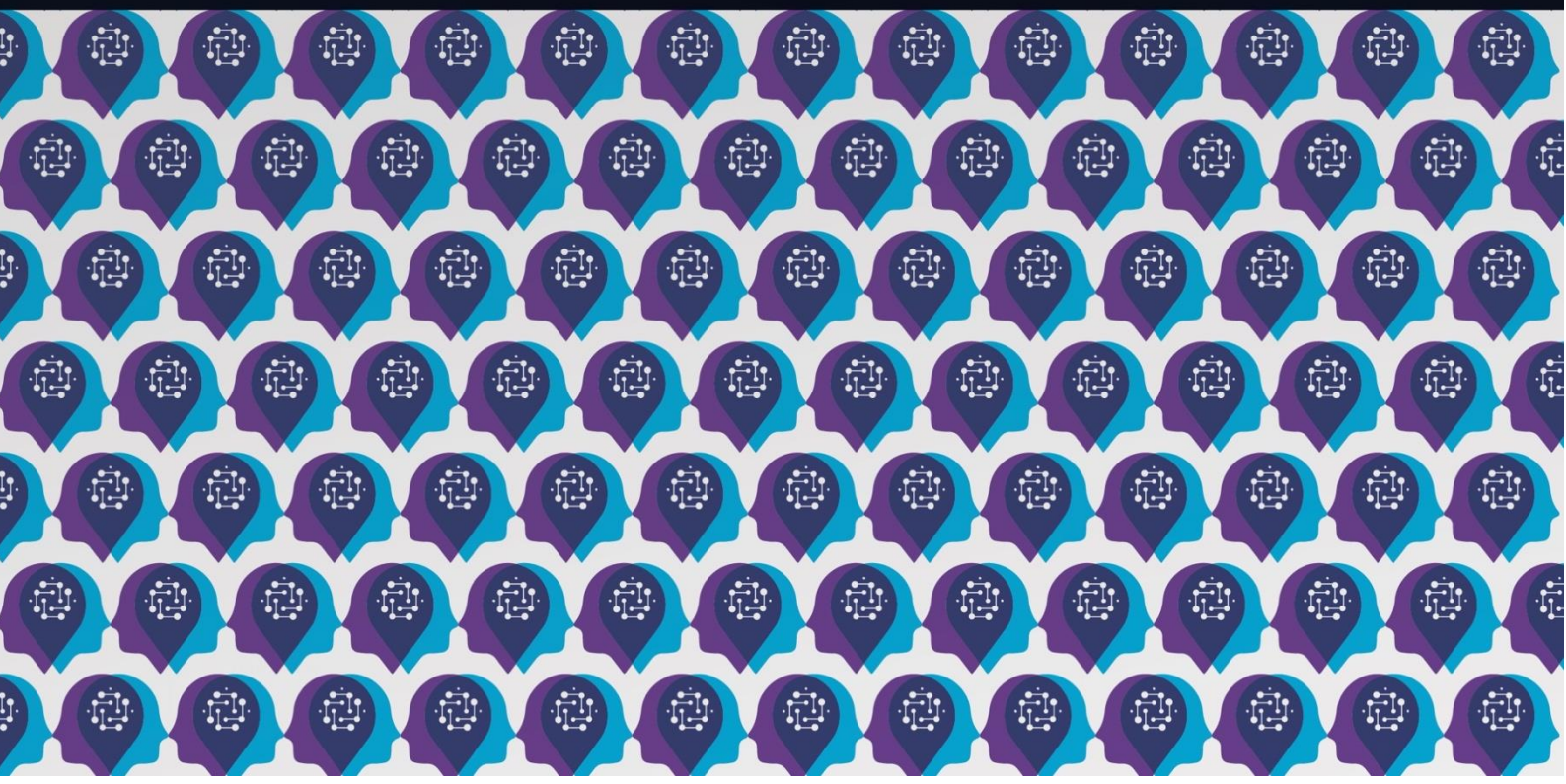
AI4Debunk

D8.4 Initial calculation of a score representing the amount of disinformation in the data

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

D8.4 INITIAL CALCULATION OF A SCORE REPRESENTING THE AMOUNT OF DISINFORMATION

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 4
Due date	31/10/2025
Submission date	24/10/2025
Deliverable lead	Partner UMONS
Version	1
Authors	Kevin El Haddad (Partner UMONS)
Contributors	Roberto Caldelli (Partner CNIT), Stefano Berretti (Partner MICC-UNIFI), Axel Brando (Partner BSC), Alvaro Parafita (Partner BSC), Jamal Nasir (Partner UoG), Qazi Alamgir (Partner UoG)
Reviewers	Chun Fei Lung (Partner HU)
Abstract	This report details the initial implementation of the DisinfoScore (DS) aggregation mechanism, a core component of the AI4DEBUNK decision support system for combating disinformation. The platform integrates a suite of multimodal analysis modules, including deepfake detection (text, image, audio) and cross-modal coherence

Funded by the European Union

checking. The current methodology calculates the final DS by normalizing all activated module outputs to a [0, 1] scale and computing a uniform weighted average. This document outlines the primary limitations of this approach, namely the simplistic, non-adaptive weighting scheme and the significant challenge of system evaluation due to the lack of a suitable multimodal dataset. Future work is defined, prioritizing a migration to a more robust and flexible agent-based orchestration model to enhance modularity and introduce explainability.

Keywords	Disinformation, Multimodal Analysis, Decision Support System, Score Aggregation, Deepfake Detection, Cross-Modal Coherence, Explainable AI (XAI)

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
0.2	27/10/2025	Reviewed by internal reviewers	HU
0.5	29/10/2025	Implementation of suggestions	UMONS
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

Funded by the European Union

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 D8.4: Initial calculation of a score representing the amount of disinformation in the data. https://ai4debunk.eu/wp-content/uploads/2025/11/AI4Debunk-Deliverable-8.4.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
XAI	Explainable AI
DS	DisinFoscore

Funded by the European Union

EXECUTIVE SUMMARY

The AI4Debunk project aims to create a decision support system to help users identify disinformation, built on a modular platform that integrates multiple specialized analysis tools. These modules include similarity estimation, a deepfake detection suite, and a cross-modal coherence checker. To provide a unified assessment, the platform aggregates outputs from all activated modules into a single "DisinfoScore". In the current implementation, this score is calculated as a simple weighted average of normalized outputs from each module. A key limitation of this version is the use of a uniform weighting scheme, which assumes each module provides an equal contribution. Furthermore, robust evaluation is hindered by the lack of a suitable, multimodal dataset. Future work will address these limitations by transitioning to a more flexible agent-based orchestration system. This new architecture will improve modularity and introduce critical explainability features. Finally, a bespoke evaluation dataset is actively being constructed to enable rigorous validation of the platform's performance.

Funded by
the European Union

1 INTRODUCTION

The AI4Debunk project is designed to develop a sophisticated decision support system that assists citizens and media professionals in the assessment of online content for potential disinformation. The core of this initiative is a multimodal, adaptable, and modular platform, the architecture of which is depicted in Figure 1. This platform integrates a suite of specialized analysis modules, each targeting a distinct facet of disinformation.

The current implementation of the platform comprises the following core modules:

- **Similarity Estimation Module:** This module ingests a given news artifact and queries a proprietary knowledge base to retrieve semantically similar articles. The retrieved set is partitioned into articles that support and those that oppose the reference content, providing crucial context (detailed in deliverable D8.1).
- **Deepfake Detection Suite:** This is a collection of specialized tools for detecting synthetic media across different modalities, as delineated by the orange boundary in Figure 1 (see deliverable D8.1 for a comprehensive overview).
    - **Textual Deepfake Detection:** Analyzes text fragments (e.g., social media posts, article excerpts) to verify authenticity. The output is designed to be richer than a simple binary classification, providing supporting evidence at the phrase or word level to facilitate further investigation.
    - **Image/Video Deepfake Detection:** Analyzes visual media for signs of manipulation by AI. The output includes localization heatmaps indicating manipulated regions, a binary authenticity assessment, and a probabilistic confidence score. For video inputs, the analysis is performed on a frame-by-frame basis.
    - **Audio Forgery Detection:** Processes audio streams from video files or standalone audio clips to detect synthetic or manipulated speech (e.g., voice cloning).
- **Cross-Modal Coherence Analysis Module:** This module evaluates the semantic consistency between an image and its associated text (e.g., caption, legend). It outputs a quantitative score representing the semantic distance between the two modalities, where a higher value indicates greater incoherence.

The types and modalities of the modules' input are predefined by the developer in this architecture. A fundamental design principle is that an incoming news artifact is processed only by modules whose input modality requirements are met. These concurrently processing modules are referred to as "activated modules."

Funded by the European Union

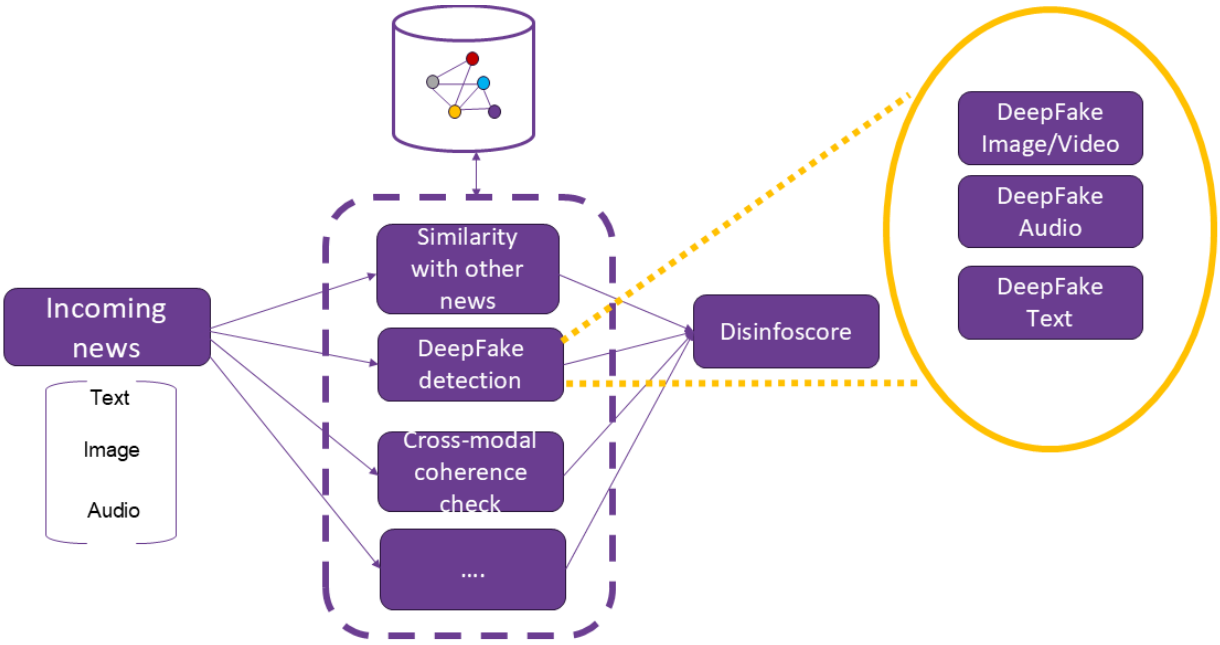


FIGURE 1 AI4DEBUNK PLATFORM

2 DISINFOSCORE AGGREGATION

The platform synthesizes the outputs from all activated modules into a single, unified metric termed the "DisinfoScore" (DS). This score provides a top-level indicator of the likelihood that a given news artifact contains disinformation. The aggregation is performed via a weighted average of the normalized outputs from each activated module, as illustrated in Figure 2.

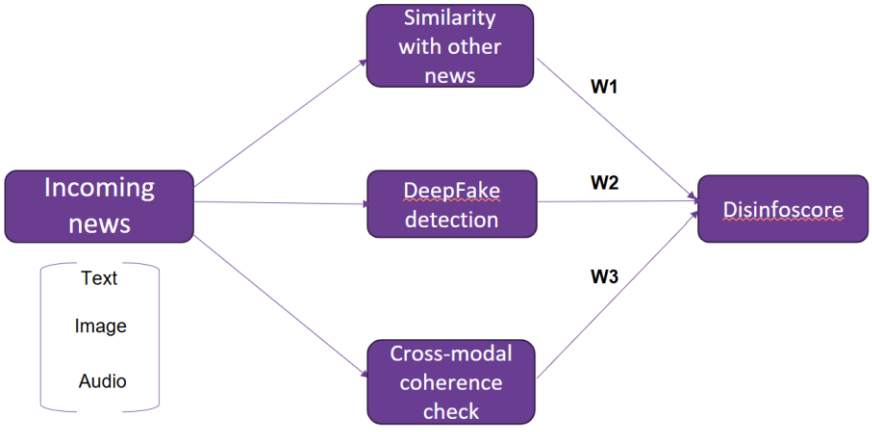


FIGURE 2 DISINFOSCORE CALCULATION

Funded by the European Union

2.1 OUTPUT NORMALIZATION

To ensure commensurability between the heterogeneous outputs of the different modules, a normalization step is applied to transform each output into a standardized score, S_i, within the range [0, 1]. This score represents the module's assessed likelihood of disinformation. The normalization schemes are as follows:

- **Similarity Estimation:** The normalized score is derived from the proportion of *supporting* articles (retrieved from the knowledge base) that are themselves verified instances of disinformation.
- **Deepfake Detection:** The module's probabilistic confidence score, when available, is used directly as the normalized score. In cases where only a binary output is provided (0 for authentic, 1 for deepfake), this value is used.
- **Cross-Modal Coherence:** The calculated semantic distance between the image and text is directly utilized as the normalized score, as it is already scaled within the [0, 1] range.

2.2 SCORE CALCULATION

The final DisinfoScore, DS, is computed as the weighted average of the normalized scores from the N activated modules:

$$DS = \sum_{i=1}^{N} W_i . S_i$$

In the current implementation, a uniform weighting scheme is employed. The weight for each activated module, W_i, is calculated as:

$$W_i = \frac{1}{N}$$

where N is the total number of activated modules for the given news artifact.

3 LIMITATIONS AND FUTURE DIRECTIONS

The initial implementation of the DisinfoScore calculation presents several limitations that will be addressed in future work packages.

A primary limitation is the use of a uniform weighting scheme, which presupposes that each module contributes equally to the final disinformation assessment. This is a significant simplification. Future work will explore two potential enhancement strategies:

1. **Empirical Weight Tuning:** A more sophisticated approach involves training the weights W_i on a pre-annotated, multimodal dataset. This would allow the system to learn the relative

Funded by the European Union

importance of each module's output in identifying disinformation. However, this method has a critical drawback: it is not scalable or adaptable. The AI4Debunk platform is designed for modularity, allowing for the addition, removal, or modification of analysis modules. A statically trained model would require complete re-training—and potentially new data collection—following any change in the platform's module configuration.

2. **Agent-Based Orchestration:** This is the designated path for future development. In this paradigm, each analysis module is encapsulated as an autonomous "Agent." A higher-level "Orchestrator Agent" will be responsible for dynamically querying the available agents, aggregating their outputs, and deriving the final DisinfoScore. This architecture offers two principal advantages:
    - **Enhanced Modularity:** The system becomes resilient to changes in the set of available modules. The Orchestrator can adapt its aggregation strategy based on the agents that respond.
    - **Explainability (XAI):** The Orchestrator can be designed to generate a reasoning trace, providing a transparent explanation of how it arrived at its final score based on the evidence provided by the individual agents.

A second limitation is the current lack of comprehensive system evaluation due to the scarcity of suitable datasets. While numerous disinformation datasets exist, finding one that is both multimodal and aligns with the specific input requirements of our diverse modules has proven challenging. To address this, work has been initiated in WP8 to construct a bespoke evaluation dataset by aggregating and curating several existing resources. This effort will be continued and finalized in WP9 to enable rigorous performance validation.

4 CONCLUSION

This report has detailed the initial implementation of the DS aggregation mechanism, which serves as the central decision support metric for the AI4Debunk platform. The current system successfully integrates normalized outputs from a diverse suite of analysis modules—including similarity estimation, deepfake detection across multiple modalities, and cross-modal coherence—using a uniform weighted average.

While this approach provides a functional baseline, its limitations are explicitly acknowledged. The primary drawbacks are the non-adaptive, uniform weighting of module contributions, and the significant challenge of system evaluation due to the lack of comprehensive, suitable multimodal test data.

The future direction for this work package is clear. One of the explored options would be to pivot from the current static model to a more dynamic and intelligent agent-based orchestration. This advanced architecture is the designated path forward as it inherently provides greater modularity, adaptability to platform changes, and a crucial framework for system explainability (XAI). The parallel effort to construct a bespoke, multimodal evaluation dataset, which will continue in WP9, is critical to rigorously validating the performance of this next-generation system. The successful execution of these future work items will be essential in advancing the AI4Debunk platform from its current prototype to a robust, scalable, and transparent decision support tool.

Funded by the European Union

Review Sheet of Deliverable Report

D8.4 Initial calculation of a score representing the amount of disinformation in the data

Editor(s):	Kevin El Haddad
Responsible Partner:	University of Mons (UMONS)
Status-Version:	Draft - v0.2
Date:	24/10/2025
Distribution level (CO, PU):	Public
Reviewer (Name/Organization)	Chun Fei Lung (HU)
Review date	29/10/2025

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	NA	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		The formatting is a bit off, even in the desktop version of Word.
… 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			

Funded by the European Union

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
… present consistency along the whole document in terms of English quality/style? (to avoid accidental usage of copy&paste text)	X			
About the content…				
ELEMENT TO REVIEW	**Y**	**N**	**NA**	**COMMENTS**
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)				
ELEMENT TO REVIEW	**Y**	**N**	**NA**	**COMMENTS**
Is the deliverable sufficiently innovative?				
Does the document present technical soundness and its methods are correctly explained?				
What do you think is the strongest aspect of the deliverable?				

Funded by the European Union

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
What do you think is the weakest aspect of the deliverable?				
Please perform a brief evaluation and/or validation of the results, if applicable.				
AI AND TECNOLOGICAL WPS (WP6 – WP11)				
ELEMENT TO REVIEW	**Y**	**N**	**NA**	**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?				The report clearly describes how the DisinfoScore is computed by combining the outputs of the platform's core modules.
What do you think is the weakest aspect of the deliverable?				
Please perform a brief evaluation and/or validation of the results, if applicable.			X	
DISSEMINATION AND EXPLOITATION WPs (WP15 – WP17)				
ELEMENT TO REVIEW	**Y**	**N**	**NA**	**COMMENTS**
Does the document present a consistent outreach and exploitation strategy?				
Are the methods and means correctly explained?				
What do you think is the strongest aspect of the deliverable?				
What do you think is the weakest aspect of the deliverable?				
Please perform a brief evaluation and/or validation of the results, if applicable.				
DISSEMINATION AND EXPLOITATION WPs (WP18)				
ELEMENT TO REVIEW	**Y**	**N**	**NA**	**COMMENTS**

Funded by the European Union

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
Does the document present the main ethical aspects regarding the use of methods and human involvement?				
What do you think is the strongest aspect of the deliverable?				
What do you think is the weakest aspect of the deliverable?				
Please perform a brief evaluation and/or validation of the results, if applicable.				

SUGGESTED IMPROVEMENTS

PAGE	SECTION	SUGGESTED IMPROVEMENT
12	3	I think I understand the concept of agent-based orchestration, but not how it would apply specifically to the computation of the DisinfoScore. We hope that the authors could explain this in a little bit more detail.

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