5. In this step, we can select any combination of our three views to be synthesized by the LLM (Figure 18).

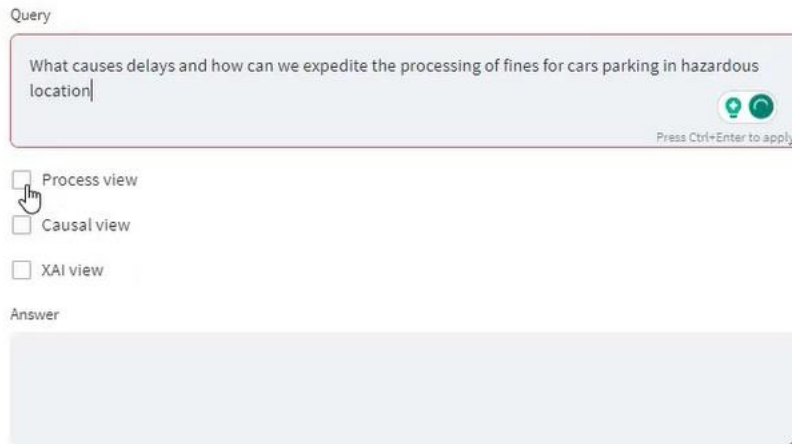


Figure 18: SAX explanation by applying LLM screenshot

For example, by selecting process only view, we get the following answer from the LLM: *“based solely on the process view, delays in processing may be attributed to any of the stages within the corresponding process variant: checking for disability certification, identification of hazardous factors, filling out hazardous circumstances, and calling for towing company”*.

The blending of causal and process views yields the following answer: *“based on the process and causal views, delays in processing may be attributed to any of the stages within the corresponding process variant: checking for disability certification, identification of hazardous factors, and calling for towing company”*. Notice that *filling out hazardous circumstances* is **not** part of the answer anymore, as it is not part of the causal path that influences the duration of the variant.

Finally, selecting all three views yields the same answer as with the process and the causal views with the addition of the *towing company* as the most important factor explaining the delays (Figure 19), although this feature was second in importance in the XAI stand-alone analysis (Figure 17).

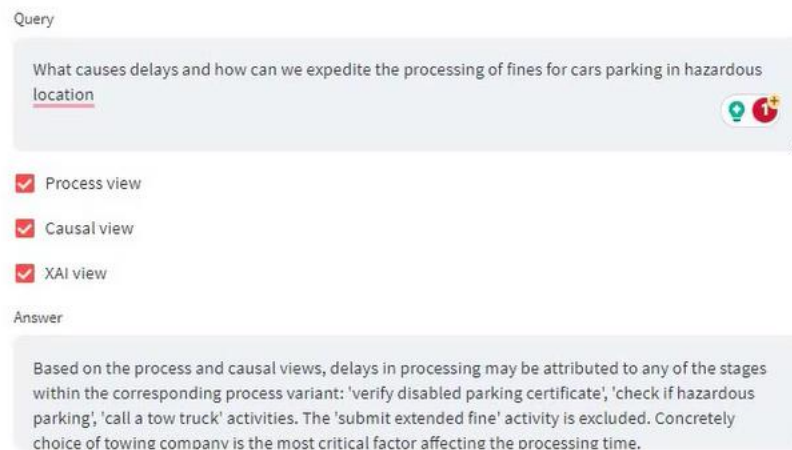


Figure 19: SAX explanation by applying selecting process, causa, and XAI views screenshot

### 3.4.5 Illustrative example 2: Waste management use case

While the previous illustrative example highlights a possible use of the SAX4BPM library where the initial input is assumed to be a conventional **process log**, many real-world systems record their surrounding conditions via a different type of data recording, one that is typically referred to as **timeseries data**. In this case, we need to transform the timeseries data into some form of a process event log that can then be input to the SAX4BPM library.

In this type of scenarios, the density of the data stems from the sampling frequency of the sensors that are used to measure the different phenomena in the environment, while there is no inherent a-priori partitioning of the data into higher level, domain meaningful events. Time series data in its simplest *single-variate* form captures a single recording value for each data point. A more advanced form is a *multivariate* timeseries in which each timestamp captures a set (vector) of recorded values  $v_1 \dots v_n$ .

Our second illustrative example revolves around the management of bins waste. Sensors located in bins are monitored every hour and the level of waste is measured for each bin (Figure 20). Our idea is to transform these measurements into higher forms of events that can then be used as event log input to assist in answering questions such as, “*will my bin waste be picked up tomorrow*”? As noted before, this service has not been anticipated during the proposal writing, but was nevertheless added to the library as some of the use cases store timeseries data. Table 2 shows the steps performed by the services used in this example.

<b>Steps demonstrated</b>	<ol style="list-style-type: none"><li>1. Loading of time-series data.</li><li>2. Data segmentation.</li><li>3. Export data to process-log.</li></ol>
<b>Services used</b>	Time4Process

Table 2: Steps and services used in the waste management example.

#### 3.4.5.1 Loading of time-series data

The first step in this illustrative example is the loading of a time-series file as an input, where the conventional schema for such a file is the tuple <timestamp, value>, where timestamp designates the time at which each individual measurement was recorded, and the value corresponds to the measurement. Each waste-bin file is a timeseries data type. Concretely, each row in the data corresponds to an hourly recording of the percentage of the level of waste inside each individual waste bin. In this particular example, the level of waste in each waste bin is recorded every single hour, and all daily measurements are stored in a single file. Such a file does not conform to the form of a conventional event log (see Figure 20). Hence the next step is to transform it into a conventional process event log.

BinId	Timestamp	LevelPercent
866349043901200	1/9/2023 0:20:06	37
866349043901200	1/9/2023 1:20:06	51
866349043901200	1/9/2023 2:20:06	51
866349043901200	1/9/2023 3:20:06	52
866349043901200	1/9/2023 4:20:06	50
866349043901200	1/9/2023 5:20:06	50
866349043901200	1/9/2023 6:20:06	50

Figure 20: snippet of the waste management time series data input

### 3.4.5.2 Data segmentation

Given a timeseries data as an input, various techniques could be employed in order to transform it into a data form that corresponds to the form of a process event log.

The `time4process` service is developed in the AI4Gov project to enable transforming timeseries data into a more conventional process event log. This transformation employs different algorithmic approaches for the segmentation of the raw timeseries data into some higher-level partitioning, marked by a finite set of labels, that would usually correspond to some domain meaningful phases along which the system traverses. Such partitioning could sometimes arise naturally from the timely distribution and clustering of the underlying data, while in certain cases, it may also require some understanding of the source domain. For example, observing the fill levels of some individual waste bin may reveal that during most weekdays, the level of the bin does not exceed 30% of its capacity, while during weekend days it goes above this threshold, getting close to its full capacity on Sundays. Such an observation may allow projecting from each individual measurement to any one of three possible labels: *low*, *medium*, and *full*, and then condensing the original data into these segmented states, retaining only the timestamps that correspond to altering between these states.

The following approaches are being explored as alternative options for the realization of the `time4process` service.

**Rule-based** – By employing domain knowledge, the raw data is mapped to a label that corresponds to a process state. For example, the fill level of the waste bin is mapped to any one of four labels {waste\_added, waste\_removed, bin\_emptied, and bin\_overflowing} conforming to the following rules (Table 3):

<pre>BinId,Timestamp,LevelPercent,EventType [...] 866349043957400,2023-09-06 22:30:06,86,waste_added 866349043957400,2023-09-06 23:30:06,97,waste_added 866349043957400,2023-09-06 23:31:06,97,bin_overflowing 866349043957400,2023-09-07 00:31:06,97,bin_overflowing 866349043957400,2023-09-07 01:31:06,97,bin_overflowing 866349043957400,2023-09-07 02:31:06,97,bin_overflowing 866349043957400,2023-09-07 03:31:06,97,bin_overflowing 866349043957400,2023-09-07 04:31:06,97,bin_overflowing 866349043957400,2023-09-07 05:30:06,83,waste_removed 866349043957400,2023-09-07 07:30:06,0,waste_removed 866349043957400,2023-09-07 07:31:06,0,bin_emptied 866349043957400,2023-09-07 10:30:06,14,waste_added</pre>	<p>(1) If: <math>waste\_level(t_i) = 0</math> AND <math>(waste\_level(t_i) - waste\_level(t_{i-1})) &lt; 0</math> then: “bin emptied”</p> <p>(2) If: <math>waste\_level(t_i) \geq 95</math> then: “bin overflowing”</p> <p>(3) If: <math>(waste\_level(t_i) - waste\_level(t_{i-1})) &gt; 0</math> then: “waste added”</p> <p>(4) If: <math>(waste\_level(t_i) - waste\_level(t_{i-1})) &lt; 0</math> then: “waste removed”</p>
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 3: Data segmentation rule-based approach for the waste management example

**Model-fit based** – another alternative we consider integrating with is the StreamStory (Stopar et al., 2019) tool by JSI that helps reveal a form of a hierarchical Markov chain based on multivariate time series data. The tool employs a methodological approach that uses clustering to construct the states and represents temporal dynamics as transitions between the states using a Markov chain. The result, as illustrated in Figure 21, is an interactive UI that allows exploration of the hierarchical set of models as different scales, each denoting a labelled graph with transitions that show the probability of altering between the states. This helps converging on the model that best reflects the process model associated with some long-term recurrent behaviour the emerges from the underlying raw data. Once the model is determined, it can be used not only to make predictions about the data, but for our purpose, to produce a process event log.

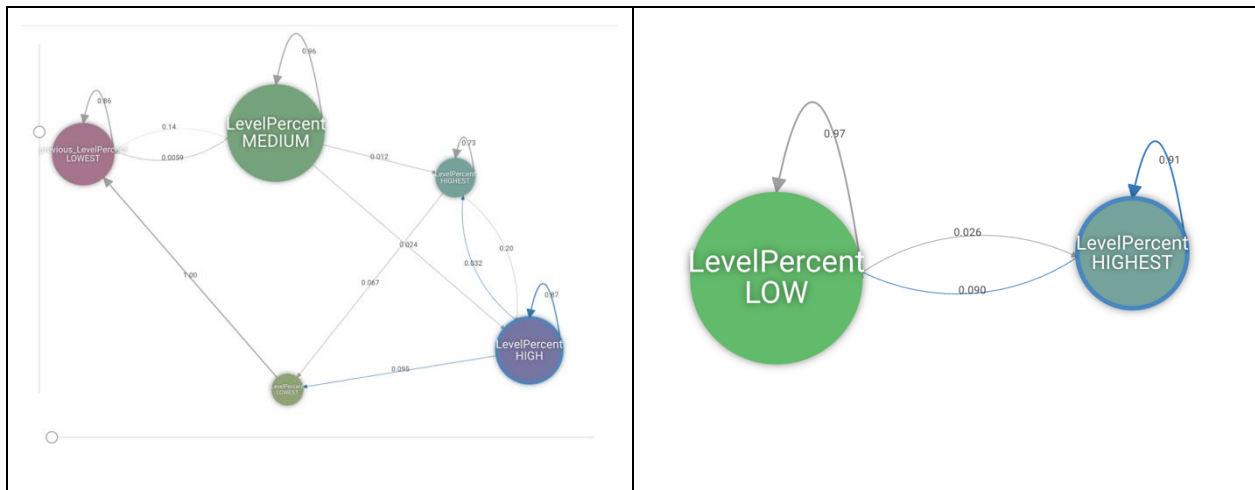


Figure 21: Data segmentation model-fit based approach for the waste management example

**Functional-fit based** – another possible realization is segmenting the data based on seasonality and distribution. For example, employing Gaussian Process Regression can partition the data based on certain periodic variables (e.g., week of day, part of day, region), while also identifying the Gaussian distribution of the points within each segment. This may also elicit a cyclic process in which each execution repetition is bound by different statistical limits (Figure 22).

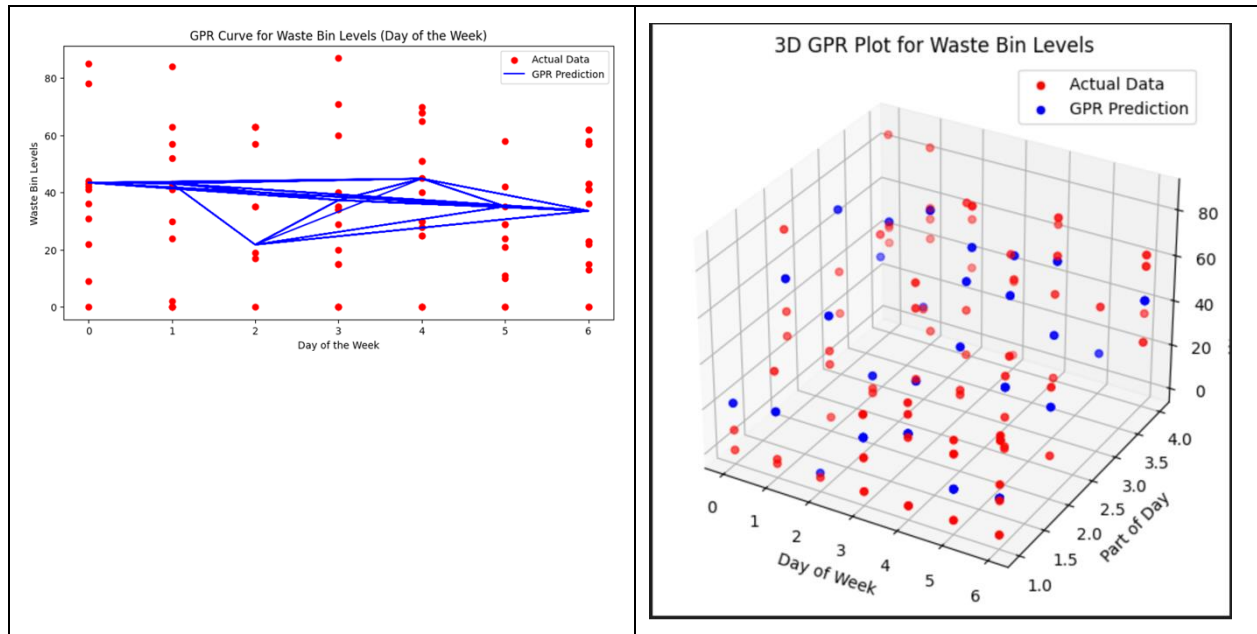


Figure 22: Data segmentation functional-fit based approach for the waste management example

#### 3.4.5.3 Export data to process log

The techniques illustrated above help associate each individual data point in the time series dataset with a corresponding label. This is a basis for then marking the points in time that correspond to transitioning between one segment and another (e.g., bin level high --> bin level full). Any such transitioning can be determined as a type of event, with the transition time reflecting the occurrence time of the event, giving the ability to transform from the original timeseries into a conventional process log format. Each such event could then be complemented by a variety of additional computed and other contextual occurrences as part of its payload, capturing different properties about each segment (e.g., average level, fill level of some other nearby bin). The eventual output is generated as a conventional process log, providing input for the subsequent steps, as illustrated in the other example.

### 3.5 Next steps with T4.2

This report describes the research around SAX, the challenges, the main concepts, as well as the current status of the SAX4BPM library that is our major asset in the project. The SAX services will

be released as open source at the end of the project. A movie demonstrating current features based on the parking tickets illustrative example can be found at: [AI4Gov demonstrators](#).

This deliverable relates to the work carried out during the first year of the project. Naturally, next steps include the continuation of our developing efforts in the implementation of the SAX4BPM library and the employment of the library in some of the use cases in the project. These efforts will be reported and demonstrated in the next version of this deliverable (D4.2 – Trustworthy, Explainable, and unbiased AI V2) at month 24.



## 4 Policy-Oriented AI and NLP Algorithms

In this section of the deliverable, the “Policy-Oriented AI and NLP algorithms” component is thoroughly specified, whilst extensive information about the source code and the corresponding user manual are provided. As also mentioned in D2.3 - Reference Architecture and Integration of AI4Gov Platform V1, the “Policy-Oriented AI and NLP algorithms” component is being developed in the context of T4.3 – “Improve Citizen Engagement and Trust utilising NLP” and consists of two (2) sub-components, namely *Policy-Oriented Analytics and AI Algorithms* and *Adaptive Analytics Framework*. The *Policy-Oriented Analytics and AI Algorithms* aims to develop several NLP algorithms in order to analyse large volumes of text data and also assist the respective AI experts. This particular subcomponent consists of the following mechanisms:

- Question Answering Service
- Time Series Analyser

The scope of the *Adaptive Analytics Framework* subcomponent is to develop the needed ML models for performing predictive analytics and optimised resource allocation to satisfy the needs of the pilots and assist policy makers.

All the above should be executed in an efficient manner, utilising the least possible number of resources.

The following sections are organized as follows. First a state-of-the-art analysis is conducted in terms of the AI algorithms used for policy making, as well as, current advancements in NLP, focusing on question-answering (QA) systems and approaches. Then, for both the subcomponents mentioned above, the architecture and internal workflow are thoroughly described. Moreover, for every subcomponent, the baseline technologies used are provided, whilst the availability and the key points for the source code are also examined. Last but not least, a user guide for each of the two (2) subcomponents is provided, focusing on the installation and use.

### 4.1 AI Algorithms for Policy Making

There have been several advancements regarding the fields of policy-oriented AI. To begin with, AI has been widely utilized in a variety of sectors in order to assist the corresponding policy makers. For instance, in the healthcare domain and during the recent COVID-19 pandemic, a variety of approaches (Rahman et al., 2021) were introduced in order to help public health experts form specific policies, protocols and interventions that could potentially protect the citizens from the pandemic. In those approaches, a variety of data analytics took place and a vast number of ML algorithms were developed (Mavrogiorgou et al., 2021), in order to monitor the progress of the pandemic and predict certain outcomes, thus enabling policy makers to make the right decisions. Towards the same direction, i.e., assisting in data-driven decision making and, thus, creating more useful policies, AI is also being utilized in a variety of sectors such as e-Governance (Alexopoulos et al., 2019), transportation (Ağbulut et al., 2022) and the environment (Hettinga et al., 2023).

With regards to the AI4Gov’s pilots and the corresponding sectors, which are water management, sustainability and tourism, there have also been some advancements in the literature. More specifically, regarding water management, the authors in (Gino Sophia et al., 2020) proposed a genetic algorithm based on a fitness function that effectively manages water distribution with regression of 98%. In (Bhardwaj et al., 2022), the authors proposed a machine learning–based framework for the assessment of water quality by utilizing a variety of ML-based algorithms such as logistic regression, naïve Bayes, ensemble-based approaches such as Random Forest and XGBoost to classify the data in appropriate classes and predict turbidity in a water sample. As for sustainability, a variety of approaches have been proposed, such as the ones presented in (Shafiq et al., 2020) and aim to the classification of Sustainable Smart Cities (SSC) network traffic. Those approaches are based on several ML techniques; most of them are more traditional while very few approaches utilize more advanced techniques such as reinforcement learning, ensemble learning and genetic algorithms. Regarding tourism, several approaches have been developed that aim to address different use case scenarios, mostly for predicting tourism flows and performing sentiment analysis such as the ones presented in (Xie et al., 2021), (Li et al., 2021) and (Puh et al., 2023).

## 4.2 Current Advancements in NLP and QA

As for NLP, there have been tremendous advancements, since the high availability of resources and data have provided new opportunities, including Large Language Models (LLMs) that are capable of performing a vast variety of tasks, such as question-answering (QA), in different domains. QA is generally divided into two (2) subcategories named Closed Domain Question Answering (CDQA) and Open Domain Question Answering (ODQA). Based on the type of answers that are provided, and more specifically, whether the answer is extracted from the text, or it is generated based on the context of the text, QA can be divided into Extractive Question Answering (EQA) and Abstractive Question Answering (AQA).

CDQA, as its name implies, refers to QA approaches that answer questions of a specific domain. This is achieved through having a domain-specific knowledge base (e.g., a knowledge base that consists of articles about the environment) (Cortes et al., 2022). ODQA is about approaches that aim to answer questions regardless of the domain that they refer to. This is achieved by having a “global” knowledge base such as the whole Wikipedia (Zhong et al., 2022). ChatGPT and Google Bard<sup>8</sup> could be identified as indicative examples of widely used ODQA tools. Depending on the user requirements and the volume of the knowledge base, both CDQA and ODQA systems may use EQA or AQA (Yoon et al., 2022).

Of course, answers provided by ODQA tools are known to be inaccurate, since the knowledge itself could be inaccurate (i.e., the Internet) and could also promote, among others, prejudices, fake information and hate speech, thus opposing a threat to the users (Ray et al., 2023). CDQA tools could also have the same disadvantages, but at a very lower scale, since the knowledge base

---

<sup>8</sup> <https://bard.google.com>

is domain-specific, smaller in size and probably not utilizing data just from the Internet (Antoniou et al., 2022). To this end, both the *Adaptive Analytics Framework* and the *Policy-Oriented Analytics and AI Algorithms* seek to implement appropriate AI techniques and NLP algorithms that satisfy the needs of the AI4Gov pilots whilst taking into consideration any potential challenges and ensuring the development of unbiased ML models.

### 4.3 Multilingual NLP

Our multilingual and multicultural societies express the need for the introduction of cross lingual and language-agnostic solutions. At the same time, the tremendous growth in the popularity and usage of social media, such as Twitter, and of applications that support citizens in their interactions with public authorities and services, as well as gather their feedback, has resulted in an immense increase in user-generated data, as mainly represented by the corresponding texts in users' posts and complaints. However, the analysis of these specific data and the extraction of actionable knowledge and added value out of them is a challenging task due to the domain diversity and the high multilingualism that characterizes these data. Hence, leveraging the potentials that can be derived from the cross lingual analysis of them is crucial in the modern policy-making domain. In that direction, researchers are constantly trying to develop the most comprehensive multilingual systems.

Several AI research teams from major pioneers, such as Google AI and Facebook AI Research, have introduced multilingual tools, corpora and sentence encoding models that are able to cover any language, thus overcoming the limitations imposed by the lack of labelled data in all languages (Zorrilla et al., 2022). The successful design and implementation of multilingual solutions rely on incorporating multilingual sentence embeddings and employing multilingual classifiers, both built upon pre-trained models and the principles of transfer learning (Artetxe et al., 2019). Current methodologies extend beyond word embeddings to enhance multilingual natural language processing (NLP) and capture deeper semantic meaning by utilizing embeddings for higher-level structures, such as sentences or even paragraphs. Existing approaches for generating such embeddings, like LASER (Artetxe et al., 2019) or MUSE (Lample et al., 2017), rely on parallel data, mapping a sentence from one language directly to another language to encourage consistency between the sentence embeddings. In addition, the authors in (Feng et al., 2020) present a multilingual BERT embedding model, called LaBSE, that produces language-agnostic cross-lingual sentence embeddings for 109 languages that is highly effective even on low-resource languages for which there is no data available during training.

One of the latest milestones in the NLP field is the introduction of BERT that enables transfer learning with large language models reaching the state-of-the-art for a great number of NLP tasks and applications (Devlin et al., 2018). In this context, several research works have proposed multilingual models based on the utilization of BERT for a wide range of cross-lingual transfer tasks. More specifically, advances in multilingual language models such as multilingual BERT (mBERT) (Pires et al., 2019) and XLM-RoBERTa that are trained on a huge corpus in over 100 languages indicate promising approaches and solutions for the implementation of multilingual

applications and have characterized as benchmarks and have introduced remarkable results in Multilingual Text Classification tasks (Wang et al., 2021). To leverage the potential of these approaches, a Multilingual Bias Classification tool is introduced under the scope of the AI4Gov project. This tool is planned to be evaluated and validated in the context of the OECD Scenario and dataset as it has been described in the context of D6.1 - "Specification of UC Scenarios and Planning of Integration and Validation Activities V1".

#### 4.4 Adaptive Analytics Framework

As also mentioned above the *Adaptive Analytics Framework* component is being developed in the context of T4.3 – “Improve Citizen Engagement and Trust utilising NLP”. The scope of this component is to develop the needed ML models for efficiently performing predictive analytics and optimised resource allocation to satisfy the needs of the pilots and assist policy makers.

In the following subsections further information about this component are provided, including the architecture and the internal workflow, the baseline technologies, the source code and a user guide.

##### 4.4.1 Architecture and Internal Workflow

The architecture of the *Adaptive Analytics Framework* is depicted in Figure 23 and thoroughly analyzed below.

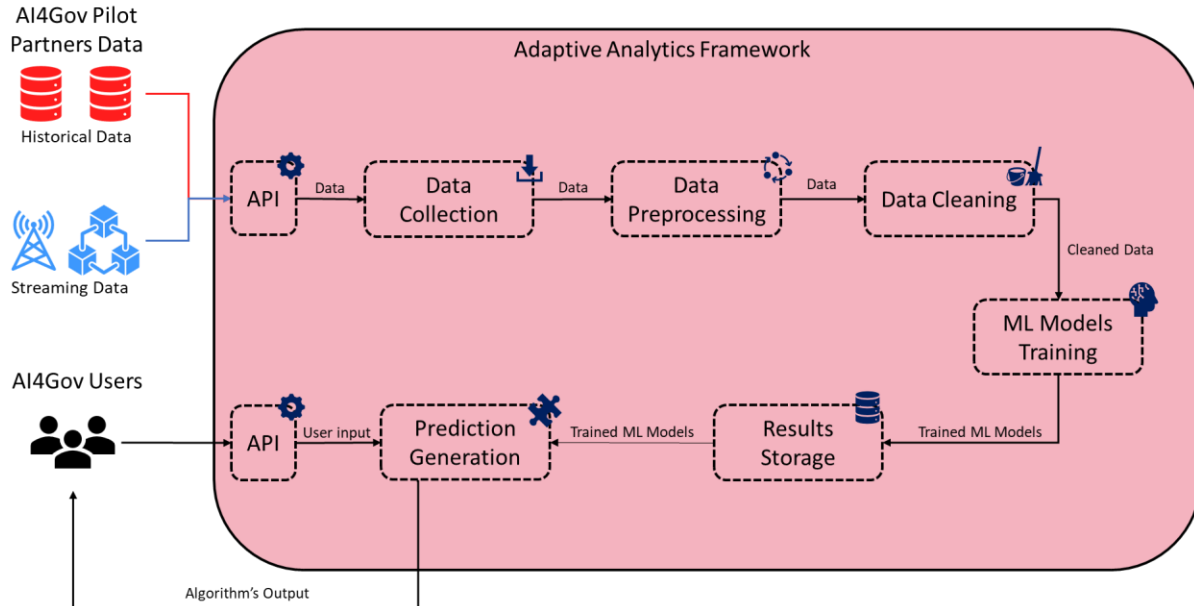


Figure 23: Architecture of Adaptive Analytics Framework Component

As shown in the figure above, the *Adaptive Analytics Framework* retrieves both historical data and streaming data from the corresponding AI4Gov pilot partners in order to analyze them and offer the appropriate predictive ML models. This component will be utilized in Pilot 1 and Pilot 3,

thus it collects and analyzes the corresponding data. More specifically, dedicated APIs (Application Programming Interfaces) have been developed in order to enable the component to retrieve the data from the project's interim repository where all the data from the corresponding data sources are being stored. After the data collection step, the data preprocessing and data cleaning steps take place. In those specific steps, the appropriate techniques are utilized in order to preprocess the data, remove/replace any erroneous values that may affect the training of the ML models, and finally reshape the datasets that are to be trained in a desired form. Since the data are cleaned, the training of the corresponding ML algorithms occurs. In this case, there have been developed a variety of ML algorithms that are trained on the datasets and the ones with the best results in terms of evaluation metrics are then stored in order to allow the users to make predictions based on the trained ML model. It is worth mentioning that the pool of ML algorithms that are tested may be expanded throughout the project, so it is not yet finalized. The users of the AI4Gov project are able to communicate with the component through dedicated APIs, either by utilizing the user interface of the Visualization Workbench, since those two components are integrated, or by performing specific HTTP requests with the appropriate parameters themselves. Based on the provided input by the user (i.e., specific parameters to use for prediction) and the trained ML model utilized, a prediction is generated which is then provided back to the user. At this point, it is also worth mentioning that in most of the cases, the users not only retrieve a prediction but also, based on the requirements of the corresponding use case, interactive diagrams and maps. More details about the usage of the component will follow in section 4.3.4.

#### 4.4.2 Baseline Technologies

*Adaptive Analytics Framework* is based on specific technologies. First of all, the component has been developed with the use of the Python<sup>9</sup> programming language, since there exists a plethora of modules available that are ideal for several ML tasks. Moreover, JavaScript<sup>10</sup> has also been utilized in order for the component to generate the interactive diagrams and maps that were mentioned above. Moreover, the component is also available as Docker Image<sup>11</sup>, thus allowing its flawless integration with the other components of the AI4Gov platform. A complete list of the Python modules that are currently being utilized by the *Adaptive Analytics Framework*, can be found below.

- *certifi*<sup>12</sup>: provides validation of SSL certificates.

---

<sup>9</sup> <https://www.python.org>

<sup>10</sup> <https://developer.mozilla.org/en-US/docs/Web/JavaScript>

<sup>11</sup> <https://www.docker.com>

<sup>12</sup> <https://pypi.org/project/certifi/>

- *Flask*<sup>13</sup>, *Flask\_Cors*<sup>14</sup>, *Requests*: utilized for developing APIs.
- *GeoPy*<sup>15</sup>: manage and analyse geographical data.
- *matplotlib*<sup>16</sup>, *plotly*<sup>17</sup>: used for creating plots, maps, diagrams both static and interactive.
- *numpy*<sup>18</sup>: used for performing mathematical operations.
- *ortools*<sup>19</sup>: provides appropriate tools for solving optimization problems.
- *pandas*<sup>20</sup>: used for data manipulation.
- *scikit\_learn*<sup>21</sup>, *scipy*<sup>22</sup>, *tensorflow*<sup>23</sup>: provide appropriate tools for the development of ML algorithms.
- *ydata-profiling*<sup>24</sup>: provided appropriate tools for performing descriptive analysis on data.

#### 4.4.3 Source Code - Availability and Key Points

The source code of the *Adaptive Analytics Framework* is available on the project’s GitLab repository under the GitLab project named “T4.3 - Improve Citizen Engagement and Trust utilizing NLP”, as shown in Figure 24.

---

<sup>13</sup> <https://flask.palletsprojects.com/en/3.0.x/>

<sup>14</sup> <https://flask-cors.readthedocs.io/en/latest/>

<sup>15</sup> <https://geopy.readthedocs.io/en/stable/>

<sup>16</sup> <https://matplotlib.org>

<sup>17</sup> <https://plotly.com/python/>

<sup>18</sup> <https://numpy.org/>

<sup>19</sup> <https://developers.google.com/optimization>

<sup>20</sup> <https://pandas.pydata.org>

<sup>21</sup> <https://scikit-learn.org/stable/>

<sup>22</sup> <https://scipy.org>

<sup>23</sup> <https://www.tensorflow.org>

<sup>24</sup> <https://github.com/ydataai/ydata-profiling>

Name	Last commit	Last update
Adaptive_Analytics_Framework	Add .gitignore in Adaptive Analytics Framework.	1 day ago
Policy_Oriented_Analytics_and_AI_Algori...	Add FunctionFindBestParams.	1 day ago
Dockerfile	Rename files and add Dockerfile.	3 weeks ago
README.md	Initial commit	4 months ago
apis.py	Add ai4gov_predict_traffic_violation_area_api	2 weeks ago
requirements.txt	Update requirements file to fix flask error.	3 weeks ago

Figure 24: Adaptive Analytics Framework Source Code on GitLab

It is also worth mentioning that in order to simplify the development process and ensure that all the user requirements are met, corresponding issues have been created in the GitLab repository, as shown in Figure 25.

Issue Title	ID	Created	Author	Labels
Predictive ML models accuracy	#19	1 month ago	Kostis Mavrogiorgos	Drinking Water, Pilot 1 (DPB), Pilot 3 (VVV and MT), Sewage Water, Waste Management
Visualization of the Adaptive Analytics Framework's results.	#18	1 month ago	Kostis Mavrogiorgos	Drinking Water, Pilot 1 (DPB), Pilot 3 (VVV and MT), Sewage Water, Waste Management
Predictive analysis	#17	1 month ago	Kostis Mavrogiorgos	Drinking Water, Pilot 1 (DPB), Pilot 3 (VVV and MT), Sewage Water, Waste Management
Connection with external APIs for streaming data	#16	1 month ago	Kostis Mavrogiorgos	Drinking Water, Pilot 1 (DPB), Pilot 3 (VVV and MT), Sewage Water, Waste Management

Figure 25: Sample of GitLab Issues Created for the Adaptive Analytics Framework

#### 4.4.4 User Guide – Installation and Use

Regarding the installation of the *Adaptive Analytics Framework*, this is a straightforward process since the component has been containerized (i.e., a corresponding Docker container has been created) and is available through the AI4Gov GitLab’s container registry.

When the installation is complete, the component will be available through a specified port, and the following APIs will be available for use.

<b>Endpoint Name</b>	<b>ai4gov_interactive_density_mapbox_api</b>
<b>Endpoint URL</b>	http://[SERVER IP]:PORT/ai4gov_interactive_density_mapbox_api
<b>Description</b>	This endpoint is responsible for creating interactive density mapboxes that showcase the geospatial evolution of data in the corresponding use cases for a specific range of dates that are defined by the “date_from” and “date_to” variables shown below. This endpoint will be used in the context of the 3 <sup>rd</sup> pilot (Municipality of Vari-Voula-Vouliagmeni and Ministry of Tourism)
<b>HTTP Method</b>	POST
<b>Parameters</b>	<p>date_from: date in YYYY-MM-DD format</p> <p>date_to: date in YYYY-MM-DD format</p> <p>use_case: available values are “citizens_feedback” and “parking_tickets”. If “citizens_feedback” is selected, then an interactive density mapbox is generated that showcases the geospatial evolution of citizens’ feedback for a specific issue in the given time range. If “parking_tickets” is selected, then an interactive density mapbox is generated that showcases the geospatial evolution of traffic violations tickets issued for a specific type of traffic violation in the given time range.</p> <p>issue_type: the issue type to use for filtering the data and providing the corresponding interactive density mapbox when the “use_case” parameter is equal to “citizens_feedback”.</p> <p>violation_type: the type of violation to use for filtering the data and providing the corresponding interactive density mapbox when the “use_case” parameter is equal to “parking_tickets”</p>
<b>Response</b>	.html file containing the corresponding interactive density mapbox

*Table 4: Description of the ai4gov\_interactive\_density\_mapbox\_api*

<b>Endpoint Name</b>	<b>ai4gov_routing_optimization_api</b>
<b>Endpoint URL</b>	http://[SERVER IP]:PORT/ai4gov_routing_optimization_api
<b>Description</b>	This endpoint is responsible for solving the capacitated vehicle routing problem for the optimization of the garbage trucks’ routes. This endpoint will also be used in the context of the 3 <sup>rd</sup> pilot (Municipality of Vari-Voula-Vouliagmeni and Ministry of Tourism)



<b>HTTP Method</b>	POST
<b>Parameters</b>	date: date in YYYY-MM-DD format for which the optimization of routes should take place
<b>Response</b>	.txt file containing the optimal routes for each vehicle

*Table 5: Description of the ai4gov\_routing\_optimization\_api*

<b>Endpoint Name</b>	<b>ai4gov_predict_traffic_violation_area_api</b>
<b>Endpoint URL</b>	http://[SERVER IP]:PORT/ai4gov_predict_traffic_violation_area_api
<b>Description</b>	This endpoint is responsible for predicting the area in the municipality of VVV where it is more likely a specified traffic violation might occur
<b>HTTP Method</b>	GET
<b>Parameters</b>	<p>algorithm_name: the name of the algorithm that will be used for making the prediction</p> <p>part_of_day: available values are 0 and corresponds to morning, 1 and corresponds to midday, 2 and corresponds to afternoon, 3 and corresponds to night), 'Violation' (– there is a dictionary that maps the integer number to the corresponding violation, in case you need it just let me know), 'Month' ('Week' (0-&gt; Weekday, 1-&gt; Weekend)</p> <p>violation: integer number corresponding to the type of the traffic violation</p> <p>month: integer number ranging from 0 (January) to 11(December)</p> <p>week: boolean value, 0 corresponds to weekday and 1 corresponds to weekend</p>
<b>Response</b>	.html file containing an interactive map showcasing the area where it is more likely the specified traffic violation might occur

*Table 6: Description of the ai4gov\_predict\_traffic\_violation\_area\_api*

A video demonstrating the abovementioned functionalities can be found [here](#). It is worth mentioning that the implementation of the *Adaptive Analytics Framework* is on-going, which means that the functionalities presented in this deliverable are subject to change, as the project progresses and other user requirements might surface from the pilot partners. In any case, in the second version of this deliverable, the finalized version of the implementation should be provided.

## 4.5 Policy-Oriented Analytics and AI Algorithms

As also mentioned previously, *Policy-Oriented Analytics and AI Algorithms* is being developed in the context of T4.3 – “Improve Citizen Engagement and Trust utilising NLP”. Its aim is to develop several NLP algorithms in order to analyse large volumes of text data and also assist the respective AI experts. This particular component consists of the following subcomponents:

- Question Answering Service: this service will provide the necessary tool for allowing the AI experts, developers, and policy makers to perform queries on the OECD papers regarding, among others, raising awareness among them of AI solutions.
- Time Series Analyser: this tool will support the analysis of time series and historical data in order to discover possible trends that will support the corresponding users in the water management cycle and the parking tickets monitoring use cases.
- Multilingual Bias Classification: this tool will support the multilingual identification and classification of bias in the OECD papers, providing all stakeholders information and enhanced knowledge of the types of bias that governments and public authorities take into consideration in their AI policies. It should be noted that, at this point of the project, a standalone implementation of this tool is available and its integration with the *Policy-Oriented Analytics and AI Algorithms* component will follow as the project progresses and will be reported in the next iteration of this series of deliverables.

Lastly, the aforementioned subcomponents should follow the guidelines proposed by the Bias Detector Toolkit component in order to address possible bias in the whole workflow of the *Policy-Oriented Analytics and AI algorithms* component.

In the following subsections, further information about this component is provided, including the architecture and the internal workflow, the baseline technologies, the source code and a user guide.

#### 4.5.1 Architecture and Internal Workflow

The architecture of the *Policy-Oriented Analytics and AI Algorithms* is depicted in Figure 26 and thoroughly analyzed below. For the shake of completeness, a simplified version of the architecture, showcasing the internal workflow of this component is also provided below.

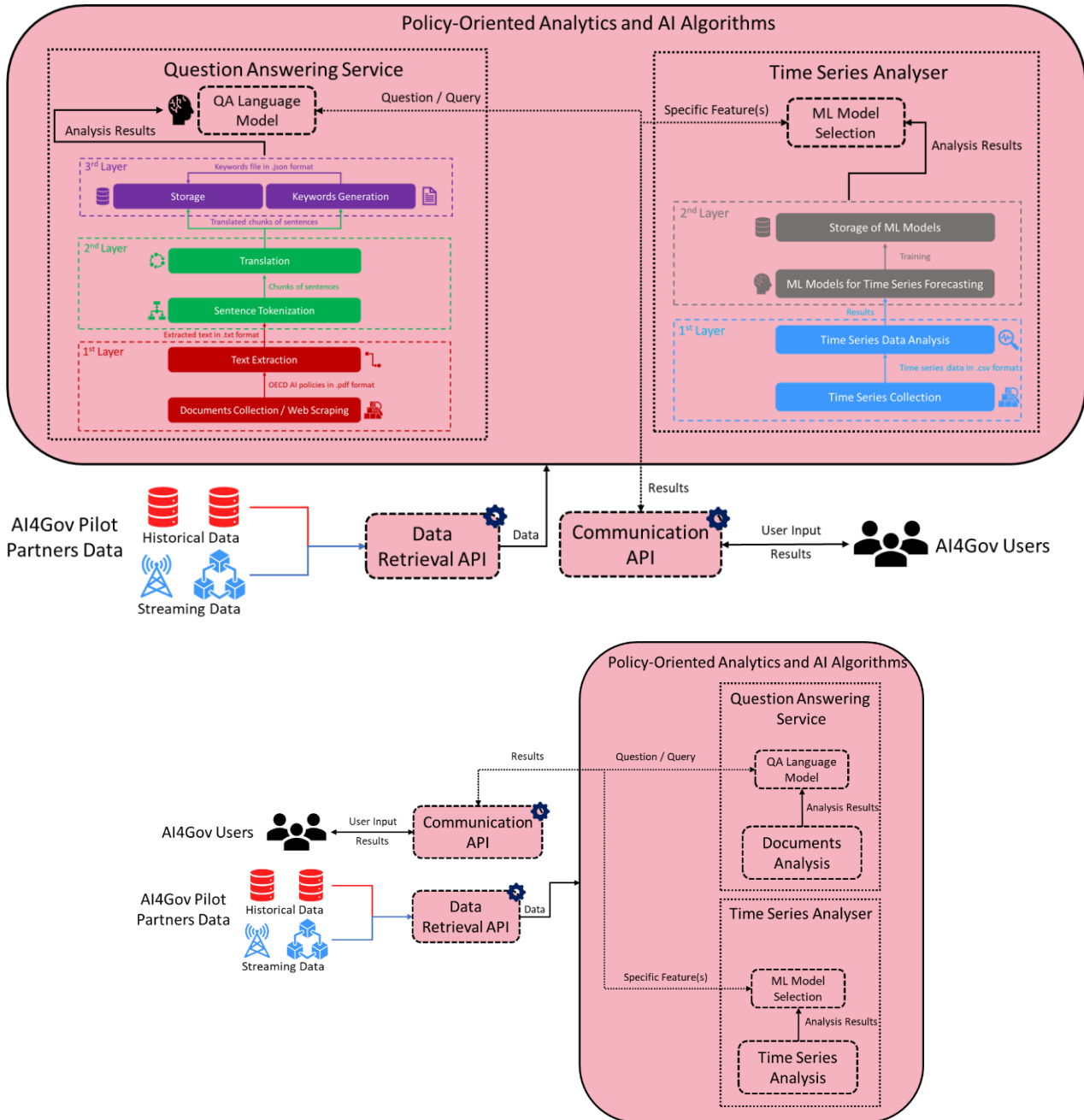


Figure 26: Architecture of Policy-Oriented Analytics and AI Algorithms Component

As mentioned previously and is also shown in figure above, the *Policy-Oriented Analytics and AI Algorithms* consists of two (2) integrated sub-components, namely the *Question Answering Service* and the *Time Series Analyser*, while the *Multilingual Bias Classification sub-component* will

be integrated in later stages. The *Question Answering Service* will be utilized in the second pilot and, more precisely, the OECD policy papers use case. The component seeks to provide the ability to ask AI-related questions to a large knowledge base that consists of official and trustworthy policies and documents of several OECD countries, which refer to the way that each country introduces AI in governance, including potential dangers and methods to mitigate them. As depicted in the figure above, this sub-component consists of three (3) layers, each of them consisting of two (2) steps. Regarding the first layer, it consists of the “Documents Collection / Web Scrapping” step and the “Text Extraction” step. The AI policy documents that construct the knowledge base of the mechanism are publicly available in the OECD AI Policy Observatory<sup>25</sup>. For every available country, there exists a web page that contains a number of useful information about the corresponding policy, except for the policy itself, which consists of one hundred (100) pages on average, such as a short description of it, the name of the responsible governmental body and the objectives of the policy document. In the context of the proposed approach, the policy document should be collected, as well as the URL (Uniform Resource Locator) where it is available. In order to retrieve the aforementioned publicly available information, and since there is no dedicated API (Application Programming Interface), the proposed mechanism performs web scraping, thus retrieving the policy documents in PDF format and the corresponding URLs. As long as the documents are stored, the text extraction step takes place. The policy documents not only contain text, but also images, tables and other text styling that are not useful for the development of the knowledge base and, thus, have to be removed. In the text extraction step, as its name implies, the text from the policy documents is extracted and then stored in plain text files, so that it can later be efficiently manipulated by the rest of the mechanism.

Given the fact that the first layer is complete, the sentence tokenization and the translation of those sentences occur. The policy documents are multilingual, meaning that either they should be translated into a specific language, or different language models should be used when performing the QA task. The latter would be quite costly in terms of resources and, as a result, it is preferred to translate all those text to a certain language, and most specifically English, since most of them are already in that particular language. However, the volume of the text extracted from each policy document is massive, so attempting to translate a whole policy text would be costly or even impossible in terms of resources. As a result, in the proposed mechanism, every text extracted from a policy document is being tokenized by sentences. Then, those sentences are translated into chunks, which is much more efficient in terms of execution time and resource allocation. It should be highlighted that the tokenization of the text is by sentence in order to ensure that the meaning of the text is not being altered in any way. For example, if it was being performed based on a default size of text chunks, then it would be highly possible that a lot of sentences would be trimmed, thus affecting the accuracy of the translation. Moreover, it should be mentioned that for the translation part, opensource APIs are utilized, which are capable of automatically identifying the language of the text and then translating it to English. The selection

---

<sup>25</sup> <https://oecd.ai/en/>

of those APIs, instead of using dedicated machine translation models locally, was made in order to reduce the cost of the mechanism in terms of resources.

Since the texts have been tokenized and translated in chunks, the third layer of the mechanism follows. This layer aims to reduce the response time of the mechanism when a user performs a question. Without this layer, the mechanism should open and search all the translated text chunks in order to find the corresponding answer, which would be catastrophic for the performance of the mechanism and could even lead to bottlenecks. In order to address this issue, the “Keywords Generation” step takes place. In this step, keywords are extracted from every chunk of the translated texts. More specifically, words of high importance such as nouns and verbs are extracted, excluding words that are not of high value, such as conjunctions. The keywords found for every chunk file, along with the corresponding chunk file name, are stored in a JSON (i.e., JavaScript Object Notation) file.

Having the aforementioned JSON file as an index means that whenever a user performs a question on the OECD knowledge base, this JSON file is searched based on the words that the question contains and, as a result, only the text chunks that contain the relative keywords are retrieved and searched for providing the answer. In order to extract the answer from the text chunks, the “deepset/roberta-base-squad2” language model<sup>26</sup> that was retrieved from the Hugging Face repository<sup>27</sup> was utilized. This model was selected because it was trained for QA tasks and it is the finetuned version of “roberta-base” language model<sup>28</sup>, meaning that it is quite efficient for limited resources. It should also be mentioned that an answer may be found in numerous text chunks. In that case, the proposed mechanism provides the answer with the highest accuracy. However, all the available answers, along with the text chunks that contain them, can also be retrieved by the user, in case the answer with the highest accuracy is not sufficient. It is worth mentioning that the language models used for translation and extraction of answers from the OECD policy papers are not fixed but will be under review during the utilization of the component in the second pilot of the AI4Gov project and, if needed, other language models will be considered.

Regarding the *Time Series Analyzer*, it will be utilized in the first and the third pilot of the AI4Gov project, where time series data are available from the pilot partners. In order to retrieve the time series data in real time and periodically from the pilot partners, a dedicated API will be developed. This API will not be connected directly to the data sources but to the interim repository of the AI4Gov project, where all the data will be available for all the technical components of the project. Since the data are retrieved from the interim repository, they are being preprocessed and analyzed by the component in order to identify the appropriate algorithms to be used for time series forecasting. Since not all the data are available at the time that this deliverable is being written, the choice of ML algorithms for the aforementioned task has not been finalized. Several

---

<sup>26</sup> <https://huggingface.co/deepset/roberta-base-squad2>

<sup>27</sup> <https://huggingface.co/tasks/question-answering>

<sup>28</sup> <https://huggingface.co/roberta-base>

algorithms can be utilized, from more statistical models like ARIMA<sup>29</sup> to more complicated algorithms, including recurrent neural networks. The exact algorithms used should be included in the second version of this deliverable. Those trained algorithms should be then stored so that they can be utilized in the corresponding pilot use case and provide a prediction for a specific time range in the future, in the form of an interactive plot. In the context of the first pilot this component will be utilized in the drinking water and the sewage water scenarios, in order to provide predictions for specific variables chosen by the users for a specific time range in the future. As for the third pilot, the *Time Series Analyzer* will be used in the waste management scenario in order to predict certain outcomes from the time series data, such as the possible time when a bin pick-up may occur. At this point, it should be mentioned that since the use case descriptions are subject to change, it is possible that the specific predictions that the *Time Series Analyzer* needs to perform may change in the second version of this deliverable.

As concerns the *Multilingual Bias Classification* subcomponent (Figure 27), it will be utilized in the context of the first pilot and, more specifically, in the OECD dataset, where the identification of bias and the classification of the different policy documents based on it is of highest importance. The latter will provide added value and improved information to the AI experts and all stakeholders by indicating the type of bias that each government takes into account in its AI-related policy actions. It should be noted that this subcomponent follows a standalone implementation at this point of the project and its integration with the overall *Policy-Oriented Analytics and AI Algorithms* and AI4Gov platform will be implemented during the next couple of months. To this end, its internal architecture and workflow are presented and described below to depict its overall functionality and internal pipeline in terms of data processing, analysis and models used so far for an initial evaluation and comparison in terms of their performance and accuracy.

---

<sup>29</sup> An AutoRegressive Integrated Moving Average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

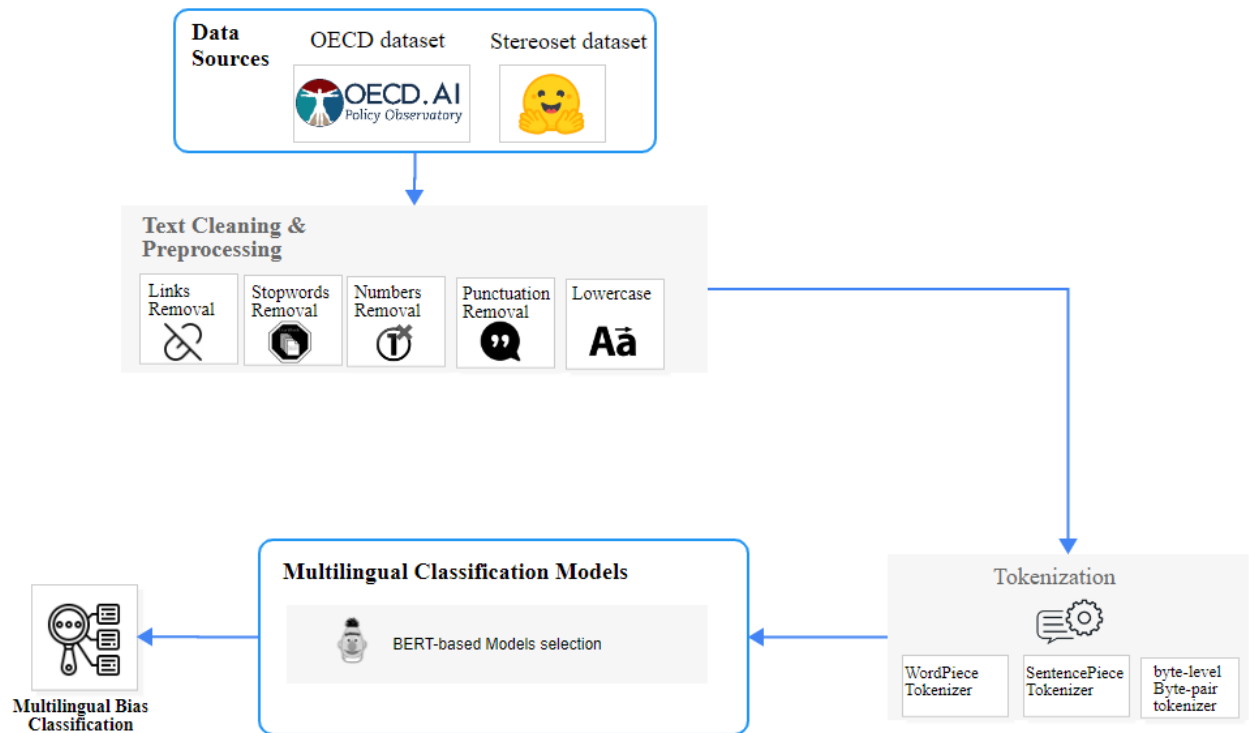


Figure 27: Multilingual Bias Classification Tool

Its main concept is based on the principles of transfer learning where the models are initially trained on a labelled dataset and then are applied on an unlabeled dataset, while some of the models are also evaluated for their performance in a zero-shot classification without any prior training. This sub-component incorporates in its different steps techniques from the fields of ML and NLP, starting from the pre-processing and cleaning of the data, where NLP techniques, such as stopwords, punctuation, lowercasing and links removal, are applied in order to provide cleaner and quality assured data. Afterwards, the tokenization phase is implemented through the utilization of multilingual word and sentence embedding and tokenization techniques by applying Byte-level Byte-pair, WordPiece and SentencePiece tokenization through the Hugging Face library. The selection of the right tokenizer is highly based on the Multilingual Classification model that is utilized and evaluated each time, as the different models handle the words and sentences in different ways. It should be noted that, so far, five (5) different BERT-based models have been evaluated for their performance providing initial results. The four (4) first models are the mBERT, XLM-R, DistilBERT, and mDeBERTa, by utilizing their corresponding through the Hugging Face library, i.e., “bert-base-multilingual-cased”, “xlm-roberta-base”, “distilbert-base-uncased”, and “mdeberta-v3-base” respectively. All these four (4) models were initially trained on the Stereoset dataset and then applied to the project’s OECD dataset following a transfer learning approach. With regards to the fifth model, it is a different version of the XLM-R model, namely the “xlm-roberta-large-xnli”, that is trained on a larger corpus (i.e., the XNLI), than the XLM-R. This model was evaluated for its zero-shot classification without prior trained in the Stereoset dataset to

showcase the applicability of such a model in a more generic and policy-oriented task. However, further finetuning, comparison and evaluation of all these models is needed and will be followed during the next months to optimize the selection phase in an automated way. Within future work this sub-component will also be integrated with the rest of the components of the AI4Gov through appropriate endpoints.

#### 4.5.2 Baseline Technologies

*Policy-Oriented Analytics and AI Algorithms* is based on specific technologies. First of all, the component has been developed with the use of the Python programming, since there exists a plethora of modules available that are ideal for several ML tasks. Moreover, JavaScript has also been utilized in order for the component to generate the interactive diagrams and maps that were mentioned above. Moreover, the component is also available as Docker Image, thus allowing its flawless integration with the other components of the AI4Gov platform. A complete list of the Python modules that are currently being utilized by the *Policy-Oriented Analytics and AI Algorithms*, can be found below.

- *certifi*: provides validation of SSL certificates.
- *Flask, Flask\_Cors, Requests*: utilized for developing APIs.
- *matplotlib, plotly*: used for creating plots, maps, diagrams both static and interactive.
- *numpy*: used for performing mathematical operations.
- *pandas*: used for data manipulation.
- *nltk*<sup>30</sup>, *pytorch*<sup>31</sup>, *transformers*<sup>32</sup>: used for applying several NLP-related algorithms and techniques.
- *langdetect*<sup>33</sup>: provides detection of the language in which a given text is written to.
- *PyPDF2*<sup>34</sup>: provides the necessary methods for extracting text from .pdf files.
- *scikit\_learn, scipy, tensorflow*: provide appropriate tools for the development of ML algorithms.
- *ydata-profiling*: provided appropriate tools for performing descriptive analysis on data.

---

<sup>30</sup> <https://www.nltk.org>

<sup>31</sup> <https://pytorch.org>

<sup>32</sup> <https://huggingface.co/docs/transformers/index>

<sup>33</sup> <https://pypi.org/project/langdetect/>

<sup>34</sup> <https://pypi.org/project/PyPDF2/>



### 4.5.3 Source Code – Availability and Key Points

The source code of the *Policy-Oriented Analytics and AI Algorithms* is available on the project’s GitLab repository under the GitLab project named “T4.3 - Improve Citizen Engagement and Trust utilizing NLP”, as shown in Figure 28.

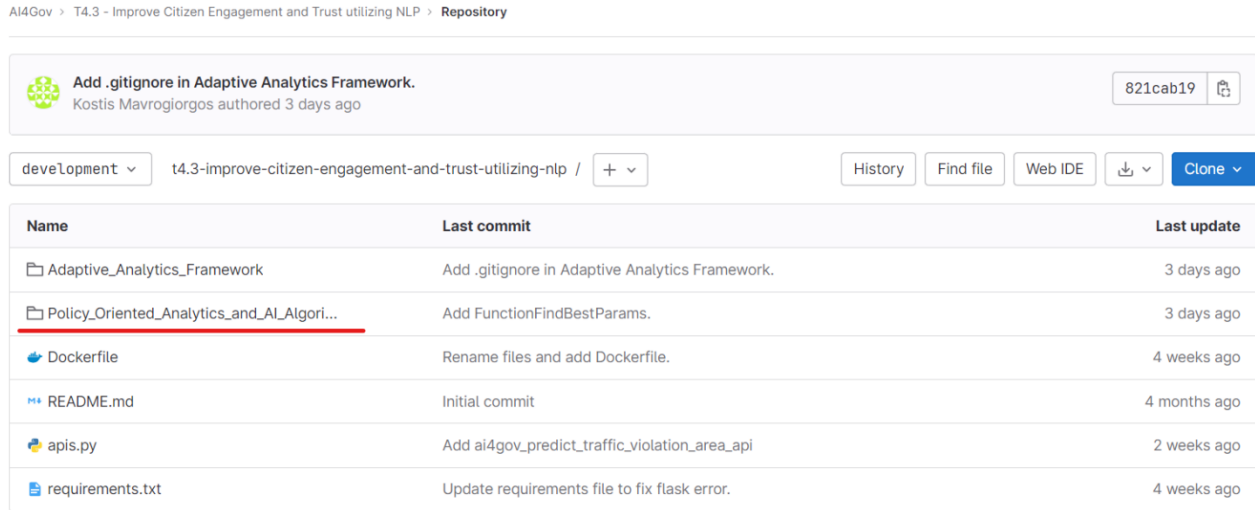


Figure 28: Policy-Oriented Analytics and AI Algorithms Source Code on GitLab

It is also worth mentioning that in order to simplify the development process and ensure that all the user requirements are met, corresponding issues have been created in the GitLab repository, as shown in Figure 29.

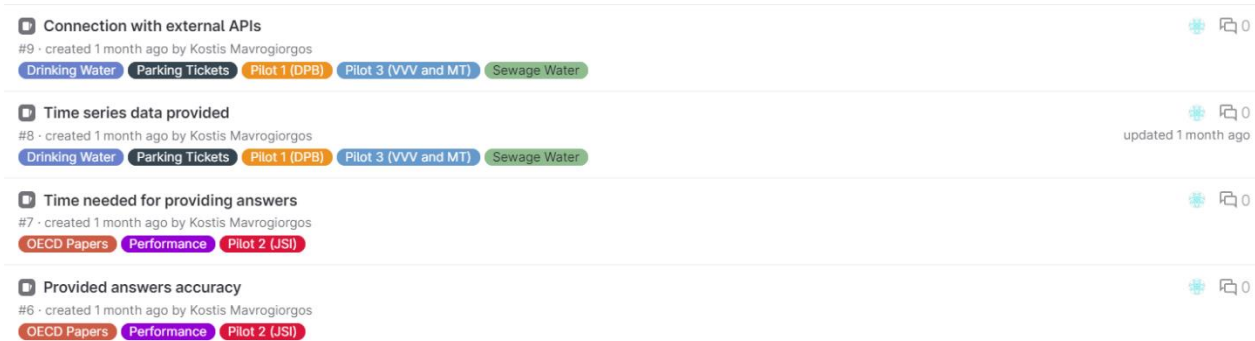


Figure 29: Sample of GitLab Issues Created for the Policy-Oriented Analytics and AI Algorithms

### 4.5.4 User Guide – Installation and Use

Regarding the installation of the *Policy-Oriented Analytics and AI Algorithms*, this is a straightforward process since the component has been containerized (i.e., a corresponding Docker container has been created) and is available through the AI4Gov GitLab’s container registry.

When the installation is complete, the component will be available through a specified port and the following APIs will be available for use.

<b>Endpoint Name</b>	<b>ai4gov_time_series_forecasting_api</b>
<b>Endpoint URL</b>	http://[SERVER IP]:PORT/ai4gov_time_series_forecasting_api
<b>Description</b>	This endpoint is responsible for performing time series forecasting for a specific variable and showcase the prediction in an interactive plot
<b>HTTP Method</b>	GET
<b>Parameters</b>	<p>date_from: date in YYYY-MM-DD format</p> <p>date_to: date in YYYY-MM-DD format</p> <p>use_case: determines the use case in order to select the appropriate algorithms for time series forecasting. Possible values are “waste_magament”, “drinking_water” and “sewage_water” that correspond to the use cases described in deliverable D6.1.</p> <p>variable_name: the name of the variable for which to perform time series forecasting</p>
<b>Response</b>	.html file containing the corresponding interactive plot that showcases the produced prediction

*Table 7: Description of the ai4gov\_time\_series\_forecasting\_api*

<b>Endpoint Name</b>	<b>ai4gov_qa_api_api</b>
<b>Endpoint URL</b>	http://[SERVER IP]:PORT/ai4gov_qa_api_api
<b>Description</b>	This endpoint is responsible for performing question answering on the OECD policies papers. As mentioned above, this API will be utilized in the context of the 2 <sup>nd</sup> pilot and the OECD use case.
<b>HTTP Method</b>	POST
<b>Parameters</b>	<p>countries: the list of the countries whose policy papers the component should search in order to provide the answer to the user.</p> <p>question: the question that the user asks</p>
<b>Response</b>	JSON response that contains the extracted answer and the corresponding text from which it was extracted

*Table 8: Description of the ai4gov\_qa\_api\_api*

A video demonstrating the abovementioned functionalities can be found [here](#). It is worth mentioning that the implementation of the *Policy-Oriented Analytics and AI Algorithms* is ongoing, which means that the functionalities presented in this deliverable are subject to change, as the project progresses and other user requirements might surface from the pilot partners. In any case, in the second version of this deliverable, the finalized version of the implementation should be provided.

#### 4.6 Next steps with T4.3

As for the next steps, as was also mentioned in the corresponding sections of this deliverable, the implementation of the Policy-Oriented Analytics and NLP Algorithms component will continue, several ML algorithms will be implemented and tested in order to provide more accurate results and the internal workflow may also be updated based on the needs of the pilots. A finalized version of this component, including fully functional demonstrators, should be provided in the next, and final, version of this deliverable.

## 5 Conclusions

This deliverable has provided an overview of the work done for T4.1, T 4.2 and T4.3 in the period of first 12 months. We have covered the progress on Bias Detector Toolkit, Situation Aware eXplainability and Policy-Oriented AI and NLP algorithms.

In the context of the Bias Detector Toolkit, the analysis encompassed a state-of-the-art examination of bias mitigation in AI, incorporating real-life examples and an in-depth depiction of the toolkit. The presentation delved into the progression of use cases, particularly focusing on the methodology applied for SDG observatories.

Concerning Situation Aware eXplainability, a comprehensive analysis of the current state-of-the-art was provided alongside a detailed exposition of the SAX4BPM library, a key asset within this domain. The presentation included two illustrative examples showcasing the application of this library in specific use cases.

For Policy-Oriented Analytics and NLP Algorithms a comprehensive state-of-the-art analysis about AI algorithms for policy making, current advancements in NLP and multilingual NLP were provided. Then, the specification for this component and the corresponding subcomponents that it consists of were presented, including architecture and internal workflow, baseline technologies utilized, availability of the source code and a comprehensive user guide describing the installation and use of the aforementioned component.

We are progressing with the development of these technical tasks, as outlined in the next steps sections, in close collaboration with WP6 to co-design solutions for the pilots.

## 6 References

A biased medical algorithm favored white people for health-care programs | MIT Technology Review. (n.d.). Retrieved December 4, 2023, from <https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/>

Ağbulut, Ü. (2022). Forecasting of transportation-related energy demand and CO2 emissions in Turkey with different machine learning algorithms. *Sustainable Production and Consumption*, (pp. 141-157).

A Popular Algorithm Is No Better at Predicting Crimes Than Random People - The Atlantic. (n.d.). Retrieved December 4, 2023, from <https://www.theatlantic.com/technology/archive/2018/01/equivant-compass-algorithm/550646/>

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>

Alexopoulos, C. L. (2019). How machine learning is changing e-government. . 12th international conference on theory and practice of electronic governance , (pp. 354-363).

Amit, G., Fournier, F., Gur, S., & Limonad, L. (2022). Model-informed LIME Extension for Business Process Explainability. <http://ceur-ws.org>

Amit, G., Fournier, F., Limonad, L., & Skarbovsky, I. (2023). Situation-Aware eXplainability for Business Processes Enabled by Complex Events. *Lecture Notes in Business Information Processing*, 460 LNBIP, 45–57. [https://doi.org/10.1007/978-3-031-25383-6\\_5/COVER](https://doi.org/10.1007/978-3-031-25383-6_5/COVER)

Antoniou, C. &. (2022). A survey on semantic question answering systems. *The Knowledge Engineering Review*.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *IBM Journal of Research and Development*, 63(4/5), 4:1-4:15. <https://arxiv.org/abs/1810.01943v1>

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1), 1–27. <https://doi.org/10.1561/22000000006>

Berg, V. d. (2022). Artificial Intelligence and the Future of Public Policy. Retrieved from [https://ec.europa.eu/jrc/communities/sites/jrccties/files/06\\_berg.pdf](https://ec.europa.eu/jrc/communities/sites/jrccties/files/06_berg.pdf)

Bhardwaj, A. D. (2022). Smart IoT and machine learning-based framework for water quality assessment and device component monitoring. *Environmental Science and Pollution Research*.

Blueprint for an AI Bill of Rights | OSTP | The White House. (n.d.). Retrieved December 4, 2023, from <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

Can you make AI fairer than a judge? Play our courtroom algorithm game | MIT Technology Review. (n.d.). Retrieved December 4, 2023, from <https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/>

Cortes, E. G. (2022). A systematic review of question answering systems for non-factoid questions. *Journal of Intelligent Information Systems*, 1-28.

Dutch DPA fines tax authority 2.75M euros. (n.d.). Retrieved December 4, 2023, from <https://iapp.org/news/a/dutch-dpa-fines-tax-authority-2-75m-euros/>

Dutch scandal serves as a warning for Europe over risks of using algorithms – POLITICO. (n.d.). Retrieved December 4, 2023, from <https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/>

Etzion, O., & Niblett, P. (2010). *Event Processing in Action*. Online, ISBN: 9781935182214, 325. <https://www.manning.com/books/event-processing-in-action>

Facing Bias in Facial Recognition Technology | The Regulatory Review. (n.d.). Retrieved December 4, 2023, from <https://www.theregview.org/2021/03/20/saturday-seminar-facing-bias-in-facial-recognition-technology/>

Fahland, D., Fournier, F., Limonad, L., Skarbovsky, I., & Swevels, A. J. (2023). *PMAI'23: Process Management in the AI era*. <http://ceur-ws.org>

First man wrongfully arrested because of facial recognition testifies as California weighs new bills | California | The Guardian. (n.d.). Retrieved December 4, 2023, from <https://www.theguardian.com/us-news/2023/apr/27/california-police-facial-recognition-software>

Fournier, F., Limonad, L., Skarbovsky, I., & David, Y. (n.d.). *The WHY in Business Processes: Discovery of Causal Execution Dependencies*. Retrieved December 4, 2023, from <https://github.com/cdt15/lingam>

Gino Sophia, S. G. (2020). *Water management using genetic algorithm-based machine learning*. *Soft computing*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, 51(5). <https://doi.org/10.1145/3236009>

Hacking AI Resume Screening with Text in a White Font - Schneier on Security. (n.d.). Retrieved December 4, 2023, from <https://www.schneier.com/blog/archives/2023/08/hacking-ai-resume-screening-with-text-in-a-white-font.html>

Hettinga, S. V. (2023). Large scale energy labelling with models: The EU TABULA model versus machine learning with open data.

I Did Nothing Wrong. I Was Arrested Anyway. | ACLU. (n.d.). Retrieved December 4, 2023, from <https://www.aclu.org/news/privacy-technology/i-did-nothing-wrong-i-was-arrested-anyway>

Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing | UCLA Law Review. (n.d.). Retrieved December 4, 2023, from <https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/>

Insight - Amazon scraps secret AI recruiting tool that showed bias against women | Reuters. (n.d.). Retrieved December 4, 2023, from <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/>

Johnson, T. L., Johnson, N. N., McCurdy, D., & Olajide, M. S. (2022). Facial recognition systems in policing and racial disparities in arrests. *Government Information Quarterly*, 39(4), 101753. <https://doi.org/10.1016/J.GIQ.2022.101753>

Leemans, S. J. J. (n.d.). Automated Process Discovery.

Li, X. L. (2021). Machine learning in internet search query selection for tourism forecasting. *Journal of Travel Research*, 1213-1231.

Liu, P., Yuan, W., Jiang, Z., Hayashi, H., Neubig, G., Fu, J., Yuan, W., Jiang, Z., Hayashi, H., Neubig, G., & Fu, J. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(195). <https://doi.org/10.1145/3560815>

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017-December, 4766–4775. <https://arxiv.org/abs/1705.07874v2>

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (n.d.). A Unified Approach to Interpreting Model Predictions. <https://doi.org/10.5555/3295222.3295230>

Machine Bias — ProPublica. (n.d.). Retrieved December 4, 2023, from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Mannhardt, F., De Leoni, M., & Reijers, H. A. (n.d.). Heuristic Mining Revamped: An Interactive, Data-aware, and Conformance-aware Miner. Retrieved December 4, 2023, from <https://fmannhardt.de/g/dhm>.

Market Guide for Event Stream Processing. (n.d.). Retrieved December 4, 2023, from <https://www.gartner.com/en/documents/4010467>

Mavrogiorgou, A. K. (2021). beHEALTHIER: A microservices platform for analyzing and exploiting healthcare data. 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS) (pp. 283-288). IEEE.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 39(1), 53–63. <https://doi.org/10.1080/10580530.2020.1849465>

- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019a). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. [https://doi.org/10.1126/SCIENCE.AAX2342/SUPPL\\_FILE/AAX2342\\_OBERMEYER\\_SM.PDF](https://doi.org/10.1126/SCIENCE.AAX2342/SUPPL_FILE/AAX2342_OBERMEYER_SM.PDF)
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019b). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/SCIENCE.AAX2342>
- Pearl, J. (2011). *Causality: Models, reasoning, and inference*, second edition. *Causality: Models, Reasoning, and Inference, Second Edition*, 1–464. <https://doi.org/10.1017/CBO9780511803161>
- Puh, K. &. (2023). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of Hospitality and Tourism Insights*, 1188-1204.
- Rahman, M. M. (2021). Machine learning on the COVID-19 pandemic, human mobility and air quality: A review. *IEEE Access*.
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope.
- Rehse, J. R., Mehdiyev, N., & Fettke, P. (2019). Towards Explainable Process Predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory. *KI - Kunstliche Intelligenz*, 33(2), 181–187. <https://doi.org/10.1007/S13218-019-00586-1/FIGURES/5>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101. <https://doi.org/10.18653/v1/n16-3020>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). “Why should i trust you?” Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Sadat Qafari, M., & van der Aalst, W. (n.d.). Root Cause Analysis in Process Mining Using Structural Equation Models. [https://doi.org/10.1007/978-3-030-66498-5\\_12](https://doi.org/10.1007/978-3-030-66498-5_12)
- Shafiq, M. T. (2020). Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustainable Cities and Society*.
- Shimizu, S. (2022). *Statistical Causal Discovery: LiNGAM Approach*. <https://doi.org/10.1007/978-4-431-55784-5>
- Shumailov, I., Shumaylov, Z., Kazhdan, D., Zhao, Y., Papernot, N., Erdogdu, M. A., & Anderson, R. (2021). Manipulating SGD with Data Ordering Attacks. *Advances in Neural Information Processing Systems*, 22, 18021–18032. <https://arxiv.org/abs/2104.09667v2>
- Shumailov, I., Shumaylov, Z., Kazhdan, D., Zhao, Y., Papernot, N., Erdogdu, M. A., & Anderson, R. (n.d.). Manipulating SGD with Data Ordering Attacks. Retrieved December 4, 2023, from [https://github.com/iliaishacked/sgd\\_datareorder](https://github.com/iliaishacked/sgd_datareorder)



Spirtes, P., Glymour, C., & Scheines, R. (2001). Causation, Prediction, and Search. Causation, Prediction, and Search. <https://doi.org/10.7551/MITPRESS/1754.001.0001>

Stopar, L., Skraba, P., Grobelnik, M., & Mladenic, D. (2019). StreamStory. *IEEE Transactions on Visualization and Computer Graphics*, 25(4), 1788–1802. <https://doi.org/10.1109/TVCG.2018.2825424>

Tax Administration fined for fraud ‘black list’ | European Data Protection Board. (n.d.). Retrieved December 4, 2023, from [https://edpb.europa.eu/news/national-news/2022/tax-administration-fined-fraud-black-list\\_en](https://edpb.europa.eu/news/national-news/2022/tax-administration-fined-fraud-black-list_en)

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble Adversarial Training: Attacks and Defenses. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. <https://arxiv.org/abs/1705.07204v5>

UK officials use AI to decide on issues from benefits to marriage licences | Artificial intelligence (AI) | The Guardian. (n.d.). Retrieved December 4, 2023, from <https://www.theguardian.com/technology/2023/oct/23/uk-officials-use-ai-to-decide-on-issues-from-benefits-to-marriage-licences>

UnitedHealth used algorithms to deny care, staff say — STAT Investigation. (n.d.). Retrieved December 4, 2023, from <https://www.statnews.com/2023/11/14/unitedhealth-algorithm-medicare-advantage-investigation/>

UnitedHealth uses AI model with 90% error rate to deny care, lawsuit alleges | Ars Technica. (n.d.). Retrieved December 4, 2023, from <https://arstechnica.com/health/2023/11/ai-with-90-error-rate-forces-elderly-out-of-rehab-nursing-homes-suit-claims/>

UnitedHealthcare accused of using AI that denies critical medical care coverage | TechSpot. (n.d.). Retrieved December 4, 2023, from <https://www.techspot.com/news/100895-unitedhealthcare-legal-battle-over-ai-denials-critical-medical.html>

Upadhyay, S., Isahagian, V., Muthusamy, V., & Rizk, Y. (2021). Extending LIME for Business Process Automation. [www.aaai.org](http://www.aaai.org)

Van der Aalst, W. (2016). Process mining: Data science in action. *Process Mining: Data Science in Action*, 1–467. <https://doi.org/10.1007/978-3-662-49851-4/COVER>

Verma, S., Lahiri, A., Dickerson, J. P., & Lee, S.-I. (2021). Pitfalls of Explainable ML: An Industry Perspective. <https://doi.org/10.24963/ijcai.2020/417>

Weske, M. (2012). Business Process Modelling Foundation. *Business Process Management*, 73–124. [https://doi.org/10.1007/978-3-642-28616-2\\_3](https://doi.org/10.1007/978-3-642-28616-2_3)

Xie, G. Q. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management*.

Yoon, W. J. (2022). Sequence tagging for biomedical extractive question answering. In *Bioinformatics* (pp. 3794-3801).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *AIES 2018 - Proceedings of the 2018 AAI/ACM Conference on AI, Ethics, and Society*, 18, 335–340. <https://doi.org/10.1145/3278721.3278779>

Zhong, W. H. (2022). Reasoning over hybrid chain for table-and-text open domain question answering. *International Joint Conference on Artificial Intelligence (IJCAI)*, (pp. 4531-4537).