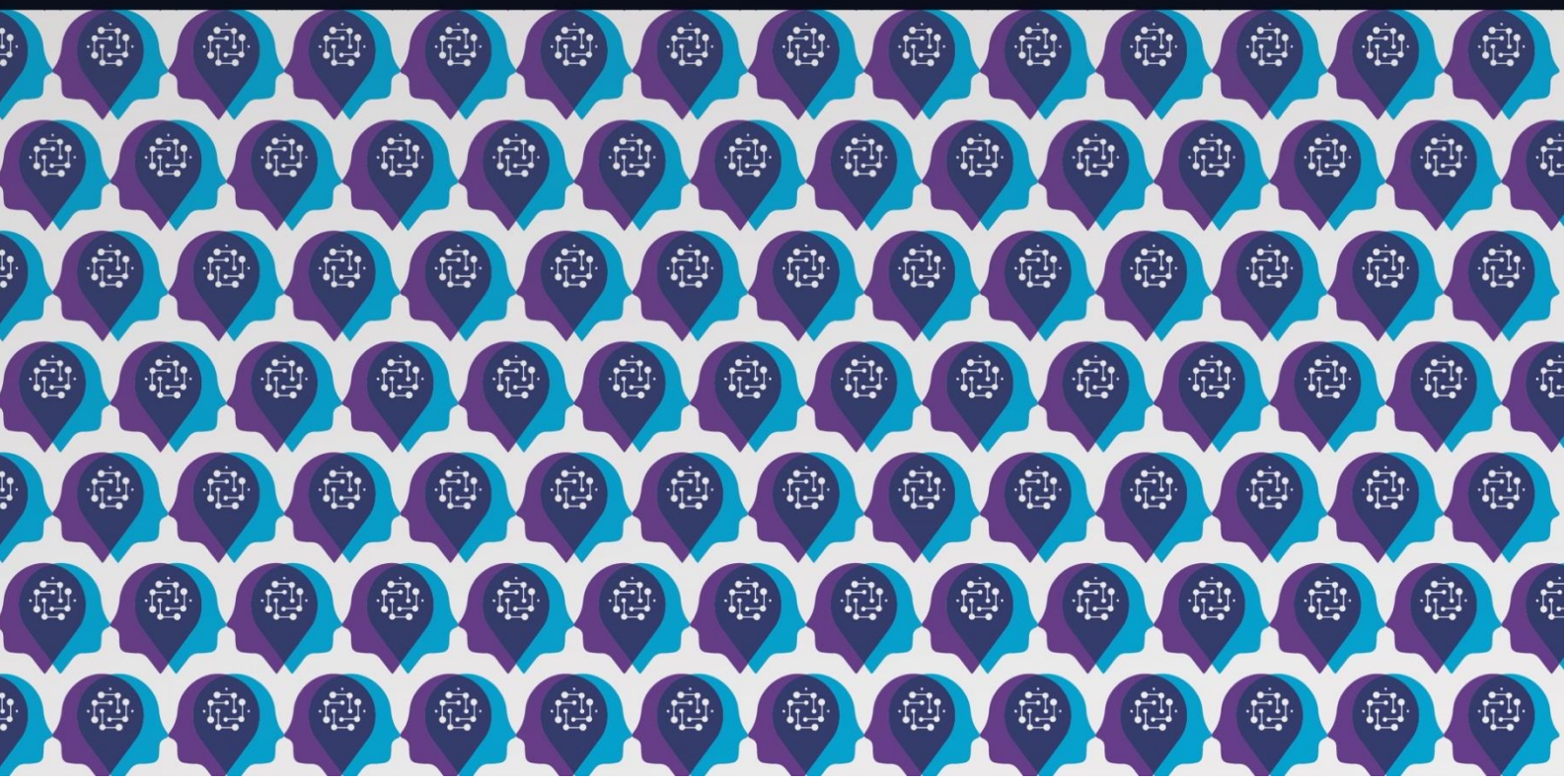
AI4Debunk

D8.5 Initial explainability module tracing back between the data and the score

October 2025

Funded by
the European Union



Grant Agreement No.: 101135757
Call: HORIZON-CL4-2023-HUMAN-01-CNECT
Topic: HORIZON-CL4-2023-HUMAN-01-05
Type of action: HORIZON Innovation Actions

D8.5 INITIAL EXPLAINABILITY MODULE TRACING BACK BETWEEN THE DATA AND THE SCORE

Project Acronym	AI4Debunk
Project Number	101135757
Project Full Title	Participative Assistive AI-powered Tools for Supporting Trustworthy Online Activity of Citizens and Debunking Disinformation
Work package	WP 8
Task	Task 5
Due date	31/10/2025
Submission date	24/10/2025
Deliverable lead	Partner UMONS
Version	1
Authors	Kevin El Haddad (Partner UMONS)
Contributors	Axel Brando (Partner BSC), Alvaro Parafita (Partner BSC)
Reviewers	Marcel Keijzer – Stichting Innovative Power
Abstract	The Horizon Europe-funded AI4DEBUNK project is developing an advanced decision support system to combat digital disinformation. This report details the system's modular, multimodal platform, which integrates specialized components for comprehensive content analysis. Key modules include semantic similarity estimation, a deepfake detection suite (covering text, image, and audio), and cross-modal coherence analysis to validate image-text relationships. The platform aggregates these analyses into a DisinfoScore, a weighted-average metric whose composition provides inherent explainability. We discuss this explainable-by-design architecture and outline future

Funded by the European Union

	work, which involves transitioning to an agentic architecture orchestrated by a Large Language Model (LLM) for enhanced contextual reasoning.
Keywords	Disinformation, Deepfake Detection, Multimodal Analysis, Explainable AI (XAI), Decision Support System, Cross-Modal Coherence, Agentic Architecture

DOCUMENT DISSEMINATION LEVEL

Dissemination level	
X	PU - Public
	SEN - Sensitive

DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributor(s)
0.1	24/10/2025	First draft	UMONS, BSC
0.2	28/10/2025	Reviewed by internal reviewers	IP
0.3	28/10/2025	Version reviewed	UMONS, BSC
1	30/10/2025	Final version	UMONS

STATEMENT ON MAINSTREAMING GENDER

The AI4Debunk consortium is committed to including gender and intersectionality as a transversal aspect in the project's activities. In line with EU guidelines and objectives, all partners – including the authors of this deliverable – recognise the importance of advancing gender analysis and sex-disaggregated data collection in the development of scientific research. Therefore, we commit to paying particular attention to including, monitoring, and periodically evaluating the participation of different genders in all activities developed within the project, including workshops, webinars and events but also surveys, interviews and research, in general. While applying a non-binary approach to data collection and promoting the participation of all genders in the activities, the partners will periodically reflect and inform about the limitations of their approach. Through an iterative learning process, they commit to plan and implement strategies that maximise the inclusion of more and more intersectional perspectives in their activities.

Funded by the European Union

DISCLAIMER

The AI4Debunk project has received funding from the European Union's Horizon Europe Programme under the Grant Agreement No. 101135757.

COPYRIGHT NOTICE

How to cite this report: Kevin El Haddad (2025). AI4Debunk D8.5: Initial explainability module tracing back between the data and the score. https://ai4debunk.eu/wp-content/uploads/2025/11/AI4Debunk-Deliverable-8.5.pdf

The AI4Debunk consortium is the following:

Participant number	Participant organisation name	Short name	Country
1	LATVIJAS UNIVERSITATE	UL	LV
2	FREE MEDIA BULGARIA SRL	EUalive	BE
3	PILOT4DEV	P4D	BE
4	INTERNEWS UKRAINE	IUA	UA
5	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR-IRPPS	IT
6	UNIVERSITA DEGLI STUDI DI FIRENZE	MICC/UNIFI	IT
6.1	CONSORZIO NAZIONALE INTERUNIVERSITARIO PER LE TELECOMUNICAZIONI	CNIT	IT
7	BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION	BSC	ES
8	DOTSOFT OLOKLIROMENES EFARMOGES DIADIKTIOY KAI VASEON DEDOMENON AE	DOTSOFT	EL
9	UNIVERSITE DE MONS	UMONS	BE
10	UNIVERSITY OF GALWAY	UOG	IE
11	STICHTING HOGESCHOOL UTRECHT	HU	NL
12	STICHTING INNOVATIVE POWER	IP	NL
13	F6S NETWORK IRELAND LIMITED	F6S	IE

Funded by the European Union

TABLE OF CONTENTS

Funded by the European Union

LIST OF FIGURES

Funded by the European Union

ABBREVIATIONS

WP	Work Package
AI	Artificial Intelligence

Funded by the European Union

EXECUTIVE SUMMARY

The AI4DEBUNK project is developing an advanced decision support system to help citizens and media professionals identify disinformation. The system is built on a multimodal, modular, and adaptable platform that integrates various specialized analytical components. Key modules include a similarity estimator that compares news against a knowledge base, a comprehensive deepfake detection suite for text, audio, and video, and a cross-modal coherence module to verify image-text alignment. These modules are dynamically activated based on the content of the incoming news item.

The modules can be used individually or to calculate the DisinfoScore, a weighted average of the results from all activated modules. This design provides inherent explainability by transparently showing which factors contributed to the final assessment. Future work under WP9 will transition the platform to a sophisticated agentic architecture. This new model will use a central "Orchestrator Agent," powered by a Large Language Model, to manage the various analysis tools. This agent will also interact directly with the knowledge graph for enhanced contextual reasoning. This evolution aims to deliver a more nuanced, articulate, and trustworthy system for securing the digital information ecosystem.

Funded by the European Union

1 INTRODUCTION

The **AI4DEBUNK** project is dedicated to the development of an advanced decision support system. This system is engineered to assist citizens and media professionals in conducting informed assessments of digital content to identify potential disinformation.

The core of the project is a **multimodal, modular, and adaptable platform**, as depicted in Figure 1. This architecture allows for flexible and targeted analysis by integrating various specialized modules, each designed to perform a distinct function. The current implementation of the platform includes the following core components:

- **Similarity Estimation Module:** This module accepts a given news item as input and queries a dedicated knowledge base to retrieve semantically similar articles. The returned results are classified based on their stance relative to the input, identifying content that either supports or opposes the reference news item. This provides crucial context by linking new information to existing, verified narratives.
- **Deepfake Detection Suite:** This suite comprises a set of specialized tools for detecting AI-manipulated media across different modalities, as highlighted in orange in Figure 1.
    - **Textual Analysis:** This component performs an authenticity verification on text-based content (e.g., articles, social media posts). Its output includes a binary classification or a probabilistic score, supplemented with highlighted phrases or words that warrant further investigation, thereby offering granular explainability.
    - **Image/Video Analysis:** This module assesses the authenticity of visual media. It delivers a multifaceted output, including localization heatmaps to pinpoint manipulated regions, a binary assessment (authentic/manipulated), and a probabilistic confidence score. For video analysis, assessments are conducted on a frame-by-frame basis.
    - **Audio Analysis:** This component processes audio signals, either from standalone files or embedded in videos, to detect synthetic or manipulated speech and other audio artifacts indicative of deepfakes.
- **Cross-Modal Coherence Analysis Module:** This module evaluates the semantic alignment between an image and its accompanying textual caption or description. It quantifies the compatibility of the visual and textual information, outputting a coherence score ranging from 0 to 1, where a lower score indicates a greater semantic distance or mismatch between the two modalities.

Upon receiving an input news item, the platform dynamically activates only the modules whose input requirements are satisfied by the modalities present in the content.

Funded by the European Union

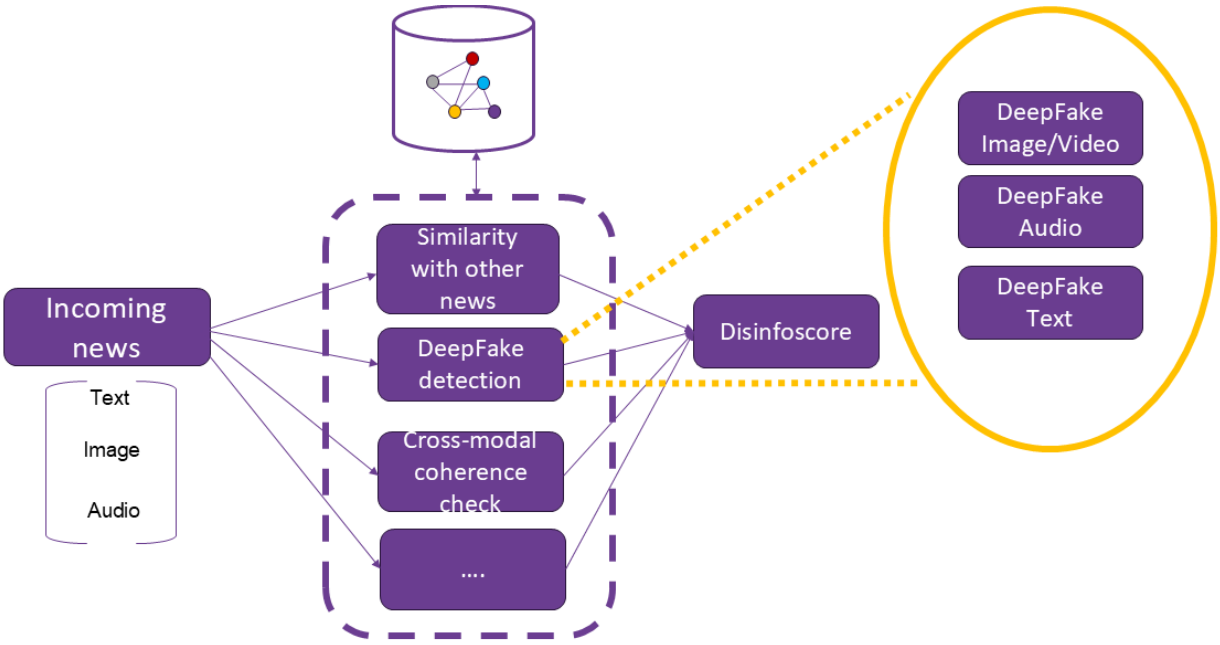


FIGURE 1 AI4DEBUNK PLATFORM

2 DISINFOSCORE CALCULATION AND EXPLAINABILITY

The platform aggregates the outputs from all activated modules into a single, composite metric termed the **DisinfoScore**. This score is calculated as a weighted average of the normalized outputs from each active module, as illustrated in the formula in Figure 2.

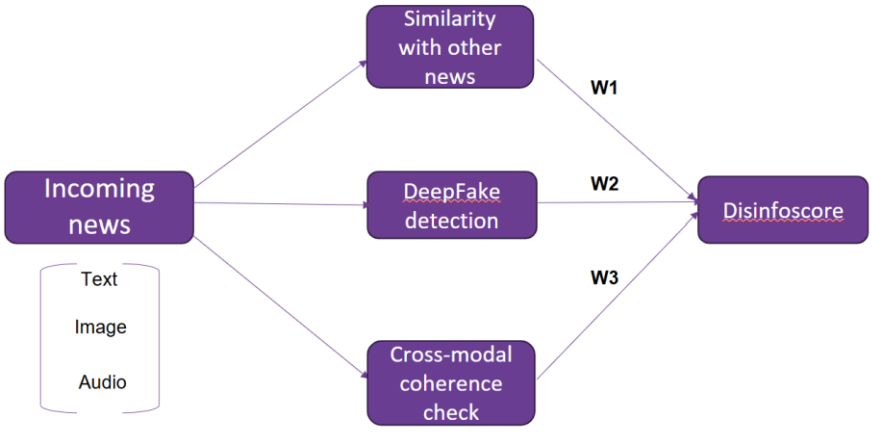


FIGURE 2 DISINFOSCORE CALCULATION

The modular architecture and the weighted aggregation method provide a high degree of **inherent explainability**. The final DisinfoScore is not a black-box output; rather, its composition is transparent. The individual scores from each module serve as direct indicators of their contribution

Funded by the European Union

to the final assessment. This allows users to precisely identify the source of potential disinformation—whether it stems from similarity to known false narratives, the presence of manipulated media by AI (deepfakes), or a logical inconsistency between the text and accompanying imagery (out-of-context information).
This approach ensures that the system's reasoning is traceable and comprehensible, and its design is inherently extensible to incorporate additional analysis modules in the future.

3 FUTURE DIRECTIONS: AN AGENT-BASED ARCHITECTURE

Future development, under Work Package 9 (WP9), will focus on transitioning the platform to a more sophisticated **agentic architecture**. In this paradigm, each analytical module will be encapsulated as an independent "Agent" or "tool." These agents will be managed by a central **"Orchestrator Agent"** powered by a Large Language Model (LLM).
This Orchestrator will dynamically invoke the appropriate tool agents to analyze incoming content, interpret their outputs, and interact directly with the project's knowledge graph for enhanced contextual reasoning. This evolution aims to produce a more nuanced final assessment, complete with a detailed, human-readable explanation of the decision-making process. Key research and development efforts will include the selection, integration, and potential fine-tuning of suitable LLMs to serve as the cognitive core of this advanced agent-based system.

The modular nature of the AI4DEBUNK Platform and the weighted average approach used to calculate the *disinfoscore*, provide by nature explainability of the decision taken. Indeed, the output scores of each module indicate the contributions of each module to the score calculation. These scores are a direct indication of what part of the input news is used to communicate disinformation: similar text that are known to be disinformation (coming from the knowledge graph developed in WP6 for example), deepfake content or Out-Of-Context image-textual description pairs in this case.
This approach generalizes to other module than the ones currently implemented in the AI4DEBUNK Platform.

4 CONCLUSION

The WP8 has successfully built a modular, and multimodal platform for the systematic analysis of disinformation. The current implementation provides a tangible decision-support tool by integrating distinct analytical vectors—semantic similarity, multi-format deepfake detection, and cross-modal coherence.

A key achievement of the current architecture is the `DisinfoScore`, which moves beyond opaque, binary assessments by providing inherent explainability. The transparent, weighted contribution

Funded by the European Union

of each module allows users to identify the specific nature of the disinformation, whether it stems from manipulated media, logical inconsistencies, or relation to known false narratives.

This solid foundation serves as the baseline for the project's next evolutionary phase. The planned transition to an agentic architecture, orchestrated by a Large Language Model, promises to significantly enhance the system's capabilities. This future-state system will not only streamline the analysis process but also introduce a more profound level of contextual reasoning and articulate, human-readable explanations. The AI4DEBUNK consortium is well-positioned to deliver this advanced, adaptable, and trustworthy system, providing a critical resource for media professionals and citizens in the ongoing effort to secure the digital information ecosystem.

Funded by
the European Union

Review Sheet of Deliverable/ Milestone Report

D8.5 Initial explainability module tracing back between the data and the score
October 2025

Editor(s):	Kevin El Haddad
Responsible Partner:	UMONS
Status-Version:	First Draft - v 0.1
Date:	24/10/2025
Distribution level (CO, PU):	Public
Reviewer (Name/Organization)	Marcel Keijzer – Stichting Innovative Power
Review date	28/10/2025

Disclaimer: This assessment reflects only the author's views and the European Commission is not responsible for any use that may be made of the information contained therein"

Funded by the European Union

Mark with X the corresponding column:

Y= yes	N= no	N = not applicable

ELEMENT TO REVIEW	Y	N	N A	COMMENTS
FORMAT: Does the document … ?				
…include editors, deliverable name, version number, dissemination level, date, and status?	X			
…contain a license (in case of public deliverables)?			X	
…include the names of contributors and reviewers?	X			
….has a version table consistent with the document's revision?	X			
… contain an updated table of contents?			X	
… contain a list of figures consistent with the document's content?	X			
… contain a list of tables consistent with the document's content?			X	
… contain a list of terms and abbreviations?	X			
… contain an Executive Summary?	X			
… contain a Conclusions section?	X			
… contain a List of References (Bibliography) in the adequate format, if relevant?			X	
… use the fonts and sections defined in the official template?	X			
… use correct spelling and grammar?	X			
… conform to length guidelines (50 pages maximum (plus Executive Summary and annexes)	X			
… conform to guidelines regarding Annexes (inclusion of complementary information)	X			
… present consistency along the whole document in terms of English quality/style? (to avoid accidental usage of copy&paste text)	X			

Funded by the European Union

ELEMENT TO REVIEW	Y	N	N A	COMMENTS
About the content…				
Is the deliverable content correctly written?	X			
Is the overall style of the deliverable correctly organized and presented in a logical order?	X			
Is the Executive Summary self-contained, following the guidelines and does it include the main conclusions of the document?	X			
Is the body of the deliverable (technique, methodology results, discussion) well enough explained?	X			
Are the contents of the document treated with the required depth?	X			
Does the document need additional sections to be considered complete?		X		
Are there any sections in the document that should be removed?		X		
Are all references in the document included in the references list?			X	
Have you noticed any text in the document not well referenced? (copy and paste of text/picture without including the reference in the reference list)			X	
SOCIAL and TECHNICAL RESEARCH WPs (WP4, 5, 12, 13, 14)				
Is the deliverable sufficiently innovative?			X	
Does the document present technical soundness and its methods are correctly explained?			X	
What do you think is the strongest aspect of the deliverable?			X	
What do you think is the weakest aspect of the deliverable?			X	
Please perform a brief evaluation and/or validation of the results, if applicable.			X	
AI AND TECNOLOGICAL WPS (WP6 – WP11)				

Funded by the European Union

ELEMENT TO REVIEW	Y	N	N A	COMMENTS
Does the document present technical soundness and the methods are correctly explained?	X			
What do you think is the strongest aspect of the deliverable?	X			The "Orchestrator Agent"
What do you think is the weakest aspect of the deliverable?			X	
Please perform a brief evaluation and/or validation of the results, if applicable.			X	
DISSEMINATION AND EXPLOITATION WPs (WP15 – WP17)				
Does the document present a consistent outreach and exploitation strategy?			X	
Are the methods and means correctly explained?			X	
What do you think is the strongest aspect of the deliverable?			X	
What do you think is the weakest aspect of the deliverable?			X	
Please perform a brief evaluation and/or validation of the results, if applicable.			X	

SUGGESTED IMPROVEMENTS

PAGE	SECTION	SUGGESTED IMPROVEMENT
		ADD ROWS AS NECESSARY

CONCLUSION

Mark with X the corresponding line.

X	Document accepted, no changes required.
	Document accepted, changes required.
	Document not accepted, it must be reviewed after changes are implemented.

Please rank this document globally on a scale of 1-5 (1 = poor, 5= excellent) – using a half point scale. Mark with X the corresponding grade.

Document grade	1	1.5	2	2.5	3	3.5	4	4.5	5
									X

Funded by the European Union