Funded by the European Union
Horizon Europe
(HORIZON-CL4-2021-HUMAN-01-27
AI to fight disinformation)

AI4TRUST

D5.8

AI4TRUST Platform Specification – Revised Version

PARTNERS

FONDAZIONE BRUNO KESSLER · CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS · UNIVERSITÀ DI TRENTO · NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS" · CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE · GDI Global Disinformation Index · sky · SAHER EUROPE · DEMAGOG · MALDITA.ES · ASTIKI MI KERDOSKOPIKI ETAIRIA KENTRO KATAPOLEMISIS TIS PARAPLIROFORISIS / CIVIL NON-PROFIT COMPANY KENTRO KATAPOLEMISIS TIS PARAPLIROFORISIS · EURACTIV · ASOCIATIA DIGITAL BRIDGE · EUROPEJSKIE MEDIA SP ZOO · FINCONS GROUP · UNIVERSITY OF CAMBRIDGE

Project acronym	AI4TRUST
Project full title:	AI-based-technologies for trustworthy solutions against disinformation
Grant info:	ID 101070190-AI4TRUST
Funding:	HORIZON-CL4-2021-HUMAN-01-27 - AI to fight disinformation (RIA)
Version:	1.0
Status	Final version
Dissemination level:	Public
Due date:	28 February 2025
Delivery date (resubmission):	28 February 2025
Work Package:	WP5
Lead partner for this deliverable:	FINC
Partner(s) contributing:	FBK, CERTH, UNITN, POLITEHNICA, SAHER
Main author(s):	Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)
Contributor(s) and reviewer(s):	Marcello Paolo Scipioni (FINC), Matteo Saloni (FBK), Vitor H. Bezerra (FBK), Riccardo Gallotti (FBK), Serena Bressan (FBK), Danilo Giampiccolo (FBK), Andrew Staniforth (SAHER), Juliet Lodge (SAHER), Evlampios Apostolidis (CERTH), Symeon Papadopoulos (CERTH), Horia Cucu (POLITEHNICA), Alexandru Caranica (POLITEHNICA), Thomas Louf (FBK), Camille Roth (CNRS), Yasmine Houri (CNRS), Stefanie Felsberger (UCAM), Niculae Sebe (UNITN), Elisa Ricci (UNITN), Elena Pavan (UNITN), Lina Livdane (GDI)

Summary of Modifications

VERSION	DATE	AUTHOR(S)	SUMMARY OF MAIN CHANGES
0.1	06/02/2025	Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)	Preliminary notes on the document for the revision.
0.2	13/02/2025	Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)	New section 3.1. Updated version of section 3.3.
0.3	18/02/2025	Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)	Updated section 2 and 3.2 with partners' contributions.
0.4	21/02/2025	Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)	Updated version of section 3.7 and 3.8
0.5	25/02/2025	Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)	Finalisation for the internal review.
0.6	26/02/2025	Riccardo Gallotti, Matteo Saloni, Danilo Giampiccolo & Serena Bressan (FBK), Evlampios Apostolidis (CERTH), Elisa Ricci (UNITN), Lina Livdane (GDI)	Internal review.
0.7	27/02/2025	Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)	Implementation of the required internal reviews and preparation of the version 0.7 of the document for final review.
1.0	28/02/2025	Riccardo Gallotti, Danilo Giampiccolo & Serena Bressan (FBK)	Final review by Project Coordinator and Project Managers.

History of Changes from D5.4 to D5.8

This deliverable of the **"AI4TRUST - AI-based technologies for trustworthy solutions against disinformation"** project, titled **"D5.8 - AI4TRUST Platform Specification - Revised version"**, is a revised version of the previously submitted "D5.4 - AI4TRUST Platform Specification", incorporating the recommendations outlined in the **"General Project Review Consolidated Report (HE)", dated 28 June 2024**, following the project's first Review Meeting. This deliverable is part of **Work Package 5 "Technical implementation of the platform & Security Framework"** (hereinafter referred to as WP5).

The following **changes** have been implemented to address the recommendations:

- **Section 1** has been updated to reflect the revised document structure, ensuring a clearer and more coherent organization.
- **Section 2** now includes both revised requirements and newly introduced ones, providing a more comprehensive foundation for the platform's development.
- **Section 3** has undergone significant revisions, incorporating new and improved subsections to enhance clarity and technical depth.
- **Section 3.1** is a newly added section detailing the Platform Roadmap, which has guided the architectural design of the AI4TRUST Platform.
- **Section 3.2** (previously Section 3.1) has been updated with refined descriptions of the platform's tools, ensuring alignment with the latest developments.
- **Section 3.3** (formerly Section 3.2) now includes an in-depth explanation of the Dataflow, illustrating the integration among platform components and linking them to the different platform versions introduced in Section 3.1.
- **Section 3.7** has been extensively revised, now providing a detailed overview of the dashboards corresponding to the different platform versions outlined in Section 3.1.
- **Section 3.8** has been refreshed with an updated and refined description.
- **Section 5** has been enhanced with improved descriptions to ensure greater clarity and consistency throughout the document.

Table of Contents

List of Acronyms

ACRONYMS	MEANING
AI	Artificial Intelligence
API	Application Programming Interface
CIB	Coordinated Inauthentic Behaviour
DAG	Directed Acyclic Graph
DoA	Description of Action
DWS	Disinformation Warning System
IP	Intellectual Property
PS	Platform Specification
SW	Software
UI	User Interface
WP	Work Package

List of Figures

List of Tables

Executive Summary

The AI4TRUST **D5.8 - "AI4TRUST Platform Specification – Revised version"** is a revised version of the previously submitted "D5.4 - AI4TRUST Platform Specification". It incorporates the **recommendations outlined in the "General Project Review Consolidated Report (HE)", dated 28 June 2024**, following the project's first Review Meeting, with newly added details of the Platform Roadmap and an enhanced description of the overall integration and Dataflow among platform components. D5.8 provides the platform's specifications as detailed in Task (T) 5.1 - AI4TRUST Platform Specification of **Work Package (WP) 5 - Technical implementation of the platform & security framework**.

This is the fifth deliverable of **Work Package (WP) 5** of the European project **AI4TRUST - AI-based-technologies for trustworthy solutions against disinformation**. This deliverable also represents the first part of the means of verification of the **first project Milestone (M1) "AI4TRUST requirements"**, together with the contents foreseen in D6.5 - "Pilot Planning Report – **Revised version"** , i.e., *"Platform and pilot requirements defined. Technological specifications and legal related requirements collected"*.

Specifically, this document aims to **describe the design** of the **AI4TRUST Platform**, based on a **comprehensive analysis** of the current **state-of-the-art solutions**, as well as the **legal and ethical implications** associated with the detection of online **disinformation and misinformation** (hereafter also referred to as **mis/disinformation**).

This **deliverable** is structured into **five main sections**:

1. **Introduction** – outlining the project and its objectives;
2. **Requirements** – providing a summary of the key requirements identified for the project;
3. **Architecture** – reporting on the definition of the platform's architecture;
4. **Integration Methodology** – describing the integration approach used in the project;
5. **Ethical, Security, and Privacy Implication** – addressing critical issues related to ethics, security, and privacy;
6. **Conclusions and Next Steps** – describing the forthcoming steps and the approach for managing the platform's evolution in subsequent iterations.

This **release** focuses on the **most critical components** of the platform's architecture and is **updated** to reflect the **advancements** of the **AI4TRUST Platform**.

1. Introduction

The AI4TRUST project aims to establish a **hybrid platform** that combines the effectiveness of advanced **artificial intelligence** (**AI) solutions** with the expertise of fact-checkers and journalists (hereafter also 'media professionals') **to fight mis/disinformation,** supporting media professionals and policymakers.

The AI4TRUST Platform will operate by monitoring various online social platforms and news data sources (e.g., web news feeds, news aggregators), effectively filtering out irrelevant information and analysing **multimodal content** (text, audio, visual) across multiple languages.

By integrating quantitative indicators about the trustworthiness of a news item and incorporating advanced approaches from the social and computational sciences, the AI4TRUST Platform will provide media professionals with reliable and explainable AI-Driven data analysis methods that can be used **to assess the credibility of news items and debunk mis/disinformation**.

To achieve its goals, the AI4TRUST platform adopts an **incremental approach**, ensuring that each phase of the R&D iterations builds upon the previous one. This structured method allows for the seamless integration of new tools and updates to existing ones. By adhering to this approach, the Platform can progressively enhance its capabilities, starting with initial AI-driven data analysis methods (AI4TRUST Platform v1 – D5.5), then adding to the platform an automated data collection and content analysis that leverages on a subset of the AI-driven data analysis methods (AI4TRUST Platform v2 – D5.6), and ultimately extending the platform with an analysis at-scale of collected data, that takes advantage of both the AI-driven data analysis methods and the automated data collection and analysis of content (AI4TRUST Platform v3 – D5.7). This step-by-step development ensures that the Platform evolves in a coherent and effective manner, continually improving its ability to tackle digital mis/disinformation.

This deliverable, **D5.8 "AI4TRUST Platform Specification – Revised Version"** (incorporating the recommendations outlined in the "General Project Review Consolidated Report (HE)", from 28 June 2024, following the project's first Review Meeting), of the **AI4TRUST Work Package 5 (WP5)** – **"Technical Implementation of the Platform & Security Framework"** – defines the platform roadmap (**Section 3.1**), platform functionalities (**Section 3.2**), and the technical overview (**Section 3.3**) based on a list of requirements (**Section 2**). It specifies the mechanisms for ingestion (**Section 3.4**), elaboration and analysis (**Section 3.5**), and the **Data Lakehouse** (**Section 3.6**), and defines how the given data are accessed (**Sections 3.7 and 3.8**), specifying the integration methodology (**Section 4**) and addressing issues related to ethics, security, and privacy (**Section 5**).

2. Requirements

This section describes the **collected requirements** that guided the design of the AI4TRUST Platform. The clear definition of these requirements and their prioritisation is essential to define the features of the platform and drive their development depending on their priority.

The prioritisation of requirements has been carried out according to the so-called **MoSCoW method**[1], a practice well-known in project management. Its name is derived from the initial of the category names used: **M**ust, **S**hould, **Co**uld and **W**on't. The categories are defined in Table 1.

Table 1 - MoSCoW Categories

CATEGORY	MEANING
Must	The requirement needs to be included for the project to be considered a success
Should	The requirement is important but not strictly necessary for the current release to be considered a success
Could	The requirement is desirable but less critical and can be considered "nice to have"
Won't	The requirement is not planned for delivery

The **current list of requirements** has been created at this stage to support the needs of the project. Their current formulation is based on (i) the described solutions in the AI4TRUST project Description of Action (DoA); and (ii) the discussions, since the beginning of the project, with AI4TRUST project partners, taking into account the planned characteristics of the platform at the current stage.

The categorisation of the requirements is defined based on their **type** (functional / non functional), and whether they are **technical** or **user** functionality/KPI oriented. Functional (F) requirements define the functionality of the system, indicating the features or more in general what needs to be accomplished, while **non-functional (NF)** requirements elaborate a specific characteristic of a system (e.g. compliance, performance, quality or KPI).

Each requirement will be **checked and revised in the next iterations**. Checks and revisions will aim to confirm existing requirements or modify them by providing more clarifications and better rewording; split existing requirements into multiple additional requirements, if need be; delete or deprecate existing requirements; or create new requirements. In the latter case, no reuse of old

[1] https://en.wikipedia.org/wiki/MoSCoW_method

numbers will be made in case of deletions, and new identifying numbers will be used for new requirements, to avoid any possible inconsistencies.

Table 2 presents the current list of requirements, used to define the platform structure.

Table 2 - Requirements List

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
01a	The AI4TRUST Platform supports data collection from the following data sources: YouTube, Telegram and news aggregators.	Must	F	User	Data collector	FBK
01b	The AI4TRUST Platform supports near-real-time data collection.	Must	NF	Technical	Data collector	FBK
02	The AI4TRUST Platform could support input from data sources other than YouTube, Telegram, web news aggregators (e.g., GDELT, news API).	Could	F	User	Data collector	FBK
03a	The data gathering process is driven by a mechanism that facilitates data collection from selected topics.	Must	F	User	Data collector	FBK
03b	The data-gathering process is defined by a set of keywords, associated with the different topics that will be taken into account.	Must	F	Technical	Data collector	FBK
05	Data cover at least eight of the most spoken languages in the EU: English, French, German, Greek, Italian, Polish, Spanish and Romanian representing 70% of the EU population in terms of first language spoken.	Must	NF	User	Data collector	FBK

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
06	Data annotated/labelled from fact-checkers' platforms, together with its metadata (e.g., URLs pointing to textual, audio or visual content, content type, classification) are accessible to consortium partners through APIs and an interactive user interface.	Must	F	Technical	Human validation	FINC
07	New fact-checked information produced by the human fact-checkers network is periodically added (once a week if manually, unlimited if automated) in the AI4TRUST Platform.	Must	F	User	Human validation	FINC
08	The data gathered from different sources are harmonised and pre-processed in nearly real-time.	Must	NF	Technical	Data collector	FBK
09	Data stored in the AI4TRUST Platform are accessible through APIs.	Must	F	Technical	API Gateway	FINC
10	The development adopts a microservice-based approach to assure the dynamic scalability of the AI4TRUST Platform.	Must	NF	Technical	Infrastructure	FINC
11	The development adopts a DevOps approach able to speed up the deployment and fully support the project's agile approach.	Must	NF	Technical	Infrastructure	FINC
12	Standards for API formalisation (e.g., OpenAPI - https://www.openapis.org/) are employed to support API implementation in different development contexts, favour the use of automatic code generation tools (e.g., service stubs) and reduce SW bugs.	Must	NF	Technical	General Platform	FINC

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
13	The AI4TRUST Platform supports ethical and security guidelines inherent with the methodological options available.	Must	NF	Technical	General Platform	SAHER
14	The AI4TRUST Platform meets the ethical, privacy and data protection requirements, also in terms of AI explainability.	Must	NF	Technical	General Platform	SAHER
15	The AI4TRUST Platform provides specific technical solutions to enhance privacy and data protection according to the current regulation (i.e., GDPR).	Must	NF	Technical	General Platform	SAHER
16	The AI4TRUST Platform supports high-performance near-real-time data processing streams.	Should	NF	Technical	Data collection	FBK
17	The AI4TRUST Platform supports non-real-time data input.	Should	F	Technical	Data collection	FBK
18	The AI4TRUST Platform supports non-real-time data processing.	Must	F	Technical	Data collection	FBK
19	The AI4TRUST Platform enables (re-)training of AI models.	Should	F	Technical	General Platform	FINC
20	The AI4TRUST Platform enables a standardised connection with processing services deployed on AI4TRUST partners' dedicated resources.	Must	F	Technical	API Gateway	FINC
21	The AI4TRUST Platform offers centralised CPU computation for the inference phase, to have an economically sustainable cloud-based data-driven platform.	Must	NF	Technical	General Platform	FBK

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
22	The GPU-based processing shall be performed only in partners' premises in order to have an economically sustainable cloud-based data-driven platform.	Must	NF	Technical	General Platform	FBK
23	The egress of RAW data is limited to have an economically sustainable cloud-based data-driven platform.	Must	NF	Technical	General Platform	FBK
24	The AI4TRUST Platform supports the storage of large quantities of data.	Should	NF	Technical	Data Lakehouse	FBK
25	Standardised schemas for data storage are defined.	Must	NF	Technical	Data Lakehouse	FBK
26	Existing data from fact-checkers are imported into the AI4TRUST Platform.	Must	F	User	Data collection	FBK
28	The AI4TRUST Platform provides low-latency access to data that need to be exposed in the UI.	Should	NF	Technical	Database for the web application	FBK
29	The data analysis and classification components are initially trained and validated on a data set curated by humans.	Must	NF	Technical	All WP3/WP4 Modules	CERTH
31	The AI4TRUST Platform supports three groups of users: 1) fact-checkers, journalists and media practitioners; 2) policymakers; 3) researchers.	Must	NF	User	General Platform	FINC

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
32	The AI4TRUST Platform enables users to visualise social networks describing statistically their overall characteristics (such as hierarchy, inequality and bias).	Must	F	User	Social Network Visualisation[2]	CNRS
33	The AI4TRUST Platform enables users to visualise differences in centrality of top nodes (i.e., compare influence in the social network).	Must	F	User	Social Network Visualisation[3]	CNRS
34	The AI4TRUST Platform enables end-users to identify coordinated inauthentic behaviours (CBI), starting from user-defined keywords and exploring the presence of potential disinformation across data sources.	Must	F	User	Coordinated Inauthentic Behaviour tool	UNITN
35	The AI4TRUST Platform enables users to check the reliability of social media, reporting statistics about unreliable contents per channel (YouTube, Telegram).	Must	F	User	Reliability State of Social Media	CNRS
36	The AI4TRUST Platform enables users to check the relevance of a content according to three criteria: language, author's virality and post's virality.	Must	F	User	Relevance evaluator	CNRS

[2] formerly Social Network Analysis
[3] formerly Social Network Analysis

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
37	The AI4TRUST Platform provides an infodemic observatory to the users, reporting language aggregated statistics of the diffusion of disinformation (such as the number of unreliable news circulating, the kind of risk index and the social media volume), through an interactive viewer and filters selectors.	Must	F	User	Infodemic Observatory Tool	FBK
38	The AI4TRUST Platform provides to the users a Disinformation Warning System (DWS).	Must	F	Technical	Disinformation Warning System (DWS)	GDI
39	The AI4TRUST Platform enables users to detect various “disinformation signals” (patterns of textual content typically found in disinformation articles) in the text of single news items.	Must	F	User	Disinformation Signals Detection	NCSR-D
40	The AI4TRUST Platform enables users to detect check-worthy claims in the text of single news items.	Must	F	User	Check-worthy claim detection	FBK
41	The AI4TRUST Platform enables users to quickly examine whether a claim has been already fact-checked in the past.	Must	F	User	Previously fact-checked claim retrieval	FBK
42	The AI4TRUST Platform is able to detect misalignment between the visual content and the associated text description in news items.	Must	F	Technical	Visual-Text Misalignment detection	CERTH

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
43	The AI4TRUST Platform enables users to detect AI-generated or manipulated images (a.k.a. deepfake images).	Must	F	User	Deepfake Image Detection	CERTH
44	The AI4TRUST Platform is able to detect various types of sensational actions/events in the visual content of images and videos.	Must	F	User	Sensational content detection	CERTH
45	The AI4TRUST Platform enables users to detect AI-manipulated videos (a.k.a. deepfake videos).	Must	F	User	Deepfake Video Detection	CERTH
46	The AI4TRUST Platform enables users to find near-duplicates of a given video on the Web, to debunk fakes that rely on the re-use of a video from the past, out of its original context.	Must	F	User	Reverse video search on the Web	CERTH
47	The AI4TRUST Platform is able to detect abnormal events in the visual content of videos.	Must	F	Technical	Video Anomaly Detection	UNITN
48	The AI4TRUST Platform enables users to transcribe speech from single news items into text, allowing textual analysis of audio contents.	Must	F	User	Speech to text	POLITEHNICA
49	The AI4TRUST Platform enables users to detect AI-generated or manipulated audio (a.k.a. deepfake audio).	Must	F	User	Audio Deepfake Detection	POLITEHNICA
50	The AI4TRUST Platform enables users to detect various types of anomalies in an audio stream.	Must	F	User	Audio Anomalies Detection	POLITEHNICA

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
51	The AI4TRUST Platform provides fact-checkers with a human validation interface for manually characterising single news items according to standardised fact-check ratings (e.g., Meta Content Ratings[4]) and fields (e.g., ClaimReview[5]).	Must	F	User	Human Validation	FINC
52	The AI4TRUST Platform provides the users with a user profile.	Must	NF	Technical	General Platform	FINC
53	The AI4TRUST Platform provides the users with a dashboard accessible upon login.	Must	F	User	General Platform	FINC
54	The AI4TRUST Platform provides the users with an interactive user interface to analyse the text, audio, video or image component of a provided news item through AI-based tools.	Must	F	User	General Platform	FINC
55	The AI4TRUST Platform supports an integrated pipeline that executes the data collection of news items, selects the content to be automatically processed, processes the content in nearly real-time with AI-based tools and stores the results (that can be then accessed through APIs and a dedicated dashboard).	Must	NF	Technical	General Platform	FINC

[4] https://transparency.meta.com/en-gb/features/content-ratings-fact-checkers-use/
[5] https://www.claimreviewproject.com/

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE	USER/ TECH.	MODULE	OWNER
56	The AI4TRUST Platform provides the users with a dashboard to monitor the collected news items, allowing to: 1) rank them according to the score of the Disinformation Warning System (DWS); 2) sort/filter them according to various parameters; 3) further analyse them through AI-based tools.	Must	F	User	General Platform	FINC
57	The AI4TRUST Platform enables users to analyse social networks describing the relationships between Telegram channels, through an interactive viewer of the networks (graphs) and filters selectors.	Must	F	User	Social Network Visualisation	CNRS

The list of requirements will be **updated periodically** throughout the development of the project, based on the feedback gathered by all partners and on the advancement of the platform implementation.

A risk related to the accessibility of source APIs needs to be addressed: although, as already included in the Grant Agreement, the possibility of access to **YouTube**'s textual data through their YouTube Researcher programme is confirmed, at the time of writing, the Crowdtangle platform offered by Meta has been discontinued[6] (on the 14th of August 2024) and the **Twitter/X API** that has been available in the past for research purposes is no longer accessible due to a change in the API access policy recently implemented by Twitter/X. This shift in API accessibility had been anticipated as a potential risk and was duly considered within our project's risk management plan. In response to this predicament, we promptly adopted a proactive stance by initiating **efforts to establish formal collaborations with alternative social media platforms** (see Section 3.2.2.1).

[6] https://transparency.meta.com/en-gb/researchtools/other-datasets/crowdtangle/

3. Architecture

This section describes the **overall architecture of the AI4TRUST Platform**, its main functionalities and components, their interactions and how the data flows between those components.

3.1. Platform Roadmap

The AI4TRUST Platform is being developed through an **incremental process, with three planned versions**. Each version incorporates components created during the respective phase of the **three scheduled R&D iterations**.

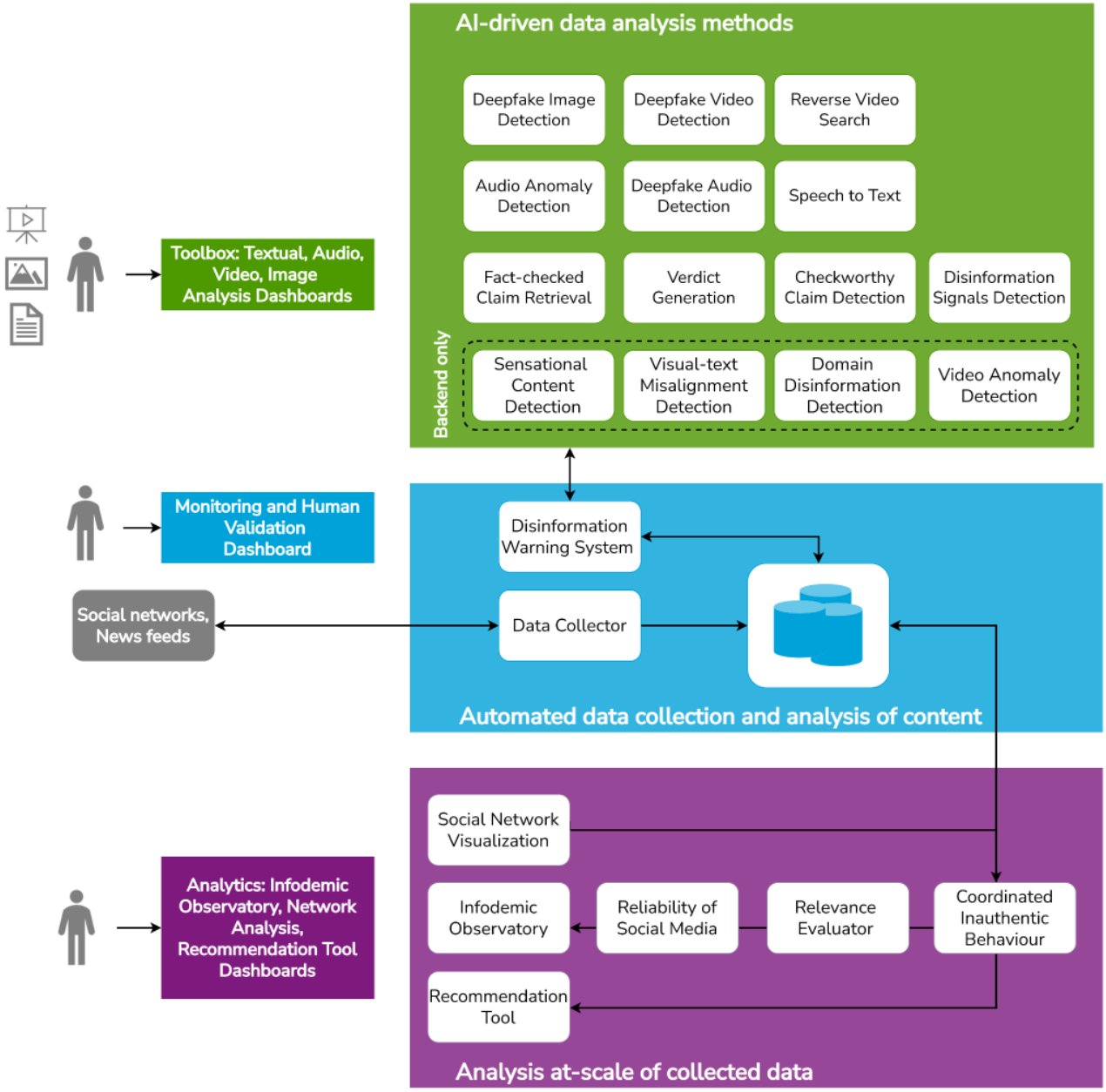

Figure 1 - AI4TRUST Platform overview

The AI4TRUST Platform is designed to **display different functionalities to users**, in which the analyses performed by a rich set of components are shown in **different dashboards** (Figure 1).

The **first section** of the AI4TRUST Platform, depicted in green in Figure 1 focuses on the analysis of individual content items: users can submit textual, audio, video or image content in specific dashboards, where AI-driven data analysis methods are employed to detect potentially misinformative content. This first set of functionalities, developed in AI4TRUST Platform v1 (Figure 2), forms a Toolbox for the analysis of user-provided content, and is described more in detail in Section 3.2.1.

The **second section** of the AI4TRUST Platform, depicted in blue in Figure 1 offers dashboards for the exploration of social media and news feeds content automatically collected and analysed. A processing pipeline takes care of automatically collecting content, having it analysed in the backend by the AI-driven data analysis tools and feeding the Disinformation Warning System, which flags potentially misinformative content. Automatically analysed content is shown to users in the Monitoring Dashboard, where expert users can augment the automatically generated analyses with human-provided feedback in the Human Validation Dashboard (see section 3.7), to refute or confirm the automated results. This second set of functionalities is added to the AI4TRUST Platform v2 (see Section 3.2.2).

The **third section** of the AI4TRUST Platform, depicted in purple in Figure 1 adds visualisation dashboards for displaying advanced analytics capabilities performed on the data collected by the automated pipelines implemented in the second section of the AI4TRUST Platform. Specific dashboards address different kinds of analysis at-scale of the collected data, which can also be accessed through APIs. This third set of functionalities is added to the AI4TRUST Platform v3 (see Section 3.2.3).

Each Platform version incorporates core functionalities that enhance the platform's overall effectiveness. The progressive development of the AI4TRUST Platform (v1, v2, v3) follows a logical and strategic process, aligned with the three Key Exploitable Results of the project (for further details please refer to D7.4[7]).

Each of the three sections of the AI4TRUST Platform introduced above is described in more detail below.

[7] D7.4 - Innovation, Exploitation and Sustainability Plan v2, due by M26 / Feb 2025 (https://ai4trust.eu/public-deliverables/)

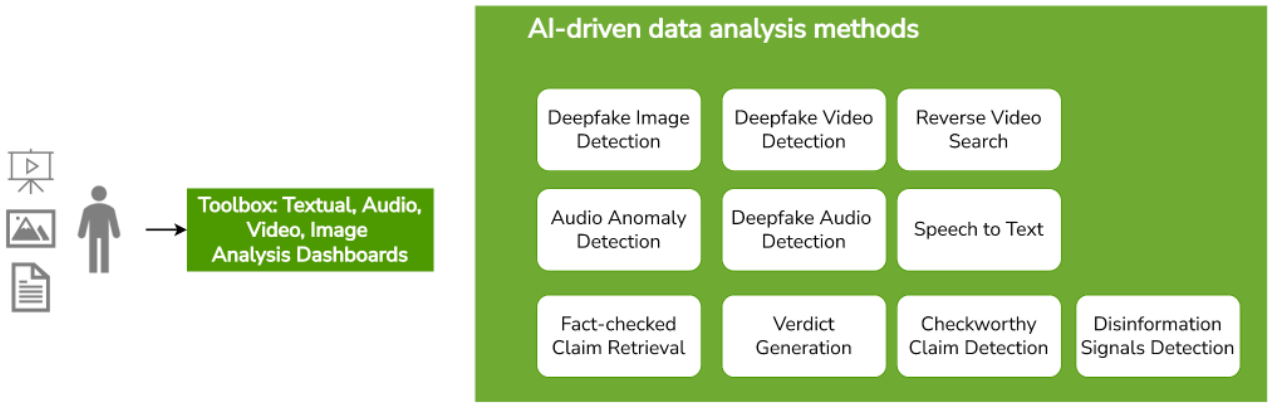

Figure 2 - AI4TRUST Platform v1 overview

The **AI4TRUST Platform v1**, described in D5.5[8], focuses on analysing individual content items using AI-driven data analysis methods developed during the first R&D phase (Figure 2). The aim is to empower users with tools that allow to: detect AI-generated or manipulated images/videos (a.k.a. deepfakes), identify cases where videos posted in the past have been re-used out of their original context to mislead the viewers about a recent event, spot AI-generated or manipulated audio streams (a.k.a. audio deepfakes), and analyse textual data to uncover disinformation indicators (such as hate speech, offensive or sensational language), or check-worthy claims. The tools envisaged for the first version are Textual analysis tools (Fact-checked Claim Retrieval, Verdict Generation, Checkworthy Claim Detection, Disinformation Signals Detection), Audio analysis tools (Audio Anomaly Detection, Deepfake Audio Detection, Speech to Text), Image analysis tools (Deepfake Image Detection) and Video analysis tools (Deepfake Video Detection, Reverse Video Search).

These tools can be directly utilised and assessed by end-users through **dedicated interactive dashboards**, exposed in a Web Application, which exposes all UI dashboard visualisations of the AI4TRUST Platform (see Section 3.7), while some of them also **serve as foundational components for more advanced automated analysis services of subsequent iterations of the platform**, such as the Disinformation Warning System (DWS, which will be integrated into the AI4TRUST Platform v2) that provides a score based on the results of the sensational actions/events in images and videos, the conceptual misalignment between the visual content of an image/video and the associated text and abnormal events in the visual content of videos.

[8] D5.5 - AI4TRUST Platform v1 (https://aiAI4TRUST Platformdeliverables/)

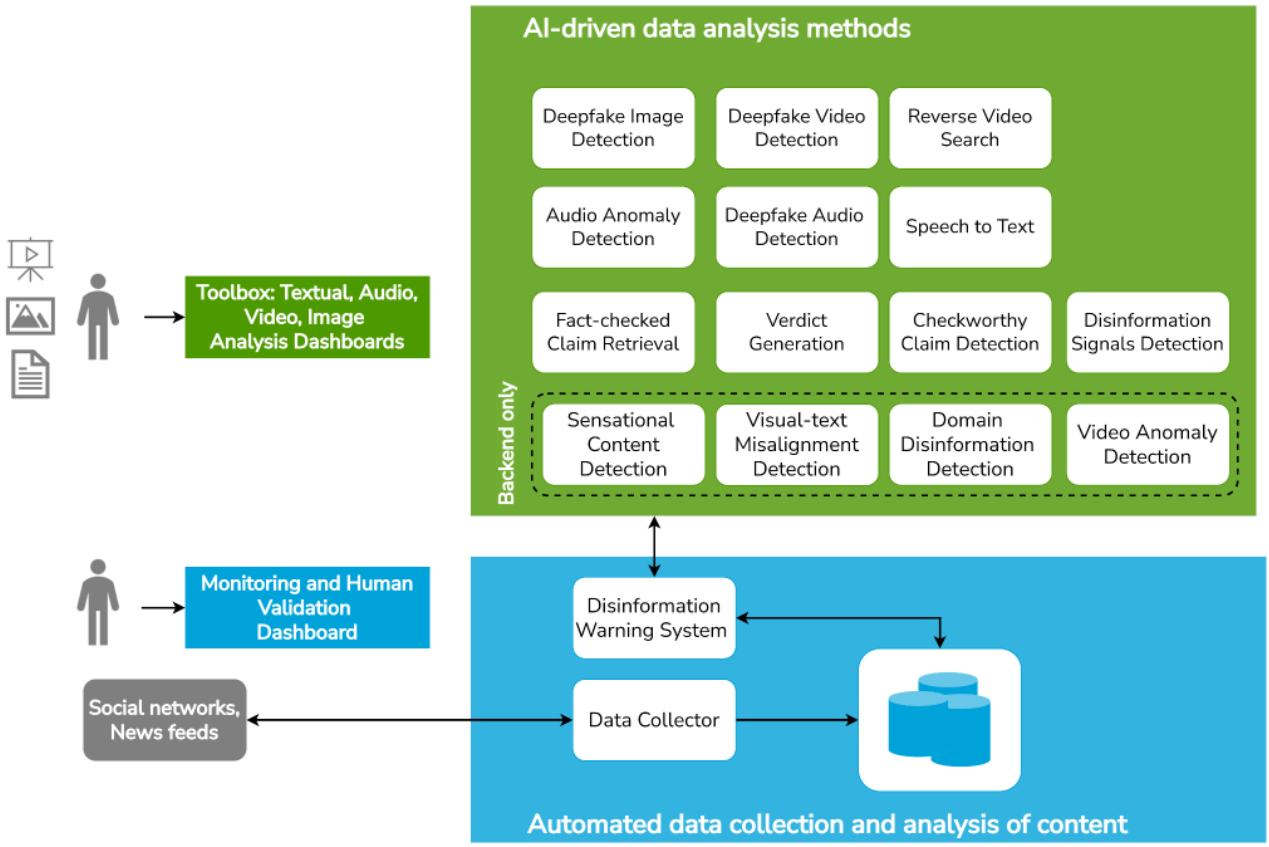

Figure 3 - AI4TRUST Platform v2 overview

The **AI4TRUST Platform v2**, as detailed in D5.6[9], extends the platform with automated collection and analysis of content from social media and news feeds, incorporating new tools developed during the second R&D iteration (depicted in blue in Figure 3) besides the tools already developed in AI4TRUST Platform v1 and further enhanced in AI4TRUST Platform v2 (depicted in green in Figure 3). For what concerns the automated data collection and analysis, content items related to three key topics - public health, climate change, and migrants - are automatically gathered and processed through a **pipeline that integrates AI-driven analysis methods** from the previous phase, along with new techniques that enable the detection of: i) sensational actions/events in images and videos, ii) conceptual misalignment between the visual content of an image/video and the associated text, and iii) abnormal events in the visual content of videos. To this end, AI-driven data analysis methods (in green) are called in the backend to **feed the Disinformation Warning System**, which flags potentially misinformative content. **A Monitoring Dashboard provides users with access to the automatically collected content and the results from the analysis pipeline**. Additionally, a Human Validation dashboard allows fact-checkers to manually review the content, supported by insights from the automated tools, and validate the findings.

[9] D5.6 - AI4TRUST Platform v2 (due by M26 - February 2025, https://aAI4TRUST Platform-deliverables/)

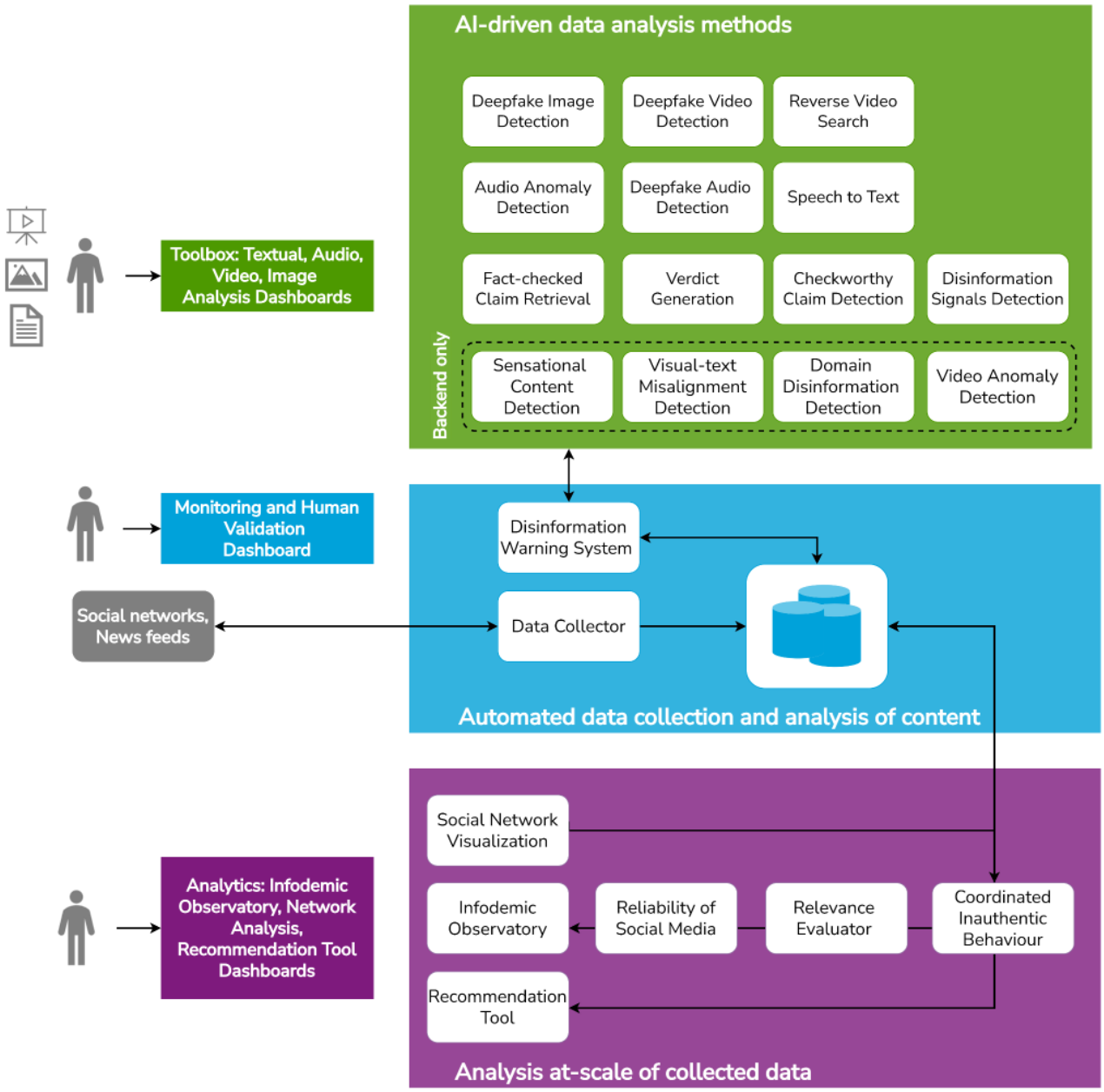

Figure 4 - AI4TRUST Platform v3 overview

The **AI4TRUST Platform v3** (due by M32 / August 2025), which will be described in D5.7[10], will build on the data collected and processed in earlier stages, **introducing advanced methods for analysis at-scale of collected data**, depicted in purple in Figure 4. This iteration will include updates to previously developed tools and new tools created during the final R&D phase. Notably, it will offer users aggregated contextual insights from historical data. Social Network Visualisation will be used to examine the spread of disinformation, while advanced data analytics, such as source reliability assessments and infodemic trend monitoring, will be addressed within the Infodemic Observatory. These features will be complemented by policy-oriented services designed to monitor

[10] D5.7 - AI4TRUST Platform v3 (due by M38 - February 2026, https://ai4trust.eu/public-deliverables/)

disinformation risks and assist policymakers in crafting more effective strategies. The results will be integrated into dedicated dashboards, providing users with graphs and visualisations that offer insights and trends across the overall dataset. Moreover, updates to AI-driven data analysis methods introduced in AI4TRUST Platform v1 are foreseen; additional methods such as the ones for explainable deepfake image/video and audio detection will be added to the Toolbox.

The **AI4TRUST architecture** described more in detail in the following sections is designed to support this roadmap by enabling a step-by-step development of its components. This **incremental approach** ensures that each phase of the **R&D iterations** builds upon the previous one, allowing for the **integration of new tools and updates of existing ones in a structured manner**. By following this roadmap, the platform can progressively enhance its capabilities, from an initial AI-driven data analysis to a more advanced analysis at-scale of collected data.

3.2. Functional Overview

The AI4TRUST project adopts a methodology through which **each of the foreseen functionalities is designed, implemented, delivered to users, and tested**. Each functionality is associated with a reference WP and Task in which it is designed and developed by each owner partner and coordinated by the related WP leaders and the Technical Coordinator, to be integrated and deployed within the overall AI4TRUST Platform.

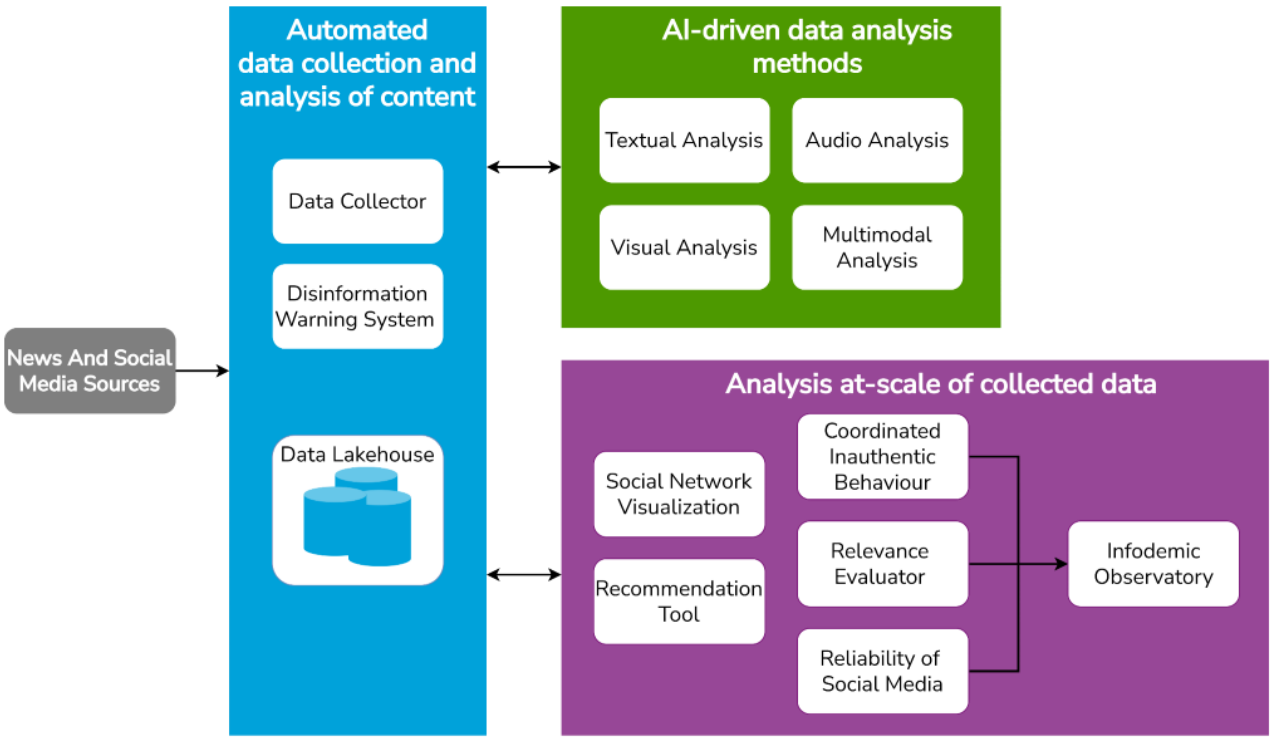

Figure 5 - Functional Overview

In order to **empower the end-users** (i.e., fact-checkers, journalists, media practitioners and policymakers) **to tackle the mis/disinformation** both at single news item level and collective level, the AI4TRUST Platform provides them with **a wide variety of functionalities** (see Figure 5), that can be categorised in:

- **AI-driven data analysis methods**, for the AI-based functionalities responsible for the textual, visual, audio and multimodal analysis of news and social media items at a one-by-one basis.
- **Automated data collection and analysis of content**, for the functionalities responsible for the automatic acquisition of news and social media items from different sources, the analysis through the Disinformation Warning System services, and the storage into the Data Lakehouse.
- **Analysis at-scale of collected data**, for the functionalities responsible for higher-level analysis happening in batch mode across the multiple news and social media items that are collected by the automated data collection, analysed through the AI-driven data analysis methods and archived in the Data Lakehouse.

The specific functionalities for the **AI-driven data analysis methods** are developed within WP3: in particular, in T3.1 - *Textual data analysis methods*, T3.2 - *Audio and Visual data analysis methods*, and T3.3 - *Multimodal data analysis* several different technologies are developed for the analysis of news and social media items, which can happen both on-demand (i.e., applied to content items provided by users) or in an automated way (i.e., automatically applied to collected data). Both the automatically collected and manually-provided AI-driven data analysis methods results can be shown to users through a series of dashboards within the Platform for direct evaluation and use. For additional information, please refer to Section 3.2.1.

Automated data collection and analysis of content activities are developed within WP2 and WP3, which takes care of automated data collection within Task 2.2 - *Social Listening and Data Streams*, Human Validation of content items within Task 2.3 - *Ground Truth*, and analysis of content within Task 3.4 - *Disinformation Warning System*. Automatically collected data is processed through a subset of tools from the AI-driven data analysis methods and used by the Disinformation Warning System (DWS) to make estimates about the trustworthiness of the relevant media items and flag suspicious content. The entire data-collection process and an in-depth explanation of the DWS and its associated AI-driven data analysis methods are further elaborated in Section 3.2.2.

Data stored during the automated data collection and analysis of content process allows to perform an **analysis at-scale of collected data**, providing insights tailored to fact-checkers, journalists, media practitioners, and policymakers. Users can interact with the analyses based on a series of filters, allowing for a more targeted examination of disinformation trends. Further details on these functionalities are provided in Section 3.2.3.

For each of the foreseen functionalities, specific requirements are identified and tracked in the Requirements table (see Table 2 - Requirements List), in accordance with the owner of the modules implementing them, which contribute to the definition of the functionality itself, and in agreement

with the related WP Leader and the Technical Coordinator. **Software modules are identified for each of the functionalities**, and their **deployment strategy** is defined depending on their specific role within the AI4TRUST Platform, taking advantage of the structure envisioned in the Technical Overview (Section 3.2). The following sections describe in detail each of these functionalities.

3.2.1. AI-driven Data Analysis Methods

The AI4TRUST Platform enables the **analysis of single news and social media content**, thanks to the extraction of indicators of which some are directly presented to the end-users through dedicated dashboards, while others operate in the background (formerly backend methods) to support the Disinformation Warning System. These **functionalities** are described in detail in the following sections. For further details, please refer to D5.5[11].

3.2.1.1. Text, Audio, Visual and Multimodal Analysis

The AI4TRUST Platform envisages a series of **AI-driven data analysis methods** that deal with different data modalities and can **empower the end-user in assessing the trustworthiness of a social network or news item** (in text, image, video and audio format). Each method can be **accessed in a backend context through API Layer** (see Section 3.8), and a subset of those methods is provided to the user through **dashboards**. More specifically, the AI4TRUST Platform provides a dedicated **Web Application** (presented in more details in Section 3.7) to the end-users containing a visual indicator of the following tools:

- **Disinformation Signals Detection:** detect various disinformation signals including hate speech, offensive language, clickbait, common disinformation signals (e.g. “Emotional Manipulation”) and unique signals of 6 prevalent manipulation tactics, namely conspiracy theory, discredit, trolling, pseudoscience, science denialism and polarization (e.g. “Secret plot by Powerful Groups” is a sign of conspiracy theories), annotating text segments of the given input text (e.g., an article) with relevant labels and providing a confidence score for each label;
- **Checkworthy Claim Detection:** detect check-worthy claims, labelling the given input text (e.g., a post on a social media platform) as “check-worthy” (i.e., factual and verifiable text that appears to be false, may be of public interest or of impact to the public, or may cause harm to the society, entities, groups, or individuals) and providing a confidence score;
- **Fact-checked Claim Retrieval:** quickly identify claims that have been already fact-checked in the past, comparing the given input text against archived fact-checked claims and providing a similarity score;
- **Verdict Generation:** generate reliable and professional responses (verdicts) about a claim; verdicts are meant to assess claim veracity using trusted sources (i.e. fact-checking articles) for the relevant background information;

[11] D5.5 - AI4TRUST Platform v1 (https://ai4trust.eu/public-deliverables/)

- **Speech to Text:** transcribe speech to text, enabling textual analysis on the given audio content;
- **Deepfake Audio Detection:** detect AI-generated or manipulated audio streams (a.k.a. deepfake audio), providing probability scores w.r.t. the manipulation of different segments of the given audio stream;
- **Audio Anomaly Detection:** detect anomalies in the audio stream, providing a list of identified splicing points with relevant confidence scores;
- **Deepfake Image Detection:** detect AI-generated or manipulated images (a.k.a. deepfake images), classifying an image as a deepfake or not after taking into account a set of state-of-the-art AI-based manipulation techniques, and providing probability scores;
- **Deepfake Video Detection:** detect AI-manipulated videos (a.k.a. deepfake videos), processing a video at the fragment level and classifying it as a deepfake or not after taking into account a set of state-of-the-art manipulation techniques, and providing corresponding probability scores;
- **Reverse Video Search:** perform reverse video search on the Web, by extracting a set of representative keyframes and using them for keyframe-based search through the relevant functionality of online search engines (e.g. GoogleLens); through this process, the user can detect near duplicates of a given video on the Web, and debunk fakes that rely on the re-use of a video from the past to present a different recent or ongoing event;

All the previously listed methods can also operate in a backend context, along with **the ones listed below, that are not accessible to users through a dashboard:**

- **Sensational Content Detection:** analyses the visual content of a given image or video and returns the most relevant action/event from a predefined list of sensational actions/events (typically found in disinformation items) along with a sensational score; the score represents the similarity of the media item (video, image, or both) to the identified action/event, with higher scores indicating greater similarity.
- **Visual-Text Misalignment Detection:** analyses the visual content of a given image or video and the provided textual description, and outputs a score indicating the level of contextual misalignment between them (the higher the score the greater the misalignment between the image/video and the associated text); based on this score, it also provides a predicted label: 0 if the items are aligned and 1 if the items are misaligned;
- **Domain Disinformation Detection:** classifies a domain as present (1) or not present (0) by extracting it from a URL using a system that indexes and labels content, leveraging models and expert validation to assess disinformation across languages.
- **Video Anomaly Detection:** analyses a given video and outputs a score in [0,1] indicating the presence of abnormal events in the input video, where 0 means the video is normal and 1 means it contains abnormal events.

For further details about the presented methods, please refer to deliverables **D3.1 – First Release of AI Tools for Disinformation Detection**[12], submitted on 30 April 2024, and **D3.2 – Second Release of the AI Tools for Disinformation Detection**, due on 30 June 2025.

[12] D3.1 - First release of AI tools for disinformation detection (https://ai4trust.eu/public-deliverables/)

3.2.2. Automated Data Collection and Analysis of Content

The AI4TRUST Platform collects data **from a series of different data sources**, such as social media content from YouTube and Telegram as well as news articles from the Web. In order to have data coming from different sources prepared for a common analysis pipeline downstream, a **normalisation phase** is necessary to align to a **common data format**.

3.2.2.1. Automated Data Collection

The data collection is based on access to social media APIs to ensure a continuous stream of data necessary for informing the **AI4TRUST Platform**. The project was initially envisioned to gather social media data from the platforms **Twitter/X**, **Facebook**, and **YouTube**, as access for research purposes was reasonably granted at the time of writing, and several members of the consortium possessed expertise in analysing data from these platforms. While data access is still granted for **YouTube**, the same is no longer true for **Twitter/X** and **Facebook**. Specifically, **Twitter/X** API access for researchers was discontinued in early **2023**, while the platform **Crowdtangle** by **Meta** was discontinued on the **14th of August 2024**. In response to these challenges, we implemented our contingency plan by immediately diversifying our data sources, including the popular messaging app **Telegram**.

In particular, we have initiated ongoing dialogues with representatives from **Bytedance/TikTok**, **Meta/Facebook**, and **Twitter/X**, reached via contacts of **GDI** (Bytedance), our Fact-Checking collaborators (Meta), or through networking events in **Brussels** (Twitter), in order to identify opportunities for more direct collaborations that would grant our project privileged access. Early discussions have revolved around the pursuit of access to the **TikTok research API**. Although this was, in principle, a promising avenue and an application form under **Article 40** of the **DSA** was made available to European researchers, to the best of our knowledge, it is important to note that this research API is still currently exclusively accessible within the **U.S.**

We consider **Tiktok as an alternative with lower priority**, since the "video + comment" nature of its content is redundant with the characteristics of YouTube, and therefore we did not prioritise this application. We also initiated dialogues with representatives from **Twitter/X**, and we proceeded with a **formal application for data access under Art 40. of the DSA**. This application got **rejected** with the following motivation: *"it does not appear that your proposed use of X data is solely for performing research that contributes to the detection, identification and understanding of systemic risks in the EU as described by Art. 34 of the Digital Services Act"*. To the best of our knowledge, similar requests have been rejected to all European applicants. Lastly, we applied for access to the new Meta Content Library Platform, but the application got rejected with the following motivation: *"certain aspects of your research agenda may violate Meta's product terms. Specifically, building a platform, model, or tool based on MCL data may be considered creating a derivative work, which is restricted by Meta"*.

In place of access to **Meta** and **Twitter**, we are currently at an advanced stage of testing the collection of data from **Telegram**. Additionally, alternative data sources to **Twitter/X**, such as **Mastodon** and **Bluesky**, are being actively considered. These platforms are currently prioritised, as they offer relevant insights into the characteristics of the social network between users sharing content, which could be pivotal for the **AI4TRUST** project's large-scale data analysis. This analysis, which builds on more than **ten years** of social media research on **Twitter**, will benefit from the inclusion of these **alternative social media platforms**.

3.2.2.2. Disinformation Warning System

The **AI4TRUST** Platform incorporates the **Disinformation Warning System (DWS)**, a component designed to identify and highlight disinforming content within media items. The **DWS** integrates outputs from a selected subset of AI-driven data analysis methods (for further details, please refer to Section 3.2.1). The technologies selected to support large-scale data analysis are: **Disinformation Signals Detection, Checkworthy Claim Detection, Sensational Content Detection, Visual-Text Misalignment Detection, Domain Disinformation Detection, and Video Anomaly Detection.** The **DWS** applies a **probabilistic model** to weigh the outputs of the aforementioned AI-driven data analysis methods, ultimately generating a **final risk assessment score**. Based on this score, content is categorised into three risk levels:

- **0.99-0.70**: High risk of disinformation
- **0.69-0.50**: Average risk of disinformation
- **0.49-0.00**: Low risk of disinformation

The model also provides **explainability metrics**, outlining which analysis components most contribute to the final risk score. The **system architecture**, as depicted in **Figure 6**, begins with a **pre-processing component** that filters incoming data by type (e.g., text or image/video). The different data modalities, collected from news and social media sources such as **YouTube**, **Telegram**, and news articles, are directed to the appropriate AI-driven data analysis method. For example, a video is analysed by the **Sensational Content Detection** and **Video Anomaly Detection** methods, while both the video and its associated title/description are analysed for **Visual-Text Misalignment**. Additionally, the video title or description is sent to the **Disinformation Signals Detection** and **Checkworthy Claim Detection**, while the URL is processed by the **Domain Disinformation Detection** to identify domains with a higher risk of disinformation.

The outputs from these various data analysis technologies serve as input to the **probabilistic model**, which assesses the risk of disinformation in the associated media item. This workflow is visually illustrated in the referenced figure.

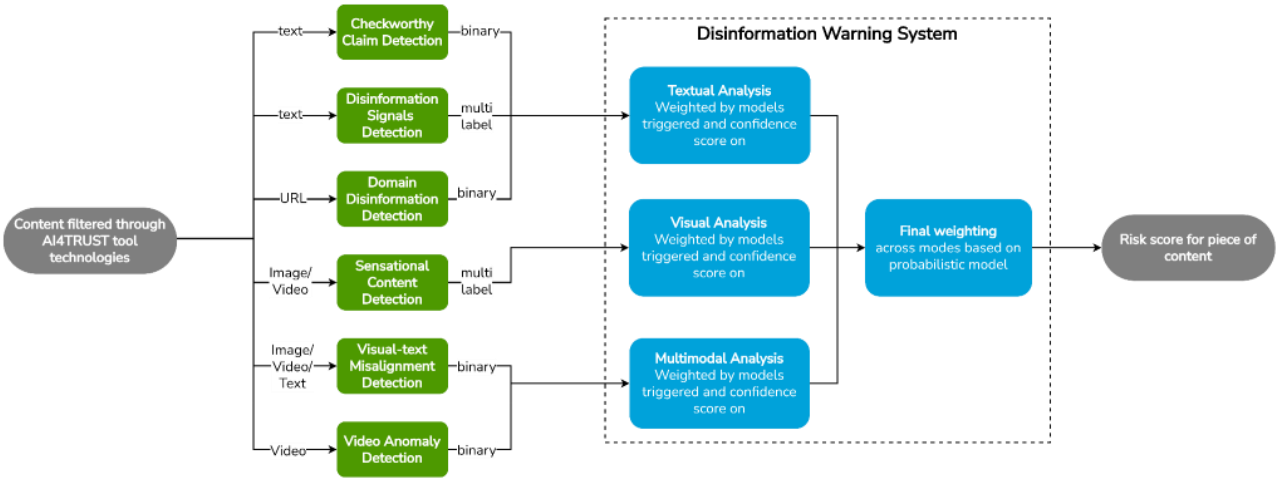

Figure 6 - Disinformation Warning System Architecture from D3.1

3.2.3. Analysis At-Scale of Collected Data

The **AI4TRUST Platform v3** will enhance the existing capabilities by extending the **AI-driven data analysis methods** and automating the data collection and analysis of content, enabling large-scale analysis of the collected data. This advancement will be made possible through the synthesis of relevant indicators derived from the collected data, which will be tailored for the **AI4TRUST Platform's end-users** (i.e., **fact-checkers**, **journalists**, **media practitioners**, and **policymakers**). The platform will allow users to filter the analysis across various dimensions, including **topic**[13], **language, source platform, period of time**. These functionalities are described in detail in the following sections.

3.2.3.1. Social Network Visualisation

The AI4TRUST Platform v3 will provide to the end-users a **"Social Network Visualisation" functionality** to evaluate how hierarchical, inequal or biased the social network is. The general aim would be to map how artefacts (such as URLs) or themes (such as embeddings) are distributed and disseminated within a given system (for instance, Telegram and its channels, focused on a certain overarching topic, as mentioned above), and to explain how they are tracked, etc. In essence, the Platform will be equipped with a **tool for the observation of specific contents**. It would be necessary for the principles to be explainable (calculations, spatialisation, etc.), which is often the case (as simple as explaining degree/strength in certain situations).

[13] for further details, please refer to "D2.1 - Design Of The Methodological Toolbox": https://ai4trust.eu/public-deliverables/

In more detail, in this tool we aim to analyse **how different types of content**, such as URLs (web links) or thematic elements (such as text embeddings), **are spread and shared** within a particular system. If we take Telegram as an example, the process and functionalities of the tool could resemble the following step-by-step explanation:

- **Data Collection:** *artefacts* (e.g., URLs) could be collected from messages exchanged on Telegram and a specific perimeter of relevant channels, while *themes (e.g., embeddings)* could be textual themes derived from the content using methods like word embeddings or topic modelling.
- **Principles and Calculations:** the *degree*, the *strength* of connections between channels and other *clustering information* will be analysed, and will represent respectively the number of connections a channel has, the weight of connections (reflecting the intensity such as how often they are receiving / sending information from one another rather than the richness of connections of a channel to others), and the way neighbours of neighbours are also connected among themselves, more broadly indicating sub-groups or communities within the larger network.

Such a module would provide a comprehensive analysis of **how specific content is disseminated across a social media platform**, such as Telegram. By making the principles clear, such as explaining degree and strength, users can better understand the dynamics of content distribution and the significance of different nodes within the network.

3.2.3.2. Coordinated Inauthentic Behaviour

The AI4TRUST Platform v3 will offer the **"Coordinated Inauthentic Behaviour" functionality** to explore with a cross-platform perspective the social and semantic dynamics that develop around potential or actual mis/disinformation. The functionality builds on the overall content-oriented approach taken by the project and allows to **explore coordination dynamics around same/similar pieces of content across different digital spaces** by checking simultaneously: i) the status of specific pieces of contents within one platform; and ii) the status of same/similar content for how they are present on different platforms.

Whether collected data for the selected platform point to any units that match the trigger words (e.g., videos in YouTube, channels in Telegram), the functionality will provide end-users KPIs such as:

- **Weight**: represents the percentage of units collected for the selected platform that match the trigger(s) chosen by the end user. This will enable the end user to evaluate how widespread is/are the semantic trigger(s) across published contents;
- **Priority**: provides a qualitative indication (YES/NO) about whether the trigger(s) is/are used in prioritised communities identified through Statistical Network Analysis. This will enable the end user to elaborate a judgement about how selected trigger(s) is/are typical of contents that exceed the normal thresholds of activity for the selected platform;

- **Social engagement**: quantitative score (ideally, 0 to 1) expressing the capacity of content units identified via the triggers to catalyse attention (e.g., number of comments received by all YouTube videos containing the triggers over the total amount of comments collected). This will enable the end users to assess the extent to which the triggers are used to identify contents that are particularly catchy or engaging for users;
- **Featuring**: list of up-to-five content units that match selected trigger(s). This will enable end users to go directly to the selected platform and check the unit contents;
- **Status in other platforms**: lists of other platform(s) where the same tigger(s) has/have been spotted. This will give the end user an input to repeat the search procedure, while at the same time exploring/assessing how much coordination is achieved across platforms via flow of same/similar contents.

3.2.3.3. Reliability State of Social Media

The AI4TRUST Platform v3 will assist the end-user in **evaluating the reliability state of social media**. This functionality will rely on a mostly quantitative analysis of the reconstructed network of social interactions on each social media. The intended use of this component is to provide the end-user with **synthetical quantitative indices of the disinformation risk** detected in given areas of social media that the platform is able to map, to bring attention to areas of interest where the risk of having disinformation circulating is higher. This simple tool will be beneficial to media practitioners, fact-checkers, journalists and policymakers, as it will paint a straightforward picture of the reliability state of interactions under study.

In practice, the **"Reliability state of social media" component** will combine three key pillars of the project: **social media data, disinformation signalling technologies, and social network tools**.

1. The reliability of given areas of interaction in the network will be assessed based on cross-sectional social media collected by the automated data collection and analysis of content (see Section 3.2.2). Given the different ontologies of each social media platform, different units of analysis will be defined for each environment. On Telegram, units of analysis are channels. On YouTube, they are videos. This means that when assessing reliability, we will work flexibly at the unit level, i.e. we will either assess the reliability of given Telegram channels or YouTube videos.
2. To get a reliable idea of the disinformation risk presented in each unit of analysis, the social media data described above will be analysed through AI-driven data analysis methods (see Section 3.2.1). This will result in an automatically annotated dataset where each piece of content (e.g. a message on Telegram, a comment or video on YouTube) will be associated with a label of risk as defined by the different AI-driven data analysis methods.
3. Finally, the reliability of social media will be assessed locally using this annotated data: we will focus on local areas in the network as defined in Task 4.2 - *Mapping of social production*

of misinformation and Social Networks Stack tools and render statistics of fraction of unreliable content per unit. On Telegram, this will allow users to identify temporal trends (e.g., x% of channels have shared misinformative content last week, which is y% more channels than last week).

The preferred homogenised output format of this pipeline will be a number, more precisely **a share of likely misinformative content** contained in the local area under study. The user could be able to compute the statistics **filtering on time, language and topic**, though the topic filter might be less robust as it requires sophisticated language analysis to identify topic categories. It is important to note that these measures will necessarily be computed at the source level (either on Telegram or YouTube at this stage, but not cross-platform), since the networks themselves (the second step of the pipeline above) will be source-specific.

3.2.3.4. Relevance Evaluator

In AI4TRUST Platform v3, it will be possible to analyse the relevance of a content through the **"Relevance Evaluator" tool**. It will determine content relevance based on three key criteria. Firstly, it will assess **whether the language is envisaged by the AI4TRUST project**. Secondly, it will evaluate the **author's virality**, considering it statistically significant if it stands out compared to the virality of other neighbouring authors within the social network. Lastly, it will analyse the **post's virality**, determining its statistical importance in relation to the virality of other neighbouring posts within the social network or by the same author. This comprehensive approach ensures that the most relevant content is identified and highlighted based on these critical factors.

3.2.3.5. Infodemic observatory

One tool provided to the end-users in the AI4TRUST Platform v3 will be an "**Infodemic Observatory"**, capable of tracking aggregated statistical information on the quantity of misleading news circulating in a certain period of time on different topics and across various social media platforms, both in absolute terms and relative exposure to the public. The observatory design is **inspired by the COVID-19 Infodemics Observatory platform**[14], developed by FBK. Social media is considered to be in an infodemic state if high volumes of mis/misinformative content are in circulation. To characterise the risk for social media users, **aggregated information will be made available to media practitioners, fact-checkers, journalists and policymakers**. This information will cover general statistics of quantitative indicators for the different media environments identified by the eight European languages included in AI4TRUST, such as the **fraction of news flagged by our AI tools as suspected disinformation**:

- The total number of news collected in the platform;

[14] https://covid19obs.fbk.eu/#/

- The total volume of messages sharing unreliable news items;
- The total exposure these messages are expected to have on social media;
- An infodemic risk index illustrating how likely it is to encounter unreliable news at that moment in time.

This information will be provided as a time series illustrating the **evolution in time** of these indicators in the **different languages, topics and platforms**. At the same time, once aggregated over all different languages, these indicators will be also **grouped geographically**.

3.2.3.6. Recommendation Tool

The **AI4TRUST Platform** aims not only to **map** but also to **counteract** existing **information disorders**, with the "**Recommendation Tool" for policymakers** playing a crucial role in this endeavour. In **AI4TRUST Platform v3**, the **Recommendation Tool** will assist **policymakers** in addressing **information gaps** regarding the current state of the **information ecosystem**, particularly in light of increasing platform restrictions on data access. Moreover, it will support the identification of **systemic challenges** and **risks to democratic societies**.

To achieve this, the **Recommendation Tool** will leverage various functionalities of **AI4TRUST Platform v3**, including **Social Network Visualisation**, **Coordinated Inauthentic Behaviour Detection**, and **Reliability State of Social Media Analysis**. The primary objective of the tool will be to **integrate and rank** aggregated insights from these platform functionalities, providing a structured assessment of their **severity level** alongside guidance on **potential mitigation measures**. Given the complexity of the issues and the diverse contexts in which they arise, the **AI4TRUST Platform** will not rely on a simple **correlation-based approach** between problems and solutions.

Instead, the **Recommendation Tool** will: 1) **Aggregate the analysis of outputs** from **social network analysis**, **coordinated inauthentic behaviour detection**, **reliability assessment**, and the **emotional state of social media**. These findings will be compiled in an **accessible format** tailored to policymakers (see **point 3** below); 2) **Classify mis/disinformation** based on **severity levels**. This classification will distinguish between various levels of **severity and legality**, including: a) **Illegal content**: b) **Legal but harmful content** targeting individuals; c) **Content posing systemic risks** to specific groups (particularly **minoritised or vulnerable communities**); d) **Threats to fundamental rights and democratic processes**, such as online harassment of voter groups or targeted disinformation campaigns against politicians aimed at disrupting or influencing elections; 3) **Generate semi-automated reports** tailored to **policymakers' needs**.

These reports will compile analyses of **disinformation campaigns**, their **virality**, the **emotional state** of discussions surrounding them, and **public reactions on social media platforms**. Additionally, they will visualise **the spread of disinformation**. The reports will not provide **automated mitigation recommendations** but will instead present policymakers with a **scientific**

overview of **existing mitigation measures** along with their **limitations**. Policymakers will have the flexibility to **request reports** based on specific **topics**, **timeframes**, and **geographic/language contexts**, either at the **national** or **EU level**, depending on their areas of responsibility and interest-

3.3. Technical Overview and Dataflow

The adopted Platform design and development approach in AI4TRUST aims to combine the strengths of **AI technologies** with the **expertise and critical thinking abilities of humans** in a platform that will operate in near real-time, targeting multiple online social platforms, and producing a series of analyses based on **multimodal-multilanguage content**. The functionalities shown in the previous sections are implemented in different components, according to the processing needs, e.g., AI-driven data analysis methods are executed in nearly real-time, while analyses at-scale of collected data are executed "offline" in batches. Moreover, components implementing the "AI-driven data analysis methods" are deployed on partners' premises, to leverage their GPU capabilities. The goal of this section is to introduce a **high-level technical representation of the AI4TRUST Platform structure**, represented in Figure 7, based on the requirements listed in Table 2 - Requirements List.

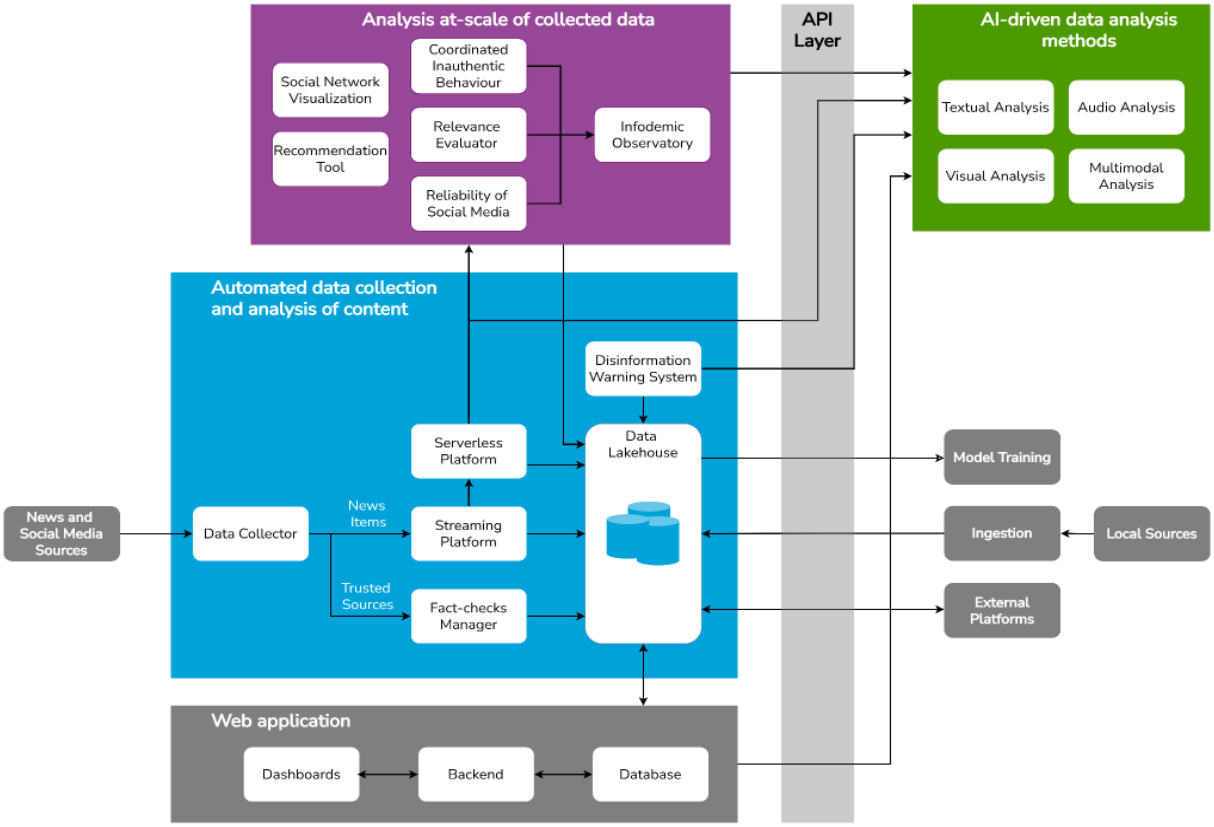

Figure 7- Architecture Overview, representing the centrally deployed components (in blue) and the components deployed on partners' premises (in purple)

As shown in Figure 7, the Platform contains both centrally deployed components (on the left) and components that run on dedicated resources (on the right). The **centrally deployed components** are based on a containerised cloud infrastructure, which allows for seamless integration and updates of individual components without affecting the entire AI4TRUST Platform. By adopting containerization (see req. #10 and #11), it is possible to manage and maintain each component dynamically, generating a platform that can scale and adapt to evolving requirements and advancements in technology. **Other components**, on the other hand, are deployed on **partners' dedicated resources**, and their specifications are described in the next sections. In this way, the identified solution manages different needs in terms of available computational power (see req. #21 and #22) by integrating already available services, safeguarding IP protection, and satisfying computational constraints (particularly regarding GPU usage).

The internal data flow of the **"Automated data collection and analysis of content"** is based on a set of **Data Collectors** used to retrieve near-real-time data from various external **sources** (see req. #01a, #01b and #02). The data collection is related to different topics and done according to a predefined set of keywords (see req. #03b) in different EU languages (see req. #05). Each collector subsequently ensures the proper routing of the collected data to the appropriate components of the platform, for further analysis.

Once the data are collected, it is routed based on its type: it is directed either to the **fact-checks manager**, which harmonises the fact-checks data's structure and content before storage (see req. #07), or to a **Streaming Platform** for data preprocessing and standardisation (see req. #08 and #25). Then, collected data are routed to "AI-driven data analysis methods" mostly deployed on AI4TRUST partners' dedicated resources (see req. #20) and to the internal **Serverless Platform** for data processing (see req. #16).

Subsequently, the collected data are stored in a **Data Lakehouse**[15] to support large volumes of data (see req. #24), to be further processed by components providing "analyses at-scale of collected data" through a series of batch analysis iterations (see req. #18, #32, #33, #34, #35, #36, #37, #38), that extracts the relevant indicators for the analytics and monitoring activities (for further details, see Section 3.2.3).

The Platform envisages a standardised subset of the data from the data lakehouse (see req. #25) to be contained in a low-latency **database** (see req. #28). This subset is selected to provide users with the data shown in dashboards exposed by a **Web Application** (see req. #31). The web application is the main interface for the users and enables them to: (i) analyse single news items (see req. #18, #39, #40, #41, #42, #43, #44, #45, #46, #47, #48, #49, #50); (ii) inspect news items ranked automatically according to their risk of manipulation (see req. #38); (iii) perform manual annotation (i.e., human validation) of single news items (see req. #7, #51); (iv) analyse social media actors and items (see req. #18, #32, #33, #34, #35, #36, #37).

[15] The Data Lakehouse augments the traditional flat data lake with modern and powerful capabilities for managing data schemas, table management, transaction processing and data versioning, all in a highly integrated environment.

Access to the data stored both in the low-latency database and in the data lake house is regulated through an API Layer (see req. #09) acting as an interface for both data retrieval used in **AI model training** (see req. #19, #29 and #31) and for interaction with the platform services (see req. #06). The API Layer approach limits the egress of RAW data (see req. #23) and provide access control mechanisms, ensuring compliance with the legal, ethical, and security standards (see req. #15). The API Layer adopts the OpenAPI formalisation (see req. #12) to enhance interoperability, streamline development processes, and ensure comprehensive documentation. External **local sources** can be ingested into the AI4TRUST Platform using ad-hoc **ingestion scripts** (see req. #17 and #26), which are executed on the partners' premises.

To ensure **compliance with legal, ethical and privacy needs** (#15), the Platform not only adheres to data protection principles, but also does not store any sensitive user data. Specifically, when processing audiovisual content (e.g., YouTube videos), only the URL is archived and transmitted to the AI-driven Data Analysis Methods. Each method retrieves the content directly from the source (e.g., YouTube), analyses it, and returns the results ensuring that the platform never stores the actual content. This approach is consistently applied to other content types**, minimising the storage of raw files while safeguarding metadata integrity** (for further details please refer to Section 3.5.4). By design, this methodology upholds privacy and data protection principles, **minimising the risk of data breaches and reinforcing user confidentiality**. Additionally, at each development phase, legal, ethical, and privacy considerations are continuously reassessed in collaboration with the relevant WP teams and SAHER Europe, AI4TRUST's partner responsible for legal and ethical compliance (for further details please see Section 5).

In detail, the sequence diagram in **Figure 8** shows the **dataflow between the described components**:

- In the **"Toolbox"**, implemented in the AI4TRUST Platform v1, the user can analyse news items (on a one-by-one basis) through the dedicated dashboards exposed by the web application, from which content is forwarded to the appropriate AI-driven data analysis methods through the API Layer, and the results are returned to be shown to the user.
- In the **"Monitoring and Human Validation"**, added in AI4TRUST Platform v2, the workflow begins with the automatic pipeline, where News and social media sources data is gathered by Data Collectors. This data is then ingested by the streaming and serverless platform, where it is structured, processed, and then stored in the Data Lakehouse. The content is further sent to the AI-driven data analysis methods, where specific AI-driven tools, that can be applied on large-scale data, examine it for spotting characteristics typically found in disinformation media items, such as unverified claims, off-topic comments, personal attacks, visually sensational or abnormal events, and conceptually misaligned pairs of visual and textual content. The results of the analyses are subsequently stored and forwarded to the Disinformation Warning System (DWS). This component identifies and surfaces disinforming content by applying an ensemble-based weighting to produce a final risk score that quantifies the likelihood of disinformation in a media item. These processed results are then reintegrated into the storage and made accessible through a Monitoring Dashboard. Fact-checkers can then access the Human Validation Dashboard to assess the DWS-

analysed content and provide annotations that are stored alongside the automated analysis results, enabling a Human and AI-driven hybrid solution.

- In the **"Analysis at scale of collected data"**, implemented in platform v3, batch analyses are executed on collected data on a scheduled basis. The data obtained from the "Automated data collection and analysis of content" components are further analysed and their results are stored on the Data Lakehouse. Analytics Dashboards show the aggregated results to the end-user.

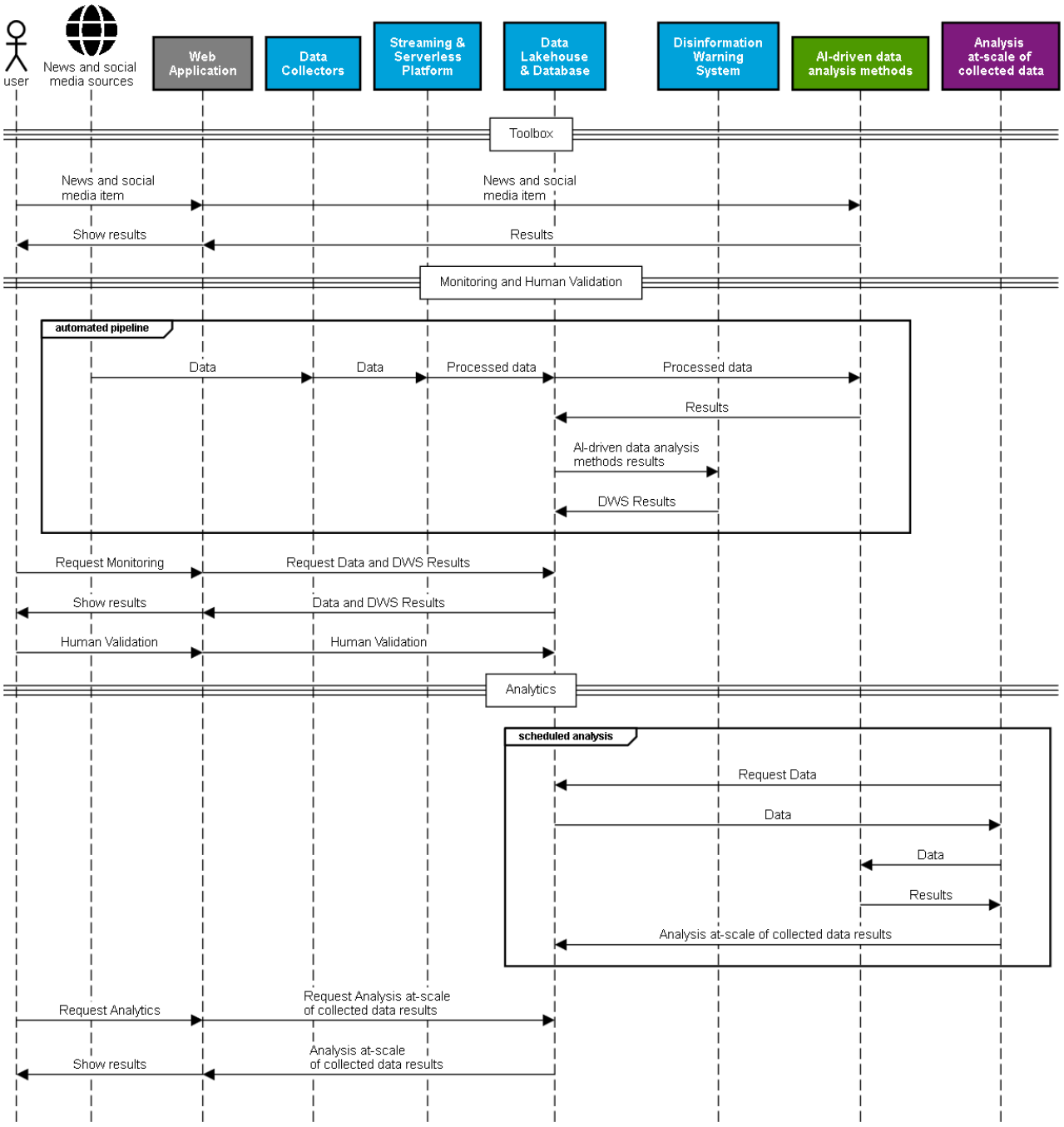

Figure 8 – Dataflow for each AI4TRUST Platform version

The **definition of dataflows**, which highlights the **interconnections** between the presented components, is implemented internally through **suitable backend logics**. This is achieved through a **comprehensive integration effort**, ensuring that the various components **communicate effectively** and operate seamlessly within the **AI4TRUST Platform**. The provided overview provides a **simplified representation of the interaction between the components**, while the **complex orchestration that is implemented in the backend by the system integrator is detailed in D5.6**[16] **and D5.7**[17].

A more in-depth overview of the components introduced in this technical overview is provided in the next sections:

- **Data Ingestion** (Section 3.4) describes how data are collected by the Data Collectors and how the external **Local Sources** are ingested;
- **Elaboration and Analysis** (Section 3.5) describes the Streaming Platform (Section 3.5.1) and the Serverless Platform (Section 3.5.2), the Analysis at-scale of collected data (Section 3.5.3) and the AI-driven data analysis methods (Section 3.5.4);
- **Data Lakehouse** (Section 3.6) describes the Data Lakehouse, its technologies and its data structures;
- **Web Application** (Section 3.7) describes its internal structure**,** and the dashboards shown to users;
- **API Layer** (Section 3.8) describes the API Layer that allows external access to the stored data.

3.4. Data Ingestion

The AI4TRUST Platform actively **monitors multimodal content** encompassing text, audio, and visual elements in **multiple languages**, and from **different sources**. Each data input is inserted into the AI4TRUST Platform based on its respective type, as illustrated in Figure 9.

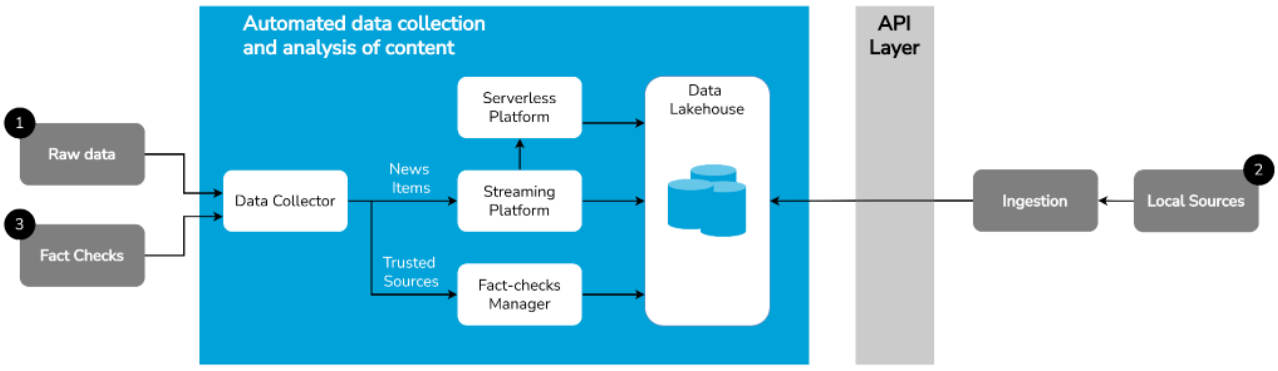

Figure 9 – Data Ingestion

[16] D5.6 - AI4TRUST Platform v2 (due by M26 - February 2025, https://ai4trust.eu/public-deliverables/)
[17] D5.7 - AI4TRUST Platform v3 (due by M38 - February 2026, https://ai4trust.eu/public-deliverables/)

The AI4TRUST Platform manages **three distinct types of input sources**, namely:

1. **News and social media sources:** original/non-processed data gathered from social media platforms, partners' platforms and web news feeds automatically collected;
2. **Local sources:** data gathered through a manually-based ingestion script, which have not been collected automatically before (e.g., data corpora collected in previous projects, externally processed data);
3. **Fact-checks:** annotated/labelled data from fact-checkers.

The **first type of input** defined as **news and social media sources** consists of data sourced in near-real time from **online platforms** (web news feeds and news aggregators), **social networks** and **data residing on partners' data platforms** (e.g., Maldita's one). Currently, our access includes News through GDELT Project[18] and then scraped from the source website, YouTube through the YouTube Researcher programme[19], and Telegram through its API[20]. However, the Twitter/X public API is presently insufficient in providing the necessary data volumes for meaningful scientific research. Consequently, as discussed in Section 3.2.2.1, it is under evaluation the potential substitution of Twitter/X within our project with the emerging competitors Bluesky or Mastodon, that share a similar ontology to the well studied Twitter/X.

Data from each source are obtained through **Data Collectors** (i.e., dedicated components for data collection). Each Data Collector automatically gathers the data according to the specified **topics and keywords** defined by the AI4TRUST consortium partners in the 8 languages covered by our project. Each collected item is then transmitted to a **streaming platform for pre-processing and data harmonisation**, which offers two possible processing paths: (i) send the data directly to the Data Lakehouse or (ii) route it to a serverless platform for further analysis. The collection and ingestion of data is privacy compliant and done securely, according to the ethical, privacy, data protection and security requirements outlined in the AI4TRUST Data Management Plan (reported in D1.2[21]).

A **second type of input** consists of the **manual entry** of **unprocessed data, processed data, or fact-checks** from **local sources**. These sources may include **existing collections, experimental preliminary results, and fact-checking exports**, ensuring the integration of diverse and contextually relevant information within the **AI4TRUST Platform**. This second data input path is manually loaded into the AI4TRUST Platform through **ad-hoc ingestion scripts** that may happen only once (in Figure 9 represented in the External section), as opposed to the collectors' approach which happens in a recurring fashion, thus enabling for **customised inclusion of structured or unstructured data** not yet available within the Platform.

The **third and final data type of input** consists of **fact-checks**, which represent the outcomes of our fact-checkers work. These fact-checks are acquired manually through the **Monitoring and**

[18] https://www.gdeltproject.org/
[19] https://research.youtubeAI4TRUST Platform
[20] .telegram.org/
[21] D1.2 -Data Management Plan (https://ai4trust.eu/public-deliverables/)

Human Validation Dashboard (see Section 3.7) **or possibly automatically**, if obtained from external local sources.

3.5. Elaboration and Analysis

The **AI4TRUST Platform** incorporates an **elaboration and analysis step** to process the **ingested data** before storing them in the **Data Lakehouse**. The **ethical and legal framework** governing the use of these data by the **technology-providing partners** of the **AI4TRUST consortium** is outlined at a general level in **Section 5** of this document. Furthermore, in compliance with the **General Data Protection Regulation (GDPR)**, **"data controllers"** (i.e., partners responsible for ensuring compliance with data processing regulations) and **"data processors"** (i.e., partners authorised to process personal data) have been formally appointed following the signing of the **Data Protection Agreements** attached to the aforementioned **D1.2**[22]. To support the harmonisation, pre-processing and processing in nearly real-time of high volumes of data, **the elaboration and analysis phase is split across several processing components:**

- **Streaming Platform:** part of the automated data collection and analysis of content, it is a platform that supports the nearly real-time preprocessing and harmonisation of the ingested data.
- **Serverless Platform:** part of the automated data collection and analysis of content, it is a platform that supports nearly real-time analysis that might occur after the ingestion or the preprocessing process.
- **AI-driven data analysis methods:** services that need to be executed on AI4TRUST partners' dedicated resources due to their hardware or IP requirements, and that are accessed by the AI4TRUST Platform through their APIs.
- **Analysis at-scale of collected data:** a customised analysis (involving specific components of the overall AI4TRUST Platform) that might occur at a later stage (i.e., not real-time).

As outlined in the **Technical Overview** section (**Section 3.3**) and illustrated in **Figure 10**, the data retrieved from **news and social media sources** by the **Data Collectors** are immediately transmitted to the **Streaming Platform**. The **Streaming Platform**, in turn, forwards the data to the **Data Lakehouse**, the **Serverless Platform**, and the **AI-driven data analysis methods** for **near real-time processing**. The results generated by both the **Serverless Platform** and the **AI-driven data**

[22] Art. 26 GDPR distinguishes between processing and joint control. In a controller-processor relationship, the processor acts solely on documented instructions from the controller and must report any changes. Typically, Art. 28(3) GDPR governs such processing, setting minimum requirements, including data type, purpose, record-keeping (Art. 30 GDPR), codes of conduct (Art. 40 GDPR), certification (Art. 42 GDPR), and data transfer principles (Arts. 44–47 GDPR). While the controller ensures compliance and serves as the primary contact for data subjects, Art. 82 GDPR establishes joint liability between the controller and processor.

analysis methods are subsequently stored in the **Data Lakehouse**, facilitating the **at-scale analysis** of the collected data, which is performed in a **non-real-time** manner.

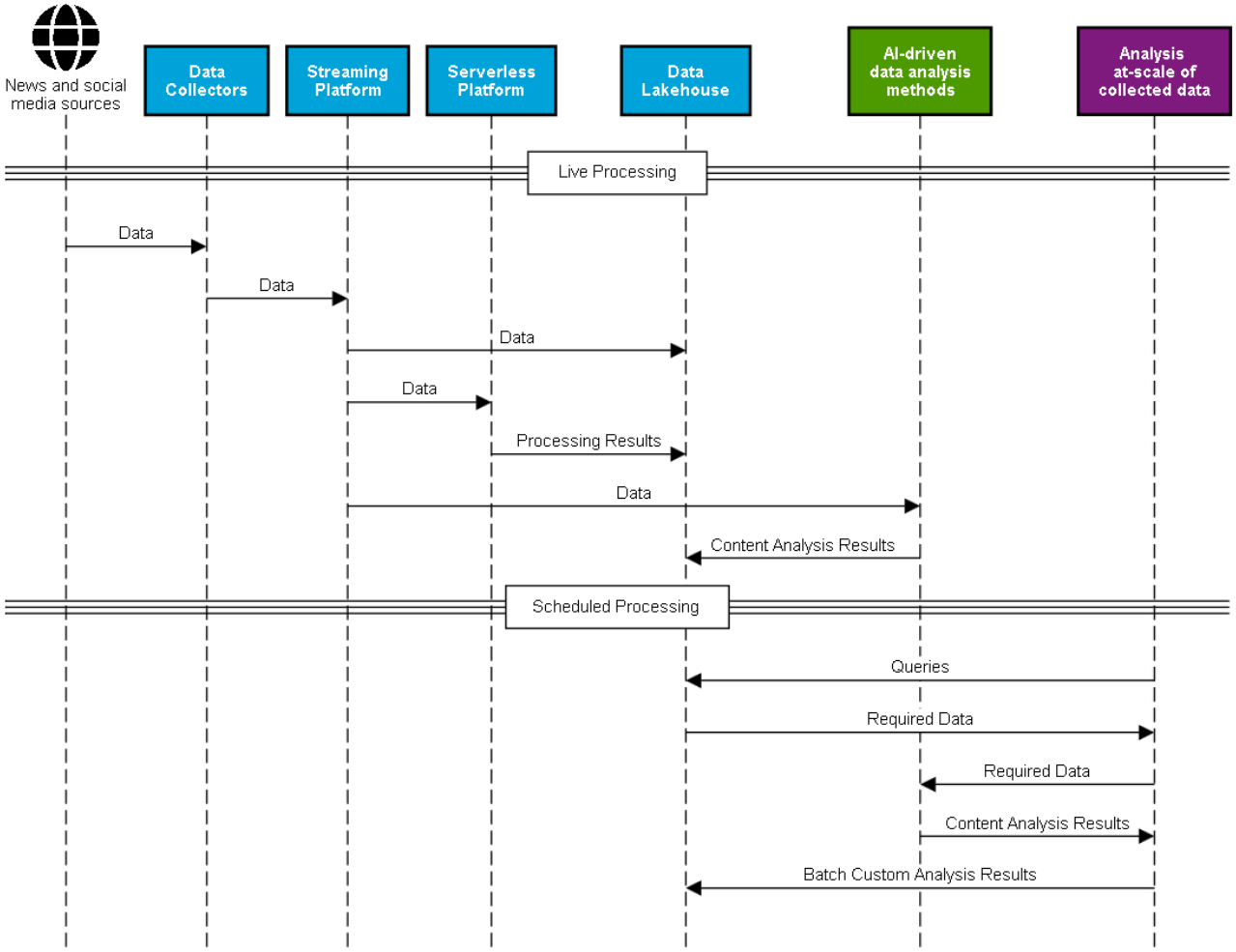

Figure 10 – Elaboration and Analyses Sequence Diagram

The details of each of the mentioned processing components are explained in the following subsections.

3.5.1. Streaming Platform

The **Streaming Platform** is the heart of the (pre)processing layer, a key component which supports all the operations and data flows defined in the design of the AI4TRUST Platform, handling all transformations and loading of datasets and models into the Data Lakehouse. At the core, the Streaming Platform receives from the **Ingestion Layer** all the data parcels, with the news and social media sources data embedded in the streaming in a properly defined way. The **Processing Layer** then applies a Directed Acyclic Graph (DAG) of operations, which are executed via atomic processors connected over the streaming topics in a mesh architecture. This approach ensures **high**

performance, easy scalability and dynamic reconfiguration, all core requirements for the AI4TRUST Platform.

Processors are written as serverless functions, which receive a block of data in input (such as a single data entity or a group) and output the transformed data as a result. The design follows the functional pattern, where processors do not keep state or conditions which could alter the processing, but are invoked in a stateless manner following the functional principles. Given a processor, by feeding a given input it always produces the same output. This principle ensures **proper reproducibility of all the data transformations**, a key aspect for data lineage and for supporting the re-processing of samples with updated processors, models or different DAGs. Given that the data flow is managed via streaming topics, the processors are executed in the serverless platform (see Section 3.5.2). The **main components** are:

- Apache Kafka[23] (with Kafka Connect), as the event streaming platform;
- Kubernetes[24], as a computer platform.

At the end of every flow, a pre-configured **sink processor** collects data associated to the predefined topics and properly persists the content both in the storage layer and in the platform catalogue.

3.5.2. Serverless Platform

The AI4TRUST Platform leverages a powerful **Serverless Platform** to provide developers with the ability to write custom processors for executing dedicated logic during the data collection and processing flow. Developers can implement their custom logic according to the specific source, data and scenarios, focusing on the business logic and not on the operational side, while the serverless platform handles the lifecycle of deployable functions, fulfilling all the operational capabilities such as deployment management, scaling, routing and networking. The **core components** are:

- Kubernetes[25], as the compute platform;
- Nuclio[26], as the serverless platform.

In the AI4TRUST Platform, the serverless layer is responsible for **running data processors**, which apply transformations or elaborations on data. Given the highly specialised needs of every data source, the AI4TRUST Platform design is aimed at enabling developers in writing serverless functions with minimal overhead, which is fully managed by the serverless platform, from building a runnable image to deploying the component, setting up the environment, wiring the connections and collecting logs and metrics.

[23] https://kafka.apache.org/
[24] https://kubernetes.io/
[25] https://kubernetes.io/
[26] https://nuclio.io/

This approach is aimed at focusing developers and researchers' efforts towards properly solving the specific problems, while all the boilerplate and management aspects are handled by the platform itself, in a transparent and observable manner. This approach **lowers the barrier to entrance for new developers**, **facilitates the growth of the AI4TRUST Platform**, and **lowers the maintenance costs**.

AI models developed and used in the AI4TRUST project are a **combination of current state-of-the-art implementations and future research and development**. As such, the AI4TRUST Platform needs a strong integration with the widely adopted and used **Python AI** stack and the common **ML platforms**.

By leveraging a **Serverless Platform** as the execution environment, the AI4TRUST Platform provides an **accessible and versatile infrastructure** that supports a broad range of tools and libraries. This approach ensures that each **AI processor** is **independently developed, deployed, and executed** in an **isolated environment**, while maintaining **direct integration** with the AI4TRUST Platform. A given AI model can be deployed as a **connected processor** within the **serverless platform**, where it **receives input data samples** and **output results** without requiring **direct access** to the streaming system or data repository. This **decouples data access and message routing** from the actual code execution, reducing complexity, facilitating processor reuse, and enabling **greater flexibility and rapid development**.

3.5.3. Analysis At-Scale of Collected Data

The **analysis at-scale of collected data** (foreseen for AI4TRUST Platform v3) is performed in the **Streaming Platform** analysing available data in batches after their collection (i.e., not real-time). In fact, the analysis at-scale of collected data may be useful to execute **further processing of available data**, which may e.g., extend the results obtained in the streaming platform or may help in gathering additional relevant information useful to end-users. The **Analytics** (Infodemic Observatory, Network Analysis, Recommendation Tool Dashboards) will be based on the results of the analysis at-scale of collected data. For further details about the analysis at-scale of collected data please check Section 3.2.3.

3.5.4. AI-driven Data Analysis Methods

The **AI-driven data analysis methods** refer to all the services that are executed on the AI4TRUST partners' dedicated resources, due to their specific hardware (e.g., GPU) or IP requirements (presented in Section 3.2.1). The AI4TRUST Platform uses the **API Layer** as an interface to access **AI-driven data analysis methods** respective functionalities and store the analysis result. To minimise egress bandwidth from the AI4TRUST Platform, whenever possible news and social media sources data is not sent to the AI-driven data analysis methods. Instead, only metadata is sent, and it is the responsibility of the AI-driven data analysis methods to "rehydrate" the data by obtaining the original news and social media sources data.

For instance, considering YouTube, the content is represented by the URL of the video, which is sent to the AI-driven data analysis methods via the API Layer. The services are responsible for downloading the video directly from YouTube, performing the analysis, and returning only the results.

3.6. Data Lakehouse

The **Data Lakehouse** is the repository for all kinds of datasets managed by the AI4TRUST Platform. Following proper design principles, it does not impose a predefined, rigid schema on data formats or structures. Nevertheless, the Data Lakehouse fully manages schema-based datasets, with advanced capabilities, such as schema verification and evolution, consistent views and even point-in-time time travel capabilities, all while supporting transactions where needed. It is an open, scalable and dynamic object storage system, paired with modern data and table formats which ensure consistency, performance, flexibility and ease of use. The **core components** are:

- Minio[27], as the open-source object store;
- Apache Parquet[28] and Apache Iceberg[29], as data and table format.
- Project Nessie[30] as data catalogue

The **Base Layer** supports the **integration of various modern query engines**, even deployed together: the data consistency is ensured by the core layer and the access protocols used by query engines, such as Apache Arrow Flight[31]. The choice of query engines depends on specific needs and can be possibly fine-tuned and even dynamically managed on a per-user/per-scenario basis. The Platform supports Dremio[32], as a user self-service query and data product platform, but additional query engines such as Trino[33] (formerly PrestoSql[34]) or Apache Spark[35] could be easily added and connected to the same sources, delivering additional capabilities to the Data Lakehouse.

Given the domains analysed for the project, and the data sources currently proposed, it is possible to define a **standardised structure for data sets**, which helps in properly managing both collected and generated data in a harmonised and compliant way. In order to properly model the data entities that live inside the Platform, it is needed to analyse the context of the project and conduct a detailed survey of the various datasets collected and features defined by AI models, with the objective of defining the proper standardisation procedure for every kind of content. Nevertheless, it can be

[27] https://min.io/
[28] https://parquet.apache.org/
[29] https://iceberg.apache.org/
[30] https://projectnessie.org/
[31] https://arrow.apache.org/
[32] https://www.dremio.com/
[33] https://trino.io/
[34] https://prestodb.io/
[35] https://spark.apache.org/

defined as a **base set of data models** used for the project, by reducing all the various complex data types to a composition of basic data models centred around the content type: 1) Core data; 2) Base data: Text/Image/Audio/Video.

By defining a core model with identification, typing and tracking metadata, content-specific base models are built for the various primitive data types: text, audio, video, image. On top of this, **complex data types can be modelled** as a composition of base data models and core data plus additional properties, which can be modelled as explicit types following a schema or as unstructured key/value pairs. This enables defining the processing and ML blocks which handle core/base data models, ensuring the **reusability and composability of the processing stack**.

Internal access to all the datasets stored inside the data lakehouse are possible via the following interfaces:

- Query execution via the query engine;
- Direct file access for both artefacts[36] and data files;
- Custom REST read-only API access for specific data models, deployed on a per case approach to facilitate integration with external processors and partners.

Every **stakeholder** within the consortium has access to one or more **interfaces**, with **strict credential management, authentication, and authorisation protocols** in place. This ensures compliance with **data licences, usage rights, and privacy regulations**, safeguarding sensitive information and maintaining controlled access to the Platform's resources.

3.7. Web Application

User interaction with data of interest is made possible through the use of a structured **Web Application**. From the technical perspective, an in-depth visualisation of the main components of the web application is shown in Figure 11.

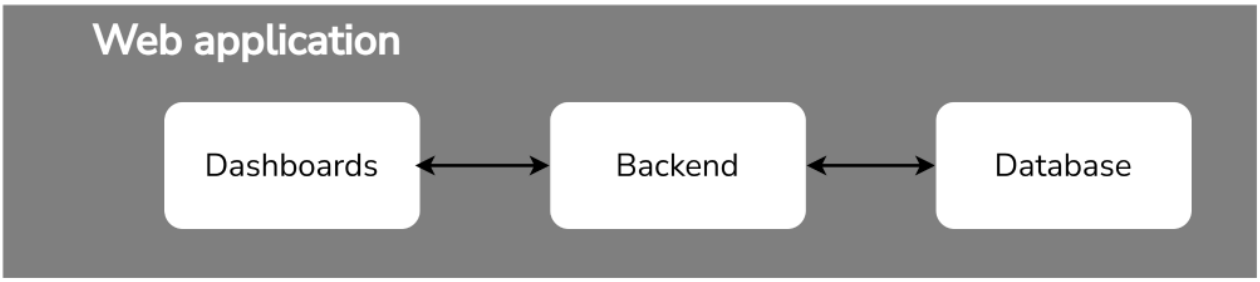

Figure 11 – Web Application

[36] Differently from the datafiles, which contains the actual data, the artefacts are used to store an auxiliary information, produced with the data processing workflows, such as logs, metrics, analysis reports, supporting intermediate datasets, etc.

The **main components** of the Web Application are:

- Database;
- Back-end;
- User Interface (UI).

A dedicated **Database** for the Web Application, different from the Data Lakehouse, is used to store a subset of the collected and analysed data, to provide low-latency access to the data of interest. Furthermore, standardised data schemas are defined to ensure consistency and facilitate seamless handling. These schemas establish a consistent structure that governs the organisation and representation of the data within the AI4TRUST Platform. The **Back-end** component contains the implementation of APIs and the integration of necessary logic to support the **UI**. This implementation enables users to interact with the data ensuring a seamless and user-friendly experience.

The Web Application is developed following the **progressive approach** mentioned in Section 3.1, through **three milestones**, each integrating and expanding upon the previous version. In the following, the new sections of the UI that are added at each version of the Platform are shown, while all preceding sections are still present and updated between versions:

1. **AI4TRUST Platform v1.** In AI4TRUST Platform v1, the Web Application provides users with a Toolbox based on the AI-driven data analysis methods, as defined in Section 3.2.1, designed to tackle disinformation online. Users interact with individual tools for tasks such as reverse video search on the Web, deepfake image, video and audio detection and check-worthy claim detection. An example of this UI is illustrated in Figure 12. For additional information about the Web Application of AI4TRUST Platform v1, please refer to D5.5[37].

[37] D5.5 - AI4TRUST Platform v1 (https://aiAI4TRUST Platformdeliverables/)

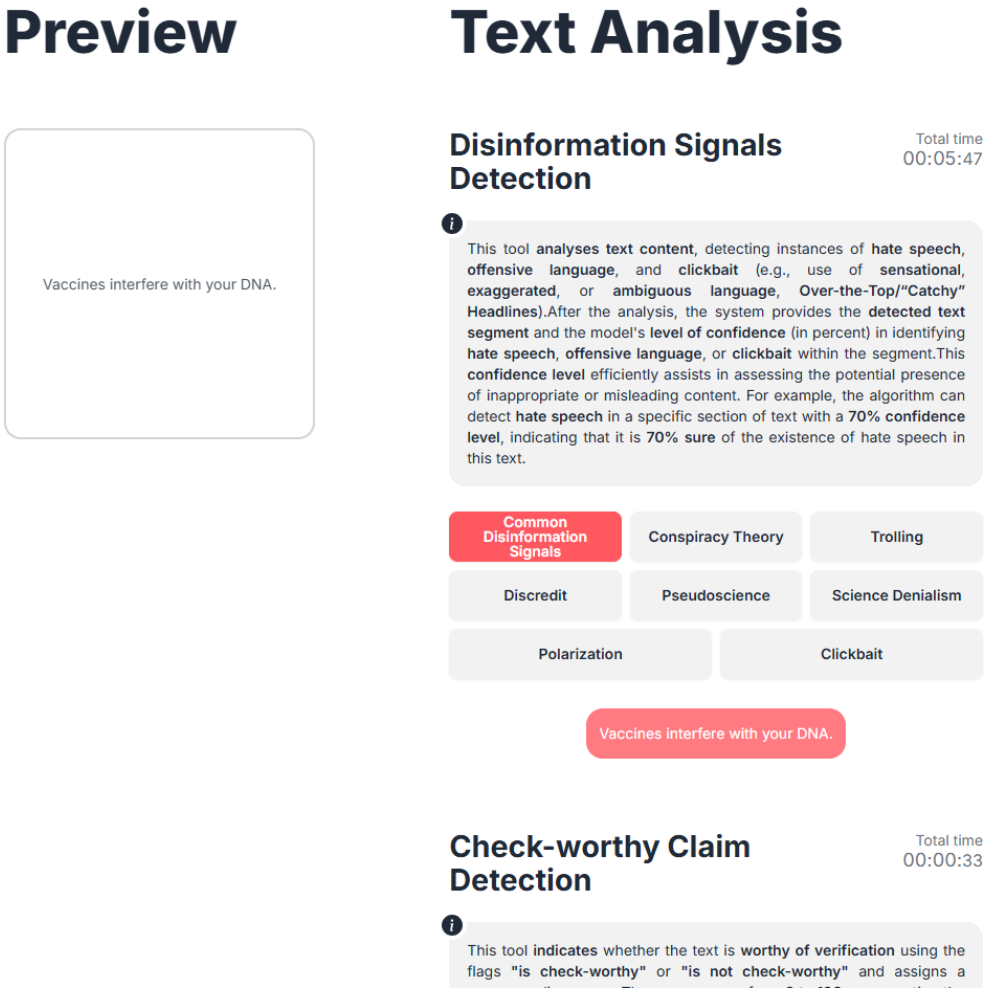

Figure 12 – Textual analysis

2. **AI4TRUST Platform v2.** Extending AI4TRUST Platform v1, the Web Application of AI4TRUST Platform v2 introduces a Monitoring and Human Validation Dashboard that acts as a bridge between end users and the automated data collection and analysis of content (see Section 3.2.2). This system enables continuous tracking and assessment of content, integrating nearly real-time insights derived from AI analysis. Additionally, it incorporates human validation, allowing fact-checkers to assess and validate DWS-analysed content, enhancing the reliability and interpretability of the analyses. The integration of human validation ensures that the system leverages both automated and expert-driven approaches, reinforcing the credibility and trustworthiness of the platform's outputs. Through the Monitoring and Human Validation Dashboard users can also leverage the previously mentioned Toolbox, applying the AI-driven data analysis methods (see Section 3.3). This integration allows for a more in-depth analysis of the content shown in the Monitoring and Human Validation Dashboard. An example of the Monitoring and Human

Validation Dashboard is shown in Figure 12. For additional details about the Web Application of AI4TRUST Platform v2, please refer to D5.6[38].

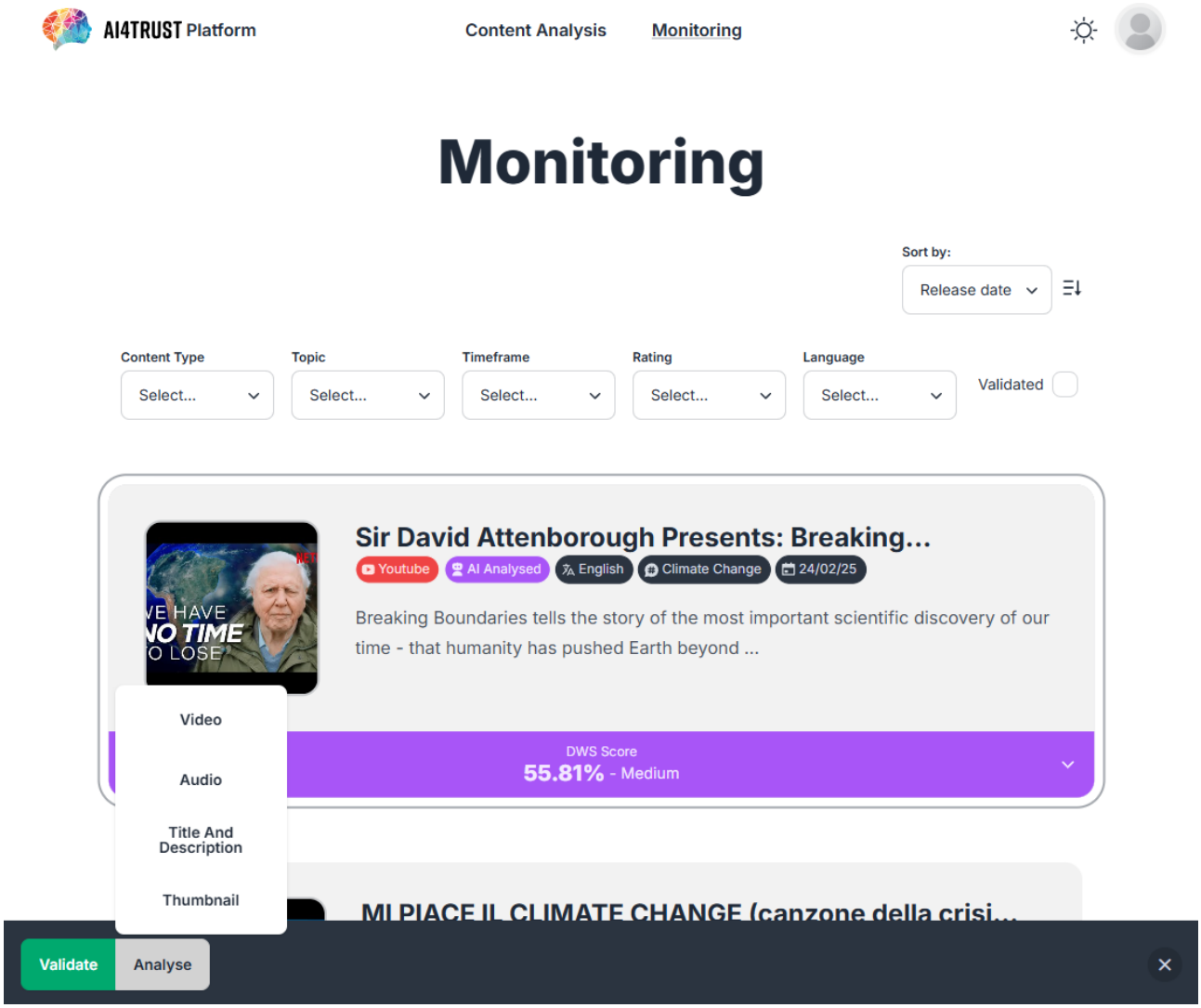

Figure 13 - Second version workflow

3. **AI4TRUST Platform v3.** The AI4TRUST Platform v3 will extend AI4TRUST Platform v2. In fact, the Web Application of AI4TRUST Platform v3 will also integrate the Analytics (Infodemic Observatory, Network Analysis, Recommendation Tool Dashboards) based on the analysis at-scale of collected data defined in Section 3.5.3. The AI4TRUST Platform v3 will consolidate the platform into a unified system where AI-driven data analysis methods, automated data collection and analysis of content and Analysis at-scale of collected data will seamlessly converge. For further details about the Web Application of AI4TRUST Platform v2, please refer to D5.7[39].

[38] D5.6 - AI4TRUST Platform v2 (due by M26 - February 2025, https://aAI4TRUST Platform-deliverables/)
[39] D5.7 - AI4TRUST Platform v3 (due by M38 - February 2026, https://ai4trust.eu/public-deliverables/)

This **incremental addition** of functionalities ensures a **structured, scalable, and user-centric evolution of the Platform**, progressively enhancing its ability to address disinformation challenges effectively.

3.8. API Layer

The **API Layer** functions as a standardized access point for **AI4TRUST functionalities**, ensuring **secure and controlled** interaction with the platform's multiple services. It enforces **multiple authentication mechanisms** (e.g., **JWT and API Key**), restricting external access and regulating **data request types and volumes**. By providing a **single-entry point** for all tools and services developed by project partners, the API Layer **enhances security** and **simplifies interactions**.

Additionally, the API Layer **ensures consistency** across integrated services by standardizing **field names, error handling, and API structures**. It consolidates multiple endpoints, simplifying API management and reducing complexity for external users. Serving as an intermediary between **external third parties and the internal S3-like API**, it enables **researchers and consortium partners** to retrieve datasets while ensuring compliance with **content usage and sharing policies**.

Following the **progressive development approach** outlined in **Section 3.1**, the API Layer evolves with each **project milestone**, integrating **new features and functionalities**. This ensures **seamless communication** between the platform's **front-end and back-end components**, fostering a **flexible and scalable architecture**. At each stage, the API Layer expands to accommodate additional **tools, services, and workflows**, while maintaining a **coherent and structured** interface aligned with the platform's objectives. For a more detailed understanding of how the API Layer has evolved and continues to support the Platform's growth, please refer to D5.5[40], D5.6[41] and D5.7[42].

4. Integration Methodology

The AI4TRUST project utilises **an iterative and incremental approach**, following **agile methodology principles** to emphasise **flexibility, collaboration, and customer satisfaction**[43]. This **human-centred method** focuses on understanding and addressing the needs, wishes, and limitations of end users.

[40] D5.5 - AI4TRUST Platform v1 (https://ai4trust.eu/public-deliverables/)
[41] D5.6 - AI4TRUST Platform v2 (due by M26 - February 2025, https://ai4trust.eu/public-deliverables/)
[42] D5.7 - AI4TRUST Platform v3 (due by M38 - February 2026, https://ai4trust.eu/public-deliverables/)
[43] https://agilemanifesto.org/

4.1. Development Platform

To enable simultaneous work among partners and developers, a **Version Control** is employed to track changes made to files. Specifically, a **Distributed Version Control** is used, which mirrors the entire codebase on every developer's computer[44], allowing offline work and the use of private repositories. Git[45], a Distributed Version Control system, is used for its key features such as: 1) Support for non-linear development through branching/merging and history tracking; 2) Handling local and remote repositories as branches that can merge with each other; 3) Compatibility with various protocols (e.g., HTTPS, FTP, SSH).

During development, the code is hosted on a **web-based Git repository manager**, which facilitates **source code management (SCM)**, **continuous integration (CI)**, **Docker registry**, **issue tracking**, and additional development tools. This infrastructure ensures **efficient version control**, **automated testing and deployment**, and **collaborative development** within the consortium.

4.2. Development Guidelines

To facilitate collaboration and ensure compatibility during environment setup, the following **guidelines** are defined for the AI4TRUST technical partners:

- **Conventional Commits:** Changes in repository files are committed, creating a snapshot of the repository. Each commit is recorded into a branch and described using the Conventional Commits[46] specification, providing clear commit history and simplifying navigation and management of commits and releases.
- **Versioning:** Tags are associated with commits when a component is ready for release or testing. Semantic Versioning is used to ensure coherence between components, as shown in Figure 14.

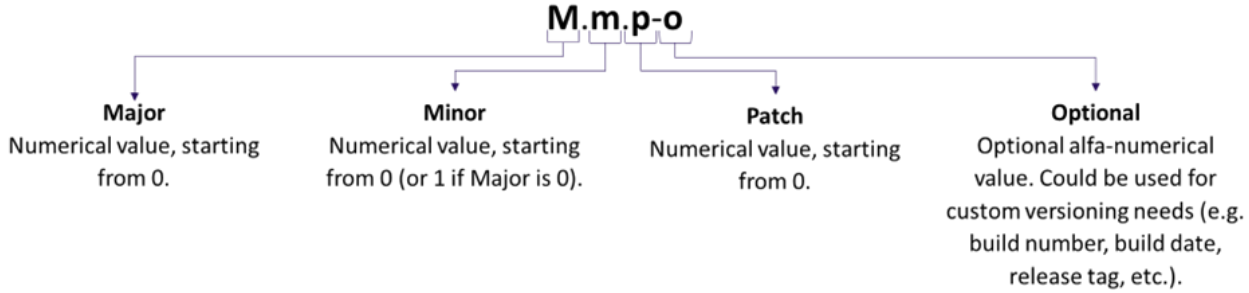

Figure 14 - Semantic Versioning structure

[44] https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control
[45] https://git-scm.com/
[46] https://www.conventionalcommits.org/

- **Branching:** Git supports creating branches to allow commits without affecting the main branch's functionality. The following branch structure for each repository is recommended:
    - **Main branch (master):** The primary branch, always in a deployable and production-ready state. Release versions are tagged following the Versioning Pattern defined above.
    - **Development Branches (dev-featureName):** Temporary branches for feature development. Multiple development branches can exist simultaneously. Upon completing a feature, **pull** any changes from the master branch and manually **merge** them into the dev-*featureName* branch if necessary. After a successful merge and testing, **push** the dev-*featureName* branch to master, then delete the dev-*featureName* branch. It is a good practice to **test** the component in the development branch before merging it to master.
- **Issue Tracking:** Issues track work, report bugs, or suggest features, with commits related to issue resolution documented in the commit message using the pattern "Close #" followed by the issue number (e.g., "Fix random disconnections (Close #127)").

4.3. Containerisation and Container Registry

AI4TRUST employs **Containerisation** for component development and **microservices deployment**. This approach enables rapid building, testing, and deployment of applications consistently across different environments by packaging software into **standardised units**. These containers encapsulate all necessary code and dependencies, ensuring **OS-level isolation**. In contrast to **virtual machines (VMs)**, which virtualise hardware, containers abstract the operating system, making them usable on both physical machines and VMs. Although containers and VMs can be used together, with containers running on either physical or virtual machines, containers primarily focus on meeting the needs of the **Application Layer** rather than the underlying hardware. To support seamless environment migration, **containerised images** are made available. These images are stored and distributed via a **Container Registry**, serving as a centralised repository for containerised images. For further details about the deployment of the AI4TRUST containers and their integration, please refer to D5.6[47] and D5.5[48].

4.4. API-First Approach

Developing different components requires a **shared communication language**, provided through **Application Programming Interface (API)**. Each component must have an API that clearly defines its functionalities. The API-First approach[49], where consistent and reusable APIs are developed

[47] D5.6 - AI4TRUST Platform v2 (due by M26 - February 2025, https://ai4trust.eu/public-deliverables/)
[48] D5.5 - AI4TRUST Platform v1 (https://aiAI4TRUST Platformdeliverables/)
[49] Understanding the API-First Approach to Building Products

using an API description language like OpenAPI, is recommended. This approach establishes a contract indicating the API behaviour and offers several advantages for the different project partners:

- **Parallel development:** Establishing a contract between services allows different teams to work on multiple APIs simultaneously. Developers can mock the API and test their components based on the defined API contract without waiting for the entire API logic to be developed.
- **Cost reduction**: Reusing and standardising API definitions across different services reduces development costs.
- **Faster time to market**: Automating API implementation and documentation based on the defined contract accelerates development cycles and thus speeds up time to market.
- **Improved developer experiences**: Well-designed, well-documented, and consistent APIs enhance developer experience by making it easier to reuse code and onboard new developers, reducing the learning curve.
- **Reduced risk of failure**: Ensuring APIs are well defined, reliable, consistent, and easy to use minimises the risk of failure.

4.5. API Definition

OpenAPI[50] is used to define interactions between AI4TRUST components. OpenAPI is a standard for describing and documenting HTTPs APIs, providing a machine-readable specification that can be used to generate documentation and is supported by a wide range of tools and platforms. An OpenAPI definition is a file written in YAML or JSON that describes a system's complete API, including:

- **Endpoints:** The specific paths (e.g., /users) and the different operations available on each endpoint (e.g., GET /users, POST /users).
- **Parameters:** Input and output parameters for each operation.
- **Authentication:** Methods required for authentication.
- **Metadata:** Additional information such as contact details, terms of use, and other information.

Figure 15 illustrates part of the **AI4TRUST API Interface**, which is generated from a YAML file that complies with the **OpenAPI v3 specifications**. The figure showcases various paths (endpoints) exposed by the API, relative to the base server URL, along with the corresponding operations (such as GET, POST, DELETE) that are used to interact with these paths. This representation provides a

[50] https://www.openapis.org/

clear overview of the structure of the API, highlighting how different HTTP methods are employed to manage and access resources within the Platform.

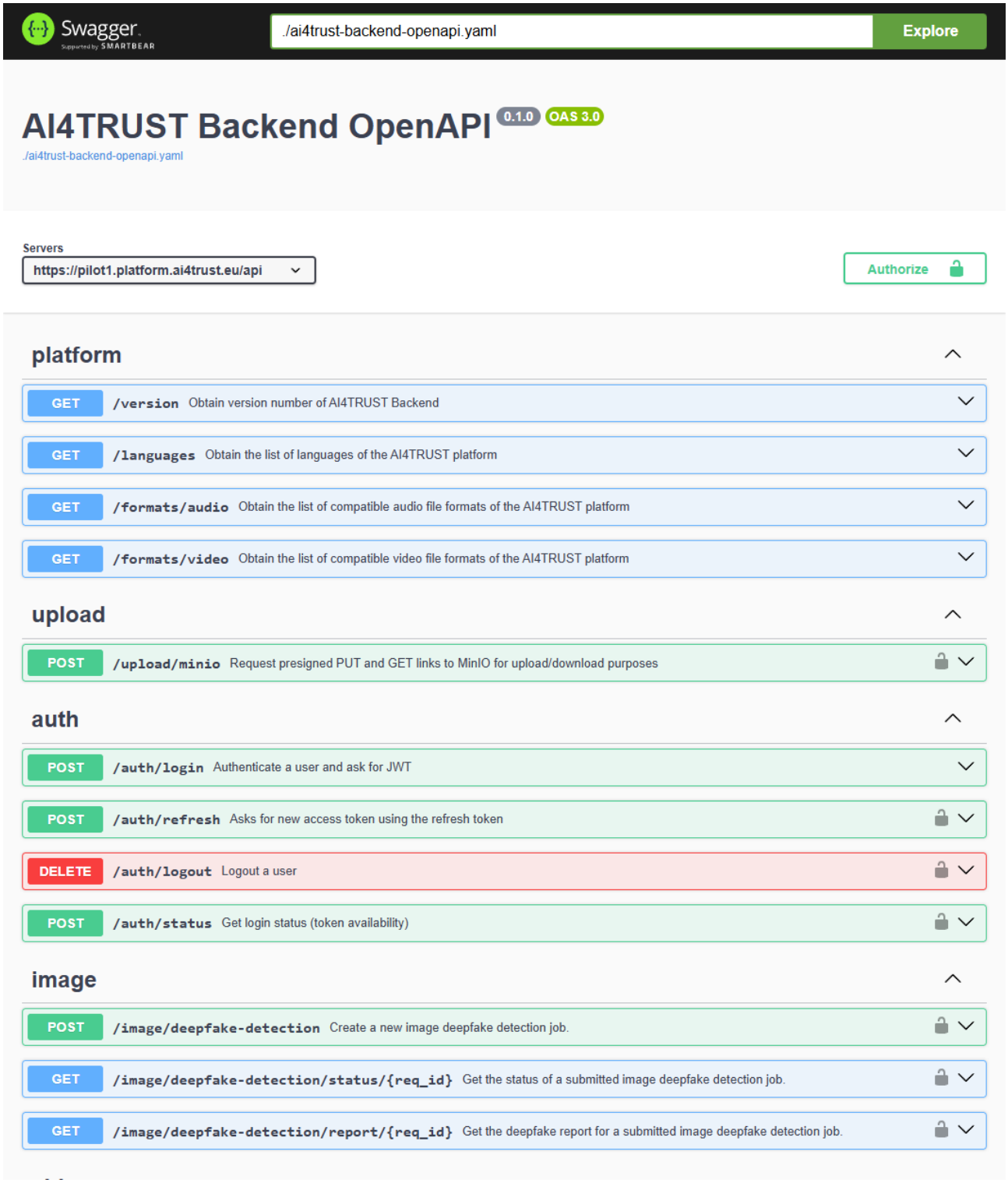

Figure 15 – OpenAPI AI4TRUST definition example

5. Ethics, Security, and Privacy Implications

Points #13 and #15 of Table 2 - Requirements List in Section 2 outline the AI4TRUST requirements in developing a proper methodology to ensure **legal, security and ethical compliance**, along with requirements related to **AI explainability**. The legal, ethical and security implications of this project extend to all aspects of data collection, processing, storage, internal and potentially external access control and credentials, as well as data destruction and/or re-use. **Section 3.3** provides a more comprehensive exploration of deployment issues in this regard. Throughout the project, **maintaining ethical and legal compliance, along with ensuring strong security, is of utmost importance**. To accomplish this, engineers and the ethics-legal team actively engage in continuous and proactive discussions, initiated from the project's outset and sustained throughout its whole duration. In addition, on-demand and regular interaction with **the internal legal-ethics team and the external Ethical Advisory Board** is facilitated. The internal team also participates in the sessions of other work packages (WPs) to offer input as needed, contribute to collaboration out of our silos and cross-WPs understanding. **Compliance is monitored and discussed** on a regular basis. This provides an additional opportunity to review the wider ethical-legal landscape and potential issues that require further reflection, such as any risks arising from the changing external political environment.

The importance of ensuring **explainability of integrated AI solutions** is addressed in these discussions recognising the changing regulatory landscape and the real-world concerns of civil society. The **ethical perspective** highlights the importance of incorporating EU values by design and considers the subsequent implications and **societal impact** of any proposed solution. It is recognised that society is heterogeneous and in terms of the development of methodologies, this diversity is reflected in the challenges posed by multiple languages for fact-checkers working on identical issues concerning the problem of classification of online disinformation and misinformation. The act of categorising online information presents indeed challenges that stem from the **diverse cultural interpretations of terms** that cannot be easily translated “literally - word by word”. These challenges highlight the need to establish a shared understanding of the specific meaning, usage, and related synonyms of these terms. In order to address this, it is crucial to consider the significant legal, security and legal matters that arise from European Union (EU) legislation and regulation. It is advised to refer to **Deliverable 1.2 - Data Management Plan** alongside this document to fully grasp these key issues.

The AI4TRUST project then sheds light on concerns that affect **public trust in the digital world**, as seen through the lens of civil society. The primary concern for civil society is to discern trustworthy **technologies** and ascertain their **reliability, authenticity, credibility, and neutrality**. Moreover, key security, ethical and privacy implications are not sufficiently captured by existing legislative and regulatory frameworks but are embedded in a **fluid and constantly evolving regulatory landscape**, of which the project is mindful. They are, however, expressed in legal requirements (most notably the GDPR) regarding the use of personal data, with or without the subject’s explicit consent; data storage; data destruction; data corruption, onward sale and re-use

(either in full or in part), purpose specification, purpose limitation and the potential impact on the personal autonomy, dignity, privacy and integrity of the individual whose data is being processed.

AI4TRUST ensures that the handling of all datasets created, processed, or re-used are informed by guidelines for **FAIR data management**. Consequently, personal data processing is aligned with **Regulation (EU) 2016/679 (GDPR)**. Briefly, the GDPR is concerned with data that can directly or indirectly identify an individual. It clearly differentiates pseudonymised data (personal data) from anonymised data (not personal data because it undergoes irreversible changes to prevent identification). AI4TRUST project specifically utilises **pseudonymised, anonymised, synthetic data** which minimises the identification of individuals and/or their activities. Each partner is thus responsible for ensuring the application of the GDPR and regularly reviewing, assessing, and updating their legal and ethical compliance, overseen by SAHER Europe. Each partner is moreover responsible for following ethics, security and data protection at their premises. Accordingly, the **AI4TRUST Data Management Plan (D1.2)** determines the data points at which any transition from pseudonymised to anonymised data may be undertaken for both internal and external use. This complies with GDPR requirements.

The AI4TRUST Consortium has reached a consensus on implementing a range of technical and organisational measures, clearly defined in the **Consortium Agreement** (CA - Annex to D1.1 - Project Management Plan):

- **Pseudonymising online data where anonymisation is considered impossible.** Therefore, the following subsequent measures are implemented: (i) perform the research, specifically by collecting and utilising online data in the most minimal and necessary manner (referred to as the 'Data Minimisation Principle'); (ii) secure that online data originates from public sources; (iii) ensure that all legal requirements are implemented, including requirements stemming from the GDPR and from national legislation. This entails establishing an effective legal basis for processing the data. Additionally, prior to sharing any personal data, it is necessary to execute appropriate data sharing agreements, such as Joint Controllership Agreements or Data Processing Agreements.
- **Discussing other operational, managerial, and technical measures** to be implemented.
- **Monitoring in a continuous manner the compliance in the use of personal data.** This includes a careful assessment of all data sets, to assess the risk of re-identification and whether they can be deemed anonymised based on legislation and state-of-the-art literature.

More specifically, the GDPR clearly prioritises the use of **pseudonymisation as a valuable tool** for data protection and management purposes. Pseudonymisation indeed renders data records unidentifiable, effectively reducing the risk of unauthorised access or exposure of sensitive information. However, it also allows authorised data processors and controllers to access and manage the data. **The AI4TRUST project accords with the GDPR's recommendations on using pseudonymised data:**

- **Art. 4(5) GDPR** defines pseudonymisation as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of

additional information provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".

- **Art. 6(4) GDPR (Recital 26)** permits pseudonymised data to be reversed with a proper key by authorised personnel. Unintentional pseudonymised data leakage may incur legal consequences for the host organisation and have harmful, adverse effects on the individual human subject. Therefore, guaranteeing the security of pseudonymised data must be a high priority for those using and/or sharing it for whatever purposes.
- **Recital 28 GDPR**: "The application of pseudonymization to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations".

In summary, the GDPR considers **pseudonymisation as a suitable method** to achieve the following objectives: (i) safeguarding data subjects and data controllers while ensuring compliance; (ii) protecting personal data on legacy systems against unauthorised access; (iii) keeping a backup to temporary store original data when personal data is being anonymised. It ultimately signals data protection by design and default (as required by Art. 25 GDPR). **Data subjects' right to be forgotten** is also seen to be enhanced by pseudonymisation, especially when automated. The processes for complying with this are to be addressed.

In addition, the GDPR expects data controllers and organisations to: (i) **protect individuals' data** and places responsibility on them for evaluating risks (under Recital 83, committing them to implement **mitigating measures, such as encryption**); and (ii) **combat risks by ensuring a high level of security** that is suitable for the identified risks, as required by Article 32 of the GDPR to ultimately ensure **confidentiality, integrity and resilience processing**. In order to accomplish these objectives, the AI4TRUST project focuses on addressing and continuously evaluating the requirements for extracting data both for and from fact-checking purposes, as well as from processing.

Ethical considerations and adherence to the values and principles of the European Union serve as the foundations for all EU funded projects. Consequently, the processing of data must adhere to the highest and rigorous ethical standards and the applicable EU, international and national law on ethical principles. Ongoing **AI4TRUST ethical reflection expands upon legal compliance** and involves the application of a series of standard questions as well as more specific questions related to its Work packages and Tasks. These relate to the use of information and context specification; privacy; human autonomy and integrity (e.g., proportionality of AI apps); control, influence, and power (e.g., transparency, accountability, responsibility); impact on social contact patterns; gender, minorities, non-discrimination, bias and justice; human values; sustainability; bias and viability.

Partners are thus required to **assess the potential risks associated with data processing activities**, particularly those involving large-scale processing of personal data or extensive monitoring of publicly accessible areas. AI4TRUST personal data processing are specifically safeguarded by: (i) identifying the kind of personal data collected, the partners involved in the processing of personal

data, as well as the legal basis declared by partners; (ii) implementing technical and organisational measures; and (iii) data processing agreements where needed.

Whereas the legal and regulatory frameworks on pseudonymisation and anonymisation are specific, **new challenges arise from the ethical and security perspectives** from the opportunities and potential for harm arising **from the deployment of AI**. Data masking for instance is seen as a form of pseudonymisation but must be strict in terms of the fields exposed to ensure that the data cannot be associated with a specific individual. Critics note that "pseudonymisation may permit identification using indirect means. When a pseudonym is used, it may be feasible to identify the individual concerned by the data by analysing related data". This point holds great significance as it encompasses ethical requirements regarding the right of the individual human data subject to integrity, autonomy, dignity as well as privacy. It also raises issues regarding data minimisation, purpose specification and limitation and non-linkability and reuse for unknown and/or unspecified purposes, again intrinsic to ethical considerations.

The AI4TRUST Platform employs a range of methodologies and technological measures to **guarantee adherence to the above-mentioned privacy and data protection principles**, outlined by the GDPR:

1. The **Data ingestion phase** (as described in **Section 3.4** and to be implemented in Task 2.2) is conducted in a privacy-compliant and secure manner, aligning with the ethical requirements outlined in the AI4TRUST D1.2 - Data Management Plan. In fact, to ensure the safeguarding of user identification, pseudonymisation techniques, such as hashing, will be effectively employed;
2. In the **Data elaboration phase** (described in **Section 3.5**), the roles of "controllers" and "processors" are specified in compliance with GDPR regulations. The "controllers" are partners who are responsible for ensuring data processing compliance, while the "processors" are partners authorised to process personal data.
3. **Data access** (as described in **Section 3.6**) is managed through the implementation of appropriate authentication and authorization techniques. In order to achieve this, every stakeholder within the Consortium is granted access to one or more interfaces. This access is managed to ensure the proper handling of credentials, authentication, and authorisation. It is essential to emphasise that this management also encompasses a diligent adherence to data licences, usage rights, and privacy concerns.
4. **Social media data exposed** to end-users (as described in **Section 3.6**) undergoes pseudonymisation to prevent user identification - this is implemented also thanks to the pseudonymization techniques adopted in the Data ingestion phase. In fact, the user identity information is actually determined by the data imported during the Data ingestion process. If the data is already pseudonymised at the source, there is no need to implement additional techniques when generating the output.

Ultimately the AI4TRUST project lifecycle aligns with and closely corresponds to **significant legislative and regulatory initiatives undertaken within the EU, including the AI Act**. In

accordance with the AI Act, the AI4TRUST Platform has been designed to ensure **AI robustness** by adhering to **strict risk management and safety requirements for AI systems**, particularly those classified as high-risk. The Platform employs rigorous testing and validation protocols to assess the reliability and resilience of AI models, ensuring they operate effectively under varied conditions. Furthermore, AI4TRUST integrates **mechanisms for continuous monitoring to detect and mitigate potential risks** throughout the lifecycle of the AI systems. By implementing these safeguards, the Platform ensures compliance with the AI Act's provisions, particularly regarding the **minimisation of harm, prevention of biases, and protection of individuals' rights and freedoms**. The robustness of the system is essential not only for regulatory compliance but also for fostering trust among end-users and stakeholders. Moreover, special consideration will be given to related and pertinent legislation, as well as the positions and recommendations made by the **European Data Protection Supervisor (EDPS)** and other relevant civil society organisations at the European Level.

Specifically, the **AI4TRUST Platform** and its tools are designed to uphold the principles of **robustness**, **security**, and **trustworthiness** as outlined in the **AI Act**, ensuring resilience against adversarial attacks, reliability in dynamic information environments, and transparency in AI-driven decision-making. Given its role in processing vast amounts of **online social data, news feeds, and multimedia content**, the Platform integrates **fault-tolerant AI models** that can **filter, classify, and analyse content** while mitigating risks of bias, mis/disinformation, and manipulation.

Robustness is embedded at multiple levels. The **incremental development** of the Platform across three versions strengthens its resilience, progressively enhancing automated content verification, disinformation detection, and advanced analytics capabilities. AI models are optimised for **real-time and batch processing**, ensuring adaptability to evolving information landscapes. **Continuous monitoring** and **human oversight**, through media professionals, fact-checkers, and IT/social science experts, reinforce the system's ability to detect and address inaccuracies, while also contributing to the refinement of explainable AI solutions.

To safeguard against adversarial threats, the Platform employs **secure data handling mechanisms**, leveraging **state-of-the-art AI pipelines** that integrate multimodal, multilingual AI-driven data analysis methods. The inclusion of **human validation** and **interactive dashboards** ensures that insights provided to journalists, fact-checkers, policymakers, and researchers remain **accurate, transparent, and actionable**. By adhering to the AI Act's regulatory framework, **AI4TRUST not only meets compliance standards but also establishes a resilient and ethical foundation for AI-driven fact-checking and mis/disinformation monitoring.**

In this regard, the project acknowledges that **external political developments may pose challenges and potential risks** to maintaining a consistent commitment to implementing both the spirit and the letter of the anticipated privacy-, data protection, and AI-related regulations. While discussions on a possible weakening of compliance efforts take place externally, **AI4TRUST remains steadfast in its original commitment** to adhering to best practices and upholding the **highest standards of compliance**.

6. Conclusions and Next Steps

The initial version of **D5.4 - AI4TRUST Platform Specification** laid the foundation for the architectural overview of the **AI4TRUST Platform**. The recommendations outlined in the "**General Project Review Consolidated Report (HE)**," dated **28 June 2024**, following the project's first Review Meeting, have played a pivotal role in shaping the revisions presented in this **D5.8 - AI4TRUST Platform Specification – Revised Version**. Specifically, this deliverable incorporates an **enhanced description** of the overall architecture and data flow, illustrating the integration and communication among the various components.

Furthermore, it provides **additional details** regarding the **Platform Roadmap**, highlighting how each version of the Platform builds upon its predecessor. In this context, the **AI-driven data analysis methods** implemented in **AI4TRUST Platform v1** are further utilised in **AI4TRUST Platform v2**, which extends the Platform with **automated data collection** and **content analysis**. Additionally, **AI4TRUST Platform v3** introduces **analysis at-scale** of collected data, combining both the **AI-driven data analysis methods** and the **automated data collection and content analysis**. This **step-by-step progression** ensures that each version of the platform builds upon the previous one, consistently enhancing its capabilities to effectively address the issue of **digital disinformation**.

While the overview provided in this deliverable presents a simplified representation of the interaction between the components, the **complex orchestration** implemented in the backend by the **system integrator FINCONS** is made possible by a significant integration effort, enabling the different components to communicate effectively. This intricate integration process will be thoroughly detailed on **D5.6 - AI4TRUST Platform v2**, due at **M27 (March 2025)**. Further advancements of this implementation will be outlined in **D5.7 - AI4TRUST Platform v3**, due at **M38 (February 2026)**.

References

- Art. 29 Data Protection Working Party (2017) Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679
- Art. 29 Data Protection Working Party (2014), Opinion 05/2014 on Anonymisation Techniques. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
- ENISA (2019) Pseudonymisation Techniques and Best Practices, https://www.enisa.europa.eu/news/enisa-news/enisa-proposes-best-practices-and-techniques-for-pseudonymisation, https://enisa.europa.eu/topics/cybersecurity-policy/data-protection
- European Commission (2022) The 2022 Code of Practice on Disinformation https://digital-strategy-ec.europa.eu/en/policies/code-practice-disinformation
- EPRS (2021) Regulating facial recognition in the EU, Brussels. https://europarl.europa.eu/RegData/etudes/IDAN/2021/698021/EPRS_IDA(2021)698021_EN.pdf
- European Union, Charter of Fundamental Rights of the European Union, 26 October 2012, 2012/C 326/02 https://www.refworld.org/docid/3ae6b3b70.html
- EDRI (2022) Position Paper Respecting Fundamental Rights in the cross border investigation of serious crimes. https://edri.org/wp-content/uploads/2022/10/EDRi-position-paper-Respecting-fundamental-rights-in-the-cross-border-investigation-of-serious-crimes-7-September-2022.pdf
- European Group on Ethics in Science and New Technologies (2018) An Ethical, societal, and fundamental rights dimension for the EU policies, Brussels
- European Data Protection Board (2022) Guidelines on Art.60 (GDPR), Guidelines on dark patterns in social media platforms interfaces, toolbox on essential data protection safeguards for enforcement cooperation between EEA and third country SAs, https://edpb.europa.eu/news/2022/edpb-adopts-guidelines-art-60-gdpr-guidelines-dark-patterns-social-media-platform-en
- General Data Protection Regulation (GDPR) regulation (EU)2016/679 General data Protection regulation OJ L 119,04.05.2016, in force as of 25.05.2018, https://gdpr-info.eu
- Iapp. (2019). Publicly available data under the GDPR: Main considerations. [online] Available at: https://iapp.org/news/a/publicly-available-data-under-gdpr-main-considerations/
- Independent High-Level Expert Group On Artificial Intelligence (AI HLEG) (2020), The Assessment List For Trustworthy Intelligence, Artificial Intelligence (ALTAI) For Self-Assessment, 17 July 2020. Doi:10.2759/002360. See also:

https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

- Privacy International (2023) Artificial Intelligence https://privacyinternational.org/learn/artificial-intelligence