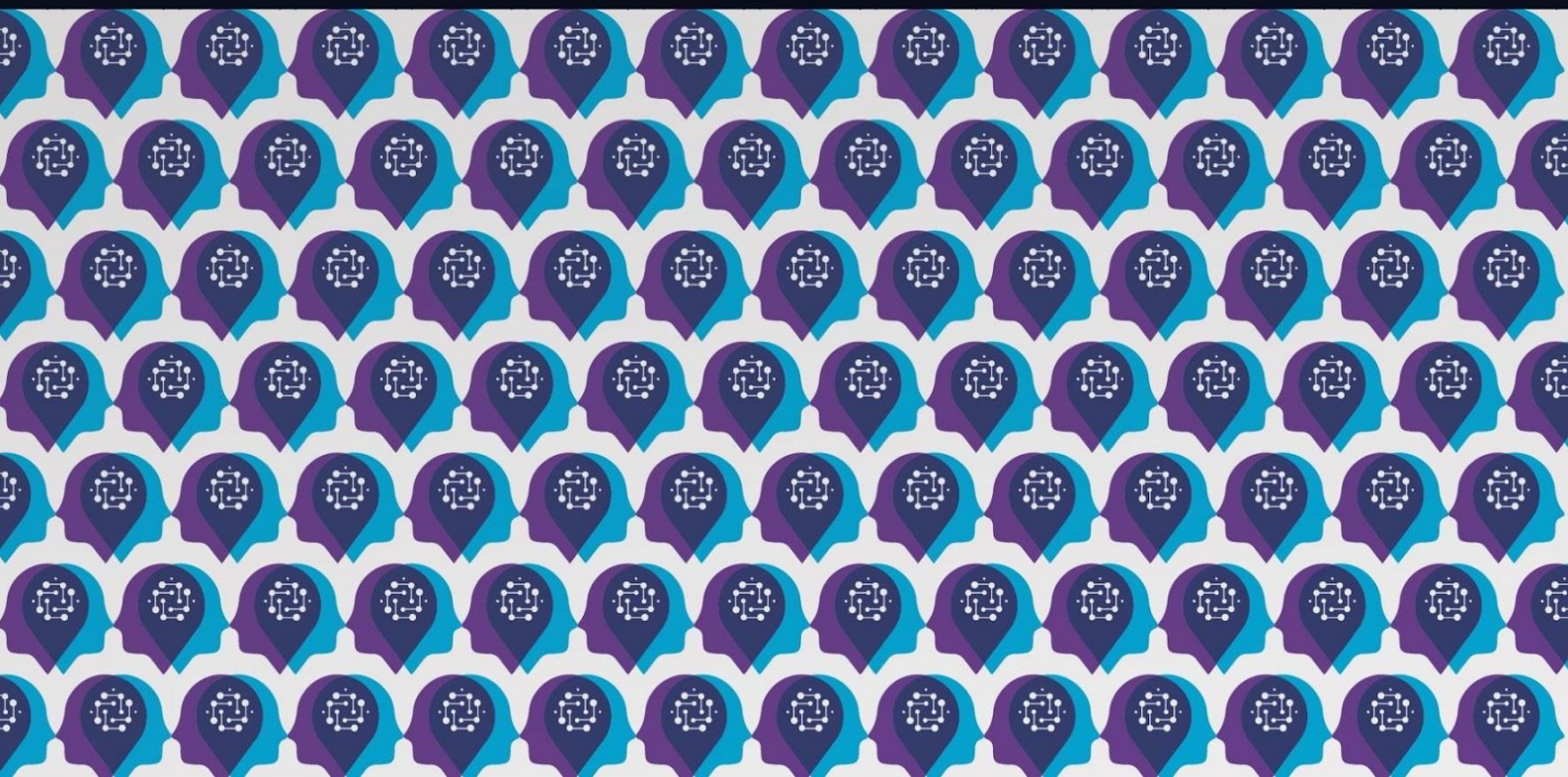
AI4Debunk

D8.3 INITIAL REPORTS ON THE TRUSTWORTHINESS OF THE DIFFERENT MODULES DEVELOPED

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

D8.3 INITIAL REPORTS ON THE TRUSTWORTHINESS OF THE DIFFERENT MODULES DEVELOPED

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 8.3
Due date	31/10/2025
Submission date	31/10/2025
Deliverable lead	BARCELONA SUPERCOMPUTING CENTER (BSC)
Version	1.0
Authors	Álvaro Parafita, Alejandro Astruc, Eddie Conti, Axel Brando (BSC)
Contributors	Kevin El Haddad (UMONS), Roberto Caldelli (CNIT), Stefano Berretti (MICC-UNIFI), Jamal Nasir (UoG), Qazi Alamgir (UoG)
Reviewers	Jan Kragt (Stichting Innovative Power)
Abstract	As disinformation spreads rapidly, amplified by AI systems capable of generating highly convincing but false content, addressing this complex, evolving threat requires multidisciplinary and modular strategies. The AI4Debunk project responds to this challenge with a flexible, scalable architecture designed to detect and mitigate disinformation early, but making these systems trustworthy is crucial for their eventual adoption. This report details novel research on Trustworthy AI topics for disinformation detection—through causal explanations, counterfactuals, and uncertainty quantification—and proposes practical strategies for making AI4Debunk's systems transparent, including

Funded by the European Union

	saliency methods, heatmaps, and model cards. These efforts lay the groundwork for integrative, cross-disciplinary approaches that enhance the project's technical robustness and societal relevance.
Keywords	trustworthiness, explainability, transparency, disinformation

DOCUMENT DISSEMINATION LEVEL

Dissemination level	
X	PU - Public
	SEN - Sensitive

DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributor(s)
0.1	03/10/2025	Final draft version	BSC, UMONS, CNIT, MICC-UNIFI, UoG
0.2	29/10/2025	Internal Quality Assessment Review	IP
0.3	30/10/2025	Implementation of suggestions	BSC, UMONS, CNIT, MICC-UNIFI, UoG
0.4	30/10/2025	Project Coordinator Review	
1.0	30/10/2025	Final version ready for submission	

Funded by the European Union

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 – organize 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 organize the inclusion of more and more intersectional perspectives in their activities.

DISCLAIMER

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

Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

COPYRIGHT NOTICE

How to cite this report: AI4Debunk (2025). **D8.3 Initial reports on the trustworthiness of the different modules developed.** https://ai4debunk.eu/wp-content/uploads/2025/11/AI4Debunk-Deliverable-8.3.pdf

AI4Debunk consortium is the following:

Participant number	Participant organization name	Short name	Country
1	LATVIJAS UNIVERSITATE / UNIVERSITY OF LATVIA	UL	LV
2	FREE MEDIA BULGARIA	EURACTIV	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 KAIVASEON 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

ABBREVIATIONS

EC	European Commission
AI	Artificial Intelligence
ML	Machine Learning
LLM	Large Language Model
XAI	eXplainable Artificial Intelligence
TWAI	TrustWorthy Artificial Intelligence
SSH	Social Sciences and Humanities
WP	Work Package
T8.3	Task 8.3 within WP8: “Creating Trustworthy AI models for fake news detection”
T9.3	Task 9.3 within WP9: “Developing Trustworthy AI models for fake news detection”

Funded by
the European Union

EXECUTIVE SUMMARY

*"A lie can travel halfway around the world
before the truth can get its shoes on."*

In the modern era of digitalisation, the spread of disinformation has become faster and more pervasive than ever. The advent of artificial intelligence systems capable of generating content that closely resembles truthful or plausible information has further complicated the identification of fake news. The consequences are far-reaching, from misleading policymakers to swaying entire populations. As noted by SSH partners in D4.1, disinformation is inherently complex and multifaceted: its sources, targets, and contexts are highly diverse. This complexity underscores the need for a 'cubist' approach: multidisciplinary collaboration and modular strategies are essential to capture the various forms disinformation can take.

In response to this challenge, AI4Debunk is developing a multidimensional, flexible, and scalable approach to better understand the nuances of how disinformation emerges. The core objective of the project is to detect false information early in order to prevent it from circulating further. Moreover, the project's flexible architecture is designed to evolve over time, adapting to the ever-changing nature of disinformation. However, as emphasized by the European Commission, AI solutions must not only be accurate but also align with ethical values, respect user privacy, and provide clear explanations of how models operate. Only in this way can trust be fostered and transparency ensured in the flow of information.

Our contribution to the project is twofold. First, we pursued research aimed at integrating trustworthiness considerations into disinformation detection anticipating challenges and opportunities. Our work includes developing causal concept explanations, novel attribution methods using counterfactuals, investigating the internal causal mechanisms behind LLM decisions, quantifying uncertainty in large generative models, and generating time-series counterfactuals. Although exploratory, these efforts lay the foundation for future cross-disciplinary research and the evolution of the project architecture. Second, we outlined practical and applicable strategies for the currently available modules considering their feasibility. These strategies include text saliency, image and video frame heatmaps, feature attribution, masking methods, and the creation of model cards (i.e., comprehensive and user-friendly descriptions of the models involved).

Together, these contributions pave the way for future research that actively involves SSH partners, broadening the scope of the project beyond purely technical dimensions by integrating socio-linguistic and cultural perspectives. This integrative approach not only enhances the project's robustness but also amplifies its potential impact on society at large.

Funded by
the European Union

1 INTRODUCTION

Disinformation has become a major concern as a form of hybrid warfare, deployed to destabilize democratic societies, foster distrust among citizens, and propagate harmful narratives. This threat has become increasingly urgent with the rise of Large Language Models (LLMs), which facilitate the generation of disinformation campaigns and enable the targeting of specific societal groups with tailored messages.

In response, the European Commission (EC) has launched several initiatives to combat disinformation, from both proactive and reactive angles. Proactive measures include media literacy programs and pre-bunking strategies designed to build societal resilience. Reactive efforts focus on identifying and debunking disinformation once it circulates.

Artificial Intelligence (AI) technologies represent both a novel threat and a potential asset in this context. On one hand, they offer new avenues, enabling the development of debunking systems that automatically detect disinformation and monitor for coordinated campaigns. However, these technologies also carry significant risks. AI systems can inherit and reproduce human biases, fail to quantify the uncertainty of their outputs, or be vulnerable to adversarial attacks that circumvent detection mechanisms.

These challenges foreground a central question: *how can we trust an AI system to determine whether a piece of content constitutes disinformation*? A system that relies on ethically questionable training data, or one that flags content indiscriminately when uncertain, risks undermining its own legitimacy. This underscores the importance of Trustworthy AI (TWAI), a field of research that aims at making AI systems transparent, fair, and accountable.

As part of Work Package 8 (WP8) of the AI4Debunk project, Task 8.3 (T8.3) focuses on evaluating the trustworthiness of the Machine Learning (ML) modules under development. Wherever feasible, the task also seeks to integrate transparency strategies into these modules. Following the completion of module development in WP8, these strategies will be implemented in T9.3 during WP9. By running in parallel with the other WP8 tasks, T8.3 facilitates the early identification of TWAI challenges within the AI4Debunk scope and ensures that TWAI principles are considered from the outset of model development.

This deliverable presents the outcomes of T8.3 and outlines the proposed strategy for implementing TWAI principles in the AI4Debunk project throughout T9.3. In the following, we will describe the WP8's modular AI architecture within which our work operates, highlight specific insights from the Social Sciences and Humanities (SSH) partners' working papers that affect TWAI methodologies for disinformation detection, and highlight key challenges in Trustworthy AI and how these inform our strategy. This section concludes with an overview of the structure of this deliverable.

1.1 WP8 Modular Architecture

WP8's technological solution for disinformation detection is built on the premise that disinformation is dynamic: manipulative strategies, news content, topics, sources and targets are in constant evolution. As a result, a monolithic system trained exclusively on today's data cannot be expected to maintain reliable performance over time.

To address this challenge, the proposed solution adopts a modular architecture composed of several AI modules, each designed to target a specific aspect of disinformation. These range from deepfake detection to cross-checking news items against alternative sources. The outputs of these modules are aggregated into a final disinformation score (*disinfoscore*), which reflects the contributions of the specific modules activated in a given instance.

The main advantage of this modular design lies in its adaptability. A module can be removed if it is shown to introduce bias, replaced if it becomes outdated or underperforms, or upgraded with improved alternatives developed by the community. Likewise, new modules can be added to capture novel forms of misinformation as they emerge. This approach makes AI4Debunk's system not only adaptable and reconfigurable but also transparent, since the contribution of each module to the final score can be made visible to the end user. In this way, the system is capable of evolving in step with the shifting landscape of disinformation, whether in the form of news articles, deepfake media, or social media content.1.2 Insights from Social Sciences and Humanities

This section summarizes the insights from the SSH partners' research relevant to the AI modules developed in WP8. These insights were incorporated to ensure that the research objectives of this task remained aligned with the broader project goals and responsive to the evolving landscape of disinformation.

Disinformation can be defined as the intentional, organized effort aimed at deceiving and potentially harming public welfare. Such harm may extend to democratic integrity, public health, and security, which underscores the need for stringent countermeasures. As proposed in D4.1, "WORKING PAPER 1: Theoretical framework for the analysis of disinformation campaigns and foreign interference in the EU policy making", disinformation threads can be identified through an explicit framework:

1. **Context**: historical, cultural, social, economic, political and global circumstances surrounding the disinformation attempt.
2. **Content**: actual content of the news item, and the way it is presented.
3. **Sources:** origins, methods of disinformation, and amplification mechanisms employed.
4. **Credibility**: linguistic and lexical structures used to enhance plausibility in the produced content.
5. **Target audience and engagement metrics**: mechanisms by which disinformation spreads.
6. **Impact**: tangible consequences of the disinformation piece in the real world.

From this framework, several disinformation detection strategies emerge, highlighting the need for **interdisciplinary collaboration** in countering this threat. Of special interest from the standpoint of the ML modules developed in WP8 are: **Context**, where background data is accumulated and leveraged through the Knowledge Graph developed in WP6; **Content**, where multiple data modalities (e.g., text, image, video, audio) are handled by specialized ML models in WP8's modular architecture; **Sources**, where source metadata stored in the Knowledge Graph provides meta-information for detection modules; and **Credibility**, where manipulation strategies are implicitly captured by the developed Deep Learning modules. As for the remaining two aspects, these are a less natural fit for the proposed ML modules and were left out of scope. However, future research efforts could integrate them into the AI4Debunk solution thanks to its modular design.

An important topic is that of **disinformation campaigns**—deliberate and coordinated efforts to spread false or misleading information intended to manipulate public opinion or behavior. From this perspective, by tracking individual instances of disinformation and unveiling patterns across them, threads emerge that indicate a coordinated attempt to promote specific narratives. Despite its importance, the consortium decided to exclude this topic from scope due to the complexity of its detection and monitoring, which exceeds the proposal's objectives. Nevertheless, anticipating a possible need in this direction in the medium term, TWAI efforts were made to accommodate this possible line of research, as discussed in Section 2.

Among other topics, D4.2 "WORKING PAPER 2: Information manipulation in the EU media ecosystem and response effectiveness" addressed **critical thinking** as a proactive approach to combating disinformation. Barak and Shahab (2023) describe critical thinking as involving several key skills: interpretation, analysis, evaluation, inference, and self-regulation. While the AI4Debunk technological

solution is primarily reactive (i.e., detecting new disinformation pieces), the inclusion of **TWAI methodologies** also fosters critical thinking by making the internal reasoning of these systems transparent, thereby contributing proactively to the fight against disinformation. By quantifying the uncertainty of automatic decision processes, users can determine when not to rely on automated systems, identifying ambiguous pieces or potential system failure modes. By producing explanations of these decisions, users can also learn to recognize the patterns that signal disinformation.

D5.1 "WORKING PAPER 3: Disinformation target groups in the EU member states, and sources and hosts of propaganda" examined the main **target groups** in disinformation attempts, identifying vulnerable populations such as youth, the elderly, minorities, rural communities, and individuals with lower levels of digital literacy. Interviewed groups deemed debunking strategies insufficient, suggesting more proactive approaches based on **critical thinking, media literacy, and civil society partnerships**. In particular, some of the challenges identified were the overreliance on reactive strategies, a deficit of trust in media credibility, and insufficient collaboration between technical and social science disciplines. These factors reinforce why multidisciplinary approaches like the AI4Debunk project are instrumental in combating disinformation.

D5.2 "WORKING PAPER 4: Narratives and foreign interference throughout Europe illustrated by case studies" analyzes case studies using the framework developed in D4.1, illustrating common features of disinformation: for example, the use of **deepfakes** and the reliance on distinctive linguistic and stylistic markers (e.g., high-density titles, emotionally charged language). On one hand, deepfake detection modules are developed as part of the WP8 modular architecture, tailored to each data modality. On the other hand, these **linguistic markers** can be identified automatically through Deep Learning models, which are general feature extractors, thereby facilitating their use in disinformation detection.

Finally, D12.2 "Resilience mechanisms triggered by the tool" outlines a comprehensive approach to enhancing the development and effectiveness of AI tools designed to detect and combat disinformation, building on resilience mechanisms and early beta testing results. The **guidelines** highlight the need for usability, inclusivity, ethical responsibility, and stakeholder involvement. They stress the importance of addressing coordinated deceptive activity, maintaining clarity in how disinformation is scored, and keeping pace with new disinformation strategies. To remain effective, these tools should integrate user feedback, update their algorithms regularly, and collaborate with experts. These guidelines are followed from the modular architecture of the disinformation detection system—enabling modules to be added, removed, or updated as needed—to the API and User Interface developed in WP10–11.

1.3 Challenges in Trustworthy Artificial Intelligence

The independent High-Level Expert Group on Artificial Intelligence set up by the European Commission presented the Ethics Guidelines for Trustworthy Artificial Intelligence in 2019. According to the guidelines, trustworthy AI should be **lawful** (respecting all applicable laws and regulations), **ethical** (respecting ethical principles and values) and **robust** (both from a technical perspective while also taking into account its social environment). They also included seven key requirements for an AI system to be deemed trustworthy:

- **Human agency and oversight**: AI should strengthen people's ability to make informed decisions while protecting their rights. This also means having clear structures for supervision, whether through direct involvement, monitoring roles, or ultimate decision-making authority remaining with humans.

- **Technical Robustness and safety**: AI must be designed to withstand errors or attacks and to operate securely. It should be dependable, precise, and consistent, with safeguards in place to reduce risks and handle failures without causing harm.
- **Privacy and data governance**: AI must respect personal data and ensure it is handled properly. This involves strong mechanisms for managing information, checking its accuracy and reliability, and ensuring that data use is lawful and appropriately controlled.
- **Transparency**: the functioning of AI, from the data it relies on to the way it is built and used, should be open to scrutiny. Systems should leave a clear record of their processes and provide understandable explanations to those affected. Users should always know when they are engaging with AI and be aware of the system's capabilities and limitations.
- **Diversity, non-discrimination and fairness**: AI should avoid reinforcing unfair treatment or exclusion. To promote inclusion, it should be usable by everyone, including persons with disabilities, and its development should involve relevant stakeholders throughout their entire life circle.
- **Societal and environmental well-being**: AI should contribute positively to communities and the planet. This means ensuring it is environmentally sustainable, mindful of its impact on ecosystems and societies, and beneficial not just for present users but also for future generations.
- **Accountability**: clear lines of responsibility must exist for the design and outcomes of AI systems. Tools for auditing how data, algorithms, and decisions are handled are essential, especially in critical applications. Additionally, people should have access to effective remedies when harm or mistakes occur.

Relative to the AI4Debunk project's outcomes, the establishment of an expert group—responsible for validating new disinformation cases submitted either by users or by the automated detection systems defined in this Work Package—ensures that human agency and oversight remain central to the process. The group's composition also helps safeguard diversity and prevent discriminatory outcomes, while reinforcing principles of fairness and accountability. Privacy and data governance are addressed in line with the Data Management Plan, which is developed and regularly updated within WP1–3 to ensure compliance with ethical and legal requirements. Technical robustness and safety are supported through the work of partners developing modules in WP8–9, who also take environmental considerations into account by seeking distilled and optimized versions of the models to minimize their computational footprint.

Our particular focus within this task, however, lies in the principle of **transparency**. Beyond regulatory compliance, transparency is essential for fostering user trust and enabling meaningful scrutiny of the system. This includes making the system's capabilities and limitations explicit through clear and understandable **documentation** and providing clear **explanations** of system decisions.

1.3.1 EXPLAINABLE ARTIFICIAL INTELLIGENCE

The eXplainable Artificial Intelligence (XAI) community has traditionally operated from two main perspectives. The first is the **ante-hoc** (intrinsic) approach, which integrates explainability mechanisms directly into the model's architecture. The second is the **post-hoc** approach, which applies model-agnostic techniques—such as calling the model or inspecting its parameters—after training, typically at inference time.

Ante-hoc approaches assume that embedding explainability into the model's decision-making process ensures *faithful* explanations (i.e., aligned with the model's underlying reasoning). However, this assumption does not always hold. For instance, Jain & Wallace (2019) argued that attention weights, despite their apparent transparency, failed to provide reliable token attributions. Similarly, Bordt et al. (2025) argued that explanation methods should be viewed as statistics of high-dimensional functions, which raises the question of how such statistics relate to the intuitive queries humans pose about models. From this perspective, attention attribution has often been misinterpreted as feature importance, leading to unfaithful explanations despite being an ante-hoc mechanism. Beyond issues of faithfulness, ante-hoc strategies present practical drawbacks: they can reduce predictive or computational performance, require significant adaptation for each new architecture, and are often infeasible in fine-tuning settings without retraining or modifying the full model. These constraints limit their adoption in rapidly evolving AI research and applied contexts.

Post-hoc approaches, by contrast, are more diverse and flexible. They can be applied to a wide range of models without altering the underlying architecture, making them highly adaptable. Yet, they too face limitations. Many are computationally expensive (e.g., Shapley Additive Explanations, which require repeated model evaluations), lack inherent faithfulness guarantees, and may produce unstable results (i.e., high variance across explanation runs). Despite these challenges, their adaptability makes post-hoc strategies the most widely used in practice.

Madsen et al. (2024) studied the traditional dichotomy between ante-hoc and post-hoc paradigms and argued for novel paradigms that challenge this dichotomy. These include: the **learn-to-faithfully-explain paradigm**, which explicitly optimizes explanation methods to improve their faithfulness; the **faithfulness-measurable model paradigm**, which introduces intrinsic constraints during training to guarantee reliable post-hoc explanations; and the **self-explaining model paradigm**, in which models produce both predictions and corresponding explanations. While these approaches address important limitations of earlier strategies, their practical feasibility and theoretical validity remain open to debate.

Regardless of how explanations are generated, a central question in XAI is determining what makes an explanation "good". This question has sparked significant debate within the XAI community (Bibal et al., 2022), particularly concerning the evaluation of explanations in terms of faithfulness. Defining clear **quality desiderata**—such as faithfulness, stability, and simplicity—offers a principled basis for comparing different approaches, rather than relying on cherry-picked examples that merely align with human intuition. However, faithfulness itself remains contested: scholars have argued that existing tests aimed at determining when explainer methods are faithful are inherently limited or even misleading (Jacovi & Goldberg, 2020), primarily due to untested assumptions or by confounding faithfulness with plausibility. This ongoing debate underscores the need to consider quality during the design of trustworthiness solutions, accounting for their underlying assumptions while making their limitations explicit.

1.4 Overview of Deliverable 8.3

Our goal in this task is twofold. First, we tracked discussions within the consortium and integrated insights from the SSH partners to identify AI needs for disinformation detection. This allowed us to anticipate advanced TWAI requirements that go beyond well-established solutions, to align research efforts accordingly. The challenges identified, together with the corresponding research carried out to address them, are presented in Section 2.

Second, we analyzed the final solutions proposed by the technical partners within WP8. This analysis served to map the different modules against TWAI considerations. We identified which TWAI mechanisms could be leveraged or embedded directly within the models themselves (e.g., saliency outputs, self-rationalizing architectures) and, where this was not feasible, selected feasible post-hoc solutions as appropriate. This feasibility-based approach allows us to prioritize strategies that are both technically sound and realistically implementable. The outcome of this analysis, along with the projected TWAI solutions to be developed in WP9 on the partners' finalized models, are detailed in Section 3.

2 TRUSTWORTHY AI RESEARCH – SUMMARY OF CONTRIBUTIONS

This section outlines key advanced technical challenges in TWAI for disinformation detection, informed by the SSH insights of the consortium. Identifying these challenges enabled us to align ongoing research with project needs while also motivating novel research specific to disinformation detection. We focus on four such challenges and the research developed around them. Rather than following a chronological order, the discussion is structured according to relevance, depth, feasibility, and alignment with the project's objectives.

The research presented here was conducted with the aim of anticipating challenges that might emerge once WP8's modules were finalized. Accordingly, much of the work has an exploratory character, serving to prepare the ground for future adaptation to the final modules developed by WP8 partners. Whether or not these insights are ultimately integrated into AIDebunk's final solution, the research advances the state of the art in TWAI applied to disinformation detection. In doing so, it provides a basis on which future projects and researchers can build. Moreover, AIDebunk's modular architecture ensures that even future modules could draw upon and benefit from these findings.

2.1 Linguistic and Lexical Markers as Concept Bases: Concept-Based Explanations

As pointed out in D4.1, "WORKING PAPER 1: Theoretical framework for the analysis of disinformation campaigns and foreign interference in the EU policy making", and in particular Zhang et al. (2018), there are certain **linguistic and lexical markers** that help identify disinformation pieces, for its aim is to provide credibility to the content being presented.

These include: information-packed titles, strategically designed to offer rapid and immediate access to the information; emotional intensity in the piece, through emotionally-marked words and those with positive or negative connotation; abundance of relative clauses, with the aim of enriching the text with apparently credible information; extensive use of the passive voice, allowing to manipulate the visibility of the agent; ambiguity and vagueness, through unspecific vocabulary, hedging words, the excessive use of personal pronouns instead of specific nouns, and higher frequency of first or second person pronouns; and common use of adjectives or superlatives, conferring a highly subjective and evaluative approach. These linguistic and lexical tricks obfuscate the illegitimacy of the piece, but serve as telltale signs of disinformation.

While manually defining such features is impractical, Deep Learning models—being universal function approximators—can infer them automatically. If a property is relevant to disinformation detection, models can potentially learn to recognize and incorporate it into their decisions. Standard explainability methods typically attribute importance to individual input dimensions (e.g., words in a text, or pixels on an image), leaving humans to interpret what these regions represent—an interpretative process that inevitably introduces human bias into the machine-generated "explanation". However, since we already know that certain linguistic or lexical markers are relevant, they can serve as a foundation for human-understandable concepts. Measuring their contribution to the final decision allows us to explain model behavior in terms of interpretable features, bridging the gap between raw attributions and human reasoning.

This approach can be referred to as **concept-based explanations**. Over the years, several strategies have been developed to leverage concepts: from probing classifiers (Kim et al., 2018; Belinkov, 2022) and Concept Bottleneck Models (Koh et al., 2020), to mechanistic interpretability methods such as Sparse Autoencoders (Bricken et al., 2023), which aim to derive monosemantic representations from otherwise polysemantic neurons. Collectively, these approaches reflect an ongoing effort to identify

abstract, semantically meaningful concepts as a core component of explanations. Once a mechanism exists to both detect these concepts in the model activation space and intervene upon them, their influence on final predictions can be systematically evaluated through attribution methods. Building on this last line of work, we contribute in two directions:

Parafita, Á., Garriga, T., Brando, A., & Cazorla, F. J. (2025). Practical do-Shapley Explanations with Estimand-Agnostic Causal Inference. Accepted at the *Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*. arXiv preprint at https://arxiv.org/abs/2509.20211
This work introduces a practical strategy for estimating do-SHAP, a variant of Shapley Values that incorporates *causal interventions* into its coalition-value function. The key idea is to adapt the widely used SHAP explainer to settings where the underlying causal graph—capturing cause–effect relations among variables—is known. In this framework, do-SHAP leverages the causal graph to compute interventional queries for each possible coalition of features or concepts. Then, these contributions are aggregated into distinct attribution scores for each feature. A straightforward approach would require the researcher to derive an observational estimand for every coalition-specific intervention. However, this quickly becomes unfeasible as the number of features increases. To address this, we propose an estimand-agnostic method: instead of manually deriving estimands, we train a single model of the data distribution that respects the structure of the causal graph. This model provides general-purpose estimation procedures that can be used to evaluate *identifiable* coalition queries directly, avoiding the need for ad hoc derivations. Finally, we also develop an efficient algorithm to detect and group coalitions that yield identical values. By eliminating redundant computations, this optimization substantially reduces the cost of do-SHAP estimation.

Causal concept explanations explicitly account for the intrinsic dependencies between concepts, allowing us to **disentangle genuine causal effects from spurious correlations**. This, in turn, provides clearer insights into how different factors contribute to a classifier's decision. Consider, for example, a piece of disinformation: the absence of *fact-checking practices* (e.g., lack of references to external evidence or citations) may lead to reduced *logical consistency* (manifested as contradictions, unsupported claims, or logical fallacies). Meanwhile, *emotional language intensity* (e.g., sensationalism or strongly-charged wording) can drive the piece's *virality potential*, a hallmark of many disinformation campaigns. Finally, the *credibility of the source* can shape both the presence of *fact-checking practices* and the extent of *sensationalist* language used.

This simple setting illustrates how causal connections naturally emerge between concepts, and how interventions on one can affect the overall likelihood of disinformation. For instance, when assessing the effect of emotionally-charged language, one must avoid conflating it with the piece's logical consistency. In a causal framework, logical consistency is only connected to emotional language through *back-door paths*, representing anticausal effects that are ignored once we intervene. By focusing on interventions, we obtain a more **faithful estimate of each concept's true influence**, yielding explanations that align more closely with real-world causal mechanisms.

Our second contribution belongs to the field of **Attribution Methods (AM)**, which aim to identify the features most influential in generating model outputs. While these methods are typically applied to the model's direct inputs, concept-based explanations, which are more interpretable for human users, can also serve as the basis for AMs.

Despite the broad array of attribution strategies developed in recent years, AMs often produce inconsistent or even conflicting results when applied to the same dataset and model. This issue,

commonly referred to as the Disagreement Problem (Harel et al., 2023, Krishna et al., 2024) suggests that there may not be a single, definitive explanation for each decision, but rather a set of complementary explanations that together describe the underlying decision-making process. In response, the research community has proposed several evaluation metrics to assess AMs, focusing on aspects such as *faithfulness* (i.e., alignment with the model's reasoning) and *stability* (i.e., reduced variability in explanation outcomes). For a comprehensive review, see Kadir et al. (2023).

In the following work, we introduce a novel AM based on **counterfactual distributions**. In the non-causal sense, a counterfactual refers to an alternative input that is similar to the one being studied, but whose differences lead to a significant change in the model's decision. By examining these counterfactuals, we can identify the features that are crucial for the model's output. Our approach generates attribution scores based on data distributions derived from these counterfactuals, providing a complementary yet faithful insight into the model's reasoning process.

Conti E., Parafita Á., Brando A. (2026). CID: Measuring Feature Importance Through Counterfactual Distributions. Under review at *Northern Lights Deep Learning Conference (NLDL).*
Assessing the importance of individual features in Machine Learning models is critical to understand the model's decision-making process. However, the lack of a definitive ground truth for feature importance underscores the need for alternative, well-founded measures. This paper introduces a novel post-hoc local feature importance method called Counterfactual Importance Distribution (CID). By generating positive and negative sets of counterfactuals (those that alter the decision and those that do not) and modeling their distributions using Kernel Density Estimation, we rank features based on a distributional dissimilarity measure. This measure is grounded in a rigorous mathematical framework and satisfies key properties required for a valid metric. We demonstrate the effectiveness of CID by comparing it to well-established local feature importance methods. Our results show that CID not only offers complementary insights but also improves performance on faithfulness metrics, offering more accurate and reliable explanations of model behavior.

Counterfactuals offer **contrastive explanations**—showing how changes in specific input features can alter a model's decision. This contrastive nature aligns well with human decision-making, making such explanations more intuitive. AMs based on counterfactuals, like the proposed CID method, provide deeper insights by examining how feature interactions influence decisions across different scenarios. This approach complements concept-based explanations, which emphasize human-understandable factors. While counterfactual-based methods offer valuable and nuanced perspectives, they remain advanced and exploratory for high-dimensional ML models, with future research needed to further refine and broaden their practical use.

2.2 Large Language Model Transparency with Attention Explanations

In recent years, the **transformer** architecture has become the dominant paradigm across multiple domains—including machine translation, text summarization, and computer vision—consistently achieving state-of-the-art results. Modern models frequently incorporate the transformer block as a central component because of its ability to capture rich dependencies within the input. By leveraging the self-attention mechanism, transformers dynamically focus on the most relevant elements of a sequence

at each attention layer, conditioned on its current context, which enables highly expressive and adaptive representations.

A major category of interpretability methods for transformers relies on attention-based explanations, where the core assumption is that the **attention weights (AWs)** in the early layers can be interpreted as indicators of each token's relative contribution to the model's output. However, this assumption has been the subject of considerable **debate**: Jain and Wallace (2019) demonstrated that AWs can be randomly scrambled without causing significant loss in performance. Similarly, Serrano and Smith (2019) showed that large portions of AWs can be zeroed out with little to no impact on performance. In contrast, Wiegreffe and Pinter (2019) argue that, for tasks where attention is indeed integral to performance, these vulnerabilities arise primarily because the rest of the model is held fixed during such manipulations.

Taken together, these findings cast doubt on the reliability of attention as an explanatory tool. If AWs can be arbitrarily modified or eliminated without meaningful consequences for performance, then the mapping between attention distributions and model decisions cannot be assumed to be uniquely determined.

This challenge raises the question of identifiability: whether attention provides a faithful account of the model's reasoning process. Jacovi & Goldberg (2020) highlight identifiability as a critical factor in assessing the explanatory validity of attention. Brunner et al. (2020) further refine the concept, defining ***AW identifiability*** as the condition in which attention weights for a given input can be uniquely recovered from the output of an attention head. On the other hand, they also introduce the notion of ***token identifiability***, which refers to the ability to reconstruct the original token from a hidden state. This concept is necessary for interpreting AWs, since deeper-layer representations heavily intermix token information. Without a degree of token identifiability, interpreting AWs as direct relations between input tokens becomes problematic.

Conti, E., Astruc, A., Parafita, A., & Brando, A. (2025). Probing the Embedding Space of Transformers via Minimal Token Perturbations. In *IJCAI 2025 Workshop on Explainable Artificial Intelligence*.
This work studies the effects of minimal token perturbations on the embedding space of transformers through a novel analytical framework. The approach is motivated by prior debates of the explanatory power of attention, where many experimental arguments overlooked both the actual flow of information within the transformer architecture and the challenges posed by token identifiability. Unlike earlier methods, the proposed framework avoids out-of-distribution instances by applying minimal token perturbations that preserve the semantics of the sequence. This design enables systematic analysis of the embedding space's sensitivity to input shifts, the propagation of token information across model layers, and the degree of token identifiability throughout the network. Our experiments reveal several key findings. First, we analyze the frequency with which tokens undergo minimal shifts, showing that rare tokens are more prone to producing larger displacements in the embedding space. Secondly, we study the propagation of perturbations across layers, demonstrating that token-level information becomes progressively intermixed in deeper representations. These results empirically support the widely held assumption that the earlier layers of a transformer provide more faithful proxies for interpretability. Overall, this work introduces the combination of token perturbations and the analysis of embedding-space shifts as a powerful tool for advancing interpretability in transformer-based models.

This work demonstrates how the often-unexplored embedding space plays a key role in providing explanations for how black-box models work. In particular, by extending this framework, we could analyze

which parts of the input are most relevant to the model's internal representations. Such analyses would not only allow us to measure the model's sensitivity, but also to identify the components that play a key role in classifying an input as disinformation.

Astruc A., Conti E., Parafita Á., Brando A. (2025). Challenging the Concept of Attention Identifiability. Under review at the *Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.
This work investigates identifiability within the attention mechanism, critically examining prior assumptions. The concept of attention identifiability was suggested to provide guarantees and mitigate concerns of unfaithfulness raised in the debate on attention. However, existing strategies to establish identifiability suffer from unaddressed limitations. We analyze the structural dependencies of AW matrices and their sensitivity to perturbations. In addition, we examine the information flow surrounding attention, highlighting the limitations of interpreting AWs in isolation. To probe these relationships, we employ two forms of token perturbations: one preserving semantic similarity to the original input, and another based on random substitutions. We then quantify the impact of these perturbations on the internal representations within attention layers. Our findings demonstrate that AWs cannot be meaningfully understood in isolation, as their identifiability is inherently entangled with the embedding space. This work contributes to the ongoing debate on the explanatory power of attention, advocating for a more nuanced analysis that incorporates both attention and embedding representations.

These contributions advance the state of knowledge in attention explainability by bridging theoretical debates with empirical insights into embedding dynamics and attention identifiability. They highlight that AWs cannot be fully understood in isolation but must be situated within the broader architecture of transformers. By systematically probing token perturbations, embedding shifts, and the structural dependencies of attention matrices, this body of work refines our understanding of **how information flows through transformer models**. Ultimately, these advances in attention explainability are critical for uncovering how attention-based architectures reach their decisions—an essential capability for the development of robust and trustworthy disinformation detection modules.

2.3 Trustworthy AI under Rising Model Complexity in Generative Systems

Modern AI systems frequently adapt large **foundation models** to specific tasks using small, domain-focused fine-tuning datasets. In AI4Debunk's case, diffusion models would have likely served as the backbone for some WP8's ML modules operating on image and video modalities (in particular, for deepfake detection). For this reason, we aligned ongoing research on this topic.

Unlike standard classifiers that output calibrated confidence scores, **diffusion models** generate high-dimensional images through long **stochastic sampling chains**. Their outputs do not yield a single interpretable probability, so assessing their reliability requires estimating uncertainty—and, crucially, distinguishing between **aleatoric uncertainty** (irreducible variability in the data) and **epistemic uncertainty** (uncertainty arising from limited or biased training). For diffusion models, this typically involves comparing distributions across generated samples or model instances, a process that can be computationally demanding.

Moreover, training objectives that approximate maximum likelihood effectively capture observable data noise (aleatoric uncertainty) but do not guarantee low epistemic uncertainty—particularly in underrepresented regions of the training distribution. Reliable measures of epistemic uncertainty are therefore highly valuable: they can flag images likely to be AI-generated, identify areas where the model has poor coverage or weak generalization, and support safer decisions (e.g., through abstention, human review, or adaptive thresholds) in downstream deepfake detectors.

Berry, L., Brando, A., & Meger, D. (2024). Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In *The 40th Conference on Uncertainty in Artificial Intelligence*.
This work proposes DECU, a practical way to measure epistemic uncertainty in diffusion image generators. The key idea is to build a lightweight ensemble of class-conditioned diffusion models by freezing almost all weights from a strong pretrained model and training only a small label-embedding module for each ensemble member; this dramatically reduces the trainable parameters (≈512k vs. 456M) and cuts training time (approx. 87% reduction), making ensembles feasible for very large generators. Uncertainty is then estimated by checking how much the ensemble members' predicted distributions disagree, using Pairwise-Distance Estimators (PaiDEs) that approximate mutual information between outputs and model weights without expensive sampling in high-dimensional spaces. On ImageNet, DECU highlights where the generator is unsure—especially under-sampled classes—offering a clear signal for risk assessment and future data collection. In short, DECU turns heavyweight uncertainty estimation for diffusion models into a computationally practical pipeline by combining frozen-weight ensembles with efficient disagreement measures.

This work was motivated by the anticipated need for uncertainty quantification within AI4Debunk's ML modules. While the approach remains exploratory—focusing on a specific class of generative models and presenting practical challenges when applied to fine-tuned AI systems—it marks an important step forward. For this reason, greater emphasis was ultimately placed on complementary strategies grounded in explainability and transparency, which will be introduced in Section 3. Even so, the contribution is significant: by advancing methods for integrating uncertainty quantification into large-scale generative systems, this work lays the groundwork for more transparent and trustworthy AI-based detection tools.

2.4 Disinformation Campaigns: Measuring Counterfactual Effects over Time

D4.1, "WORKING PAPER 1: Theoretical framework for the analysis of disinformation campaigns and foreign interference in the EU policy making" defines **disinformation campaigns** as deliberate and coordinated efforts to spread false or misleading information intended to manipulate public opinion or behavior. Such campaigns can bring specific issues to the forefront of media coverage, amplifying targeted narratives. To monitor these dynamics, it becomes particularly relevant to measure the influence of individual news items or social media posts on broader public discourse. This raises the **counterfactual** question: "how much attention would this topic have received had this piece not been published?". In other words, it seeks to estimate the *causal* effect of an intervention by comparing reality with a hypothetical alternative.

This inquiry belongs to the field of **Causal Inference**, where counterfactuals are typically examined under the assumption of i.i.d. (independent and identically distributed) data—that is, data points that are

mutually independent and drawn from the same underlying process. However, this assumption rarely holds for temporally structured data such as news, which are published sequentially, shaped by prior events, and capable of influencing future developments.

To address these challenges, we aligned prior research efforts with the objectives of AI4Debunk to explore the problem of time-series counterfactual estimation in a particular pharmaceutical setting:

Garriga, T., Sanz, G., de Cambra, E. S., & Brando, A. (2024). A Novel Application of SCMs to Time Series Counterfactual Estimation in the Pharmaceutical Industry. In *NeurIPS 2024 Causal Representation Learning Workshop*.
This work introduces a time-series causal pipeline to quantify the impact of an "event" on a topic. Built on Structural Causal Models and the abduction–action–prediction procedure, and implemented with conditional sparse autoencoders (CSAE), it infers the counterfactual trajectory from the observed series alone—without external controls. An Added Variations stress test shows the counterfactuals remain sensitive to unrelated shocks, enabling credible attribution. Validated on synthetic and real data (originally in a pharmaceutical setting), the approach transfers to disinformation monitoring by treating each publication or coordinated amplification of interest as the intervention and estimating its marginal impact on visibility.

This work represents an initial step toward detecting disinformation campaigns by estimating the counterfactual effects of specific news items or social media actions on public discourse around particular topics. The idea originated from early consortium discussions that underscored the significance of disinformation campaigns. Although the topic ultimately proved too complex to be fully addressed within the scope of this project, the findings remain valuable. They can inform and support future research, making it important to highlight this contribution.

3 TRUSTWORTHY AI STRATEGIES FOR WP8'S MODULES

This section presents the modules that compose AI4Debunk's modular architecture and outlines the TWAI strategies applied to each. The primary aim is to build on the TWAI mechanisms already embedded into these models whenever possible, and, where such mechanisms are absent, to incorporate additional technologies guided by feasibility considerations.

It is important to note, however, that not all modules support meaningful introspection. This limitation may arise from **architectural constraints** (e.g., rigid or opaque pipelines that restrict access to intermediate representations), **representational complexity** (where high dimensionality entangles parameters and features, obscuring causal or interpretable pathways), or their **training regime** (as with modules fine-tuned from large foundation models, whose internal representations remain only partially understood, making it difficult to separate pre-training effects from fine-tuning adaptations). For each module, we therefore evaluate the feasibility of applying TWAI strategies, with the aim of achieving the highest degree of transparency permitted by its design.

The structure of this section is as follows. We first present general considerations and strategies that apply to all or most modules. Next, we examine each module individually and outline the corresponding TWAI strategy. Finally, we discuss the transparency of the aggregator module, which produces the overall *disinfoscore*.

3.1 General Considerations

Given the dynamic and evolving nature of AI4Debunk's modular architecture, intrinsic approaches to explainability are of limited use unless the modules themselves are designed with built-in interpretability guarantees. Imposing such intrinsic constraints on a module's architecture would not only risk reducing its predictive accuracy and computational efficiency, but would also tightly couple the explainability solution to that specific implementation. As a result, any architectural modification, update, or replacement of the module would invalidate the intrinsic mechanism, requiring the explainability framework to be rebuilt from the ground up.

By contrast, post-hoc approaches offer greater flexibility and adaptability. They can be applied across different models and architectural revisions without the need to redesign the underlying modules. This portability makes them far better suited to AI4Debunk's modular design, where components may evolve over time or be swapped out entirely. For this reason, our strategy is to **leverage intrinsic mechanisms only when they are natively available** in a given module, **while relying on post-hoc methodologies as the default** means of ensuring transparency and interpretability.

Finally, to ensure transparency of the whole system, we will devise **factsheets**[1] for each submodel that constitutes AI4Debunk's modular solution. These will be brief one-page documents that present key information visually and concisely, aiming at providing non-technical users with a clear understanding of the system's components and their role in decision-making, without requiring them to engage with technical documentation. In this way, AI4Debunk offers practical accessibility, facilitating trust and usability for diverse audiences, including vulnerable subgroups.

[1] The European Union frequently provides information in this format as seen in: https://commission.europa.eu/strategy-and-policy/eu-budget/eu-borrower-investor-relations/factsheets_en

3.2 GRAPH SIMILARITY

Identifying disinformation is inherently challenging, as its characteristics evolve continuously. However, a reliable approach involves comparison with other known news items. If the piece under analysis closely resembles previously identified disinformation—especially in structure, content, and recurring claims—this provides strong grounds for classifying it as disinformation.

The Graph Similarity module builds on the continuous adaptation work on T6.4. This task ensures that AI4Debunk's databases are regularly updated through 1) feedback collected from the user interface, 2) ML/AI modules designed to detect emerging patterns in new data, and 3) subsequent validation by a dedicated debunking committee. Using these dynamic updates, the Graph Similarity module selects and compares news items that are similar to the one being studied. When such items are both thematically aligned (i.e., raising the same points) and already recognized as disinformation, the system can flag the new piece accordingly.

In summary, the module operates through a structured chain of steps, ensuring systematic comparison and classification:

- **Topic identification**. The input text is first categorized as *"Climate Change"*, *"Russia–Ukraine War"*, or *"Other"*. The classification relies on cosine similarity between the CLIP embeddings of the input title and the reference embeddings for the first two topics. If the similarity difference between the two topics is below a threshold of 0.05, the input is assigned to *"Other"*. If the topic is *"Other"* the module halts as it fails to find a similar topic within the data.
- **Retrieval of related news**. Once the topic is determined, a database of news titles on the same subject is selected. Using CLIP embeddings, the system retrieves the 50 titles most similar to the input text, based on cosine similarity.
- **Classification**. Each of the retrieved titles is then passed, via a formatted prompt, to a Qwen[2] LLM model, which classifies the stance of the title with respect to the input as *"support"*, *"against"*, *"not related"*, or *"undetermined"*.
- **Fake-score computation**. Finally, the fake-score is derived as the ratio between the number of *"support"* cases and the total count of either *"support"* or *"against"* cases.

The explanation strategy adopted for this module is based on **examples**. The model's decision relies on retrieving the most similar news headlines from the database, each categorized as either *supporting* or *opposing* the input. To enhance transparency, these top-ranked headlines—together with their similarity scores—can be presented directly to the user.

This approach offers two key advantages: its intuitiveness, since by grounding the explanation in concrete, comparable cases on the same topic, users can more easily understand and assess the result; and faithfulness, since the explanation reflects the actual reasoning process of the model, ensuring that what is shown to the user aligns with how the classification was made. Additionally, provided enough background data, a confidence estimator could be added to the predictive module, generating a score based on the number of titles that are relevant (i.e., either "support" or "against") within the filtered most similar entries in the database. Intuitively, the confidence in the prediction will be lower or higher depending on the number of related entries found.

[2] https://huggingface.co/Qwen

Proposed solution

• Links to the top N **most similar headlines**, including:
  ○ Related **metadata** for each entry (e.g., title, url, source).
  ○ **Label** with whether each headline "supports" or "opposes" the analyzed headline.
• **Confidence score**, describing the number of entries found with similar headlines.

3.3 Image-Caption Coherence

Disinformation campaigns often recycle images from unrelated past events, presenting them out of context to reinforce misleading narratives. To counter this, the Image-Caption Coherence module evaluates whether a news item's image and its accompanying caption are consistent. A mismatch between the two can serve as a strong indicator of a potential disinformation attempt.

The Image-Caption Coherence score is computed by measuring the cosine similarity between the CLIP (Contrastive Language–Image Pre-training) embeddings (Radford et al. 2021) of images and texts, using the ViT-B/32 model. CLIP guarantees that both the image and the caption are represented in the same embedding space, where the distance between vectors reflects their semantic similarity. Consequently, the cosine similarity provides a direct measure of text–image alignment, and conversely, 1 - similarity represents the degree of **incoherence**, which will be the output score of the module. Thanks to its inherent interpretability, we propose presenting this **incoherence score as an explanation of the model's output**.

Additionally, to gain a more fine-grained understanding of the Image-Caption Coherence score, we propose introducing controlled perturbations at the text level by selectively **masking** specific parts of the caption. By observing how these perturbations affect the underlying decision, we can explore the evolution of the caption's embedding vector and analyze the corresponding variations in the similarity score. If effective, this approach can serve as a proxy for identifying **which text segments contribute most** significantly to image-caption alignment, thereby revealing the components that exert the greatest influence on the overall score. To complement and validate this analysis we plan to incorporate faithfulness measurements such as comprehensiveness and sufficiency (DeYoung et al. 2019). These metrics will allow us to verify that masking the tokens or segments of tokens identified as important leads to a measurable drop in the model's prediction confidence, thereby confirming that the highlighted features are indeed aligned with the model's own reasoning process.

Proposed solution
• **Incoherence score**: number between 0 and 1, higher the more in disagreement an image and its caption are. • (Tentative) **Text Saliency**: mapping of text segments to numerical scores indicating each segment's relative importance for the overall incoherence score. *This can be visualized by highlighting each text segment according to its numerical Importance Score, following a color scale. For example: "Demonstrators gather outside the Ukranian Embassy in Paris," with bolder colors representing the most importance for the model's score.*

3.4 DEEPFAKE DETECTION

Deepfakes pose a significant threat to society, politics, and business (Westerlund, 2019). By fabricating highly realistic depictions of people saying or doing things that never occurred, they enable malicious actors to spread harmful narratives with unprecedented credibility. As a tool of disinformation, they demand robust countermeasures. To address this challenge, AI4Debunk provides multiple deepfake detection systems, with at least one tailored to each data modality (e.g., text, image, video, audio). Because each modality requires different architectural designs, Trustworthy AI solutions must be tailored to the specific requirements of each model.

3.4.1 IMAGE DEEPFAKE DETECTION

The various models considered have structural differences and capture different aspects of image analysis. Below, we suggest approaches and strategies to improve their transparency and explainability.

TruFor (Guillaro, 2023) is designed to detect and localize image forgeries. Its core idea is to compare the RGB image with a learned noise representation, which acts as a fingerprint of the image. By combining these two sources of information, the model outputs an anomaly localization map that highlights suspicious regions. In addition, it uses Noiseprint++ together with the RGB image to compute a confidence map, which identifies less reliable areas of the anomaly map. Finally, the anomaly and confidence maps are aggregated to produce a global integrity score for the image.
Briefly, the architecture consists of the following steps: the process starts with the RGB image and its corresponding extracted noise fingerprint. These inputs are fed into two parallel Mix Transformer (MiT-B2) encoder branches, creating separate feature maps for visual and noise information. The feature maps are fused in a Cross-Modal Feature Rectification Module (CM-FRM) to compare visual and noise information and find discrepancies. The fused features are passed to a lightweight All-MLP decoder to generate the final localization and confidence maps. Lastly, the maps are aggregated to produce the single integrity score.

MantraNet (Wu, 2019) is an end-to-end image forgery detection and localization solution, taking an image as input to predict pixel-level forgery likelihood map. ManTraNet is composed of two sub-networks: (1) Image Manipulation Trace Feature Extractor, a VGG-style network for the image manipulation classification task, sensitive to different manipulation types (385), encoding the image manipulation in a patch into a fixed dimension feature vector; (2) Local Anomaly Detection Network, the anomaly detection network to compare a local feature against the dominant feature averaged from a local region (Z-score), whose activation depends on how far a local feature deviates from the reference feature instead of the absolute value of a local feature. An LSTM analyzes the Z-scores sequentially and the output is used by a final convolutional layer to produce the forgery localization map.

Both models can be considered inherently self-explainable, because both of their predictions are generated from maps (**anomaly map** for TruFor, or the **forgery localization map** for MantraNet) visually indicating suspicious areas in the image. Therefore, these maps already constitute faithful explanations.

Proposed solution
• **Heatmap**: pixel matrix highlighting which regions of the image influenced the decision.

CLIP_BSID's architecture (Cozzolino, 2024) relies on CLIP embeddings and uses a linear classifier to distinguish between *real* and *AI-generated* images. The training setup is straightforward: starting from N real images with their captions, a text-to-image generator is used to produce N synthetic images based on the same captions. This results in pairs of similar images (real vs generated). All images are passed through a pre-trained and frozen CLIP Image Encoder (e.g., ViT-L/14), which outputs a high-dimensional feature vector for each image. These N+N feature vectors are then used to train a linear SVM classifier, which learns to separate real and AI-generated samples.

Given the nature of this model, which relies on CLIP embeddings and an SVM classifier, it is inherently difficult to produce faithful explanations. However, a feasible approach is to provide uncertainty-based insights. In particular, the **distance** of an image **to the SVM decision boundary** can serve as a proxy for the confidence of the classification. Since the distance itself is not bounded, a possible strategy (similar to the case of conformal prediction) is to use quantiles to generate confidence estimates within [0, 1]. Moreover, as an exploratory strategy to be tested on T9.3, we propose evaluating the feasibility of masking strategies (whether masking parts of the image is compatible with the embedding procedure, to avoid out-of-distribution effects), in which case we could add **heatmap** explanations.

Proposed solution
• **Confidence score**: quantized distance to the decision boundary (between 0 and 1). • (Tentative) **Heatmap**: pixel matrix highlighting regions of the image influencing the decision.

3.4.2 Video Deepfake Detection

The models employed for this data modality do not consider audio and they will produce a "Real" or "Fake" classification for each frame of the video. Consequently, the model will calculate fake-scores for each frame. These scores allow us to localise which sections of the video have been significantly altered the most, and thus have a stronger impact in the final prediction. We can use this information to provide users with an **interactive visual representation** highlighting these frames to be reviewed at will.

The **FF++** model (Rossler, 2019) uses a ResNet-50 model pre-trained on **ImageNet** as a strong feature extraction base. The top classification layer of the ResNet is replaced, and the entire network is then fine-tuned on the **FaceForensincs++**[3] dataset. The model does not receive the whole image, instead it first extracts the cropped faces from the video using the method in Thies (2016), and then feeds it to the ResNet-50.

Wavelet-Domain ResNet50: The model starts with an RGB image. Each color channel is decomposed in 4 sub-bands (LL, LH, HL, HH) by applying a Discrete Wavelet Transform decomposition. The resulting 12 maps are stacked to form a single 12-channel input tensor, representing the image's frequency information and are subsequently fed into a ResNet-50. Lastly, the final feature vector is passed to a classification head to produce the final probability.

Both models in this module employ the ResNet-50 architecture, a deep convolutional neural network with 50 layers, widely used as a backbone for image-related tasks. While powerful, this architecture is **not**

[3] https://github.com/ondyari/FaceForensics

inherently explainable and produces high-level feature representations that are difficult to interpret directly. To ensure feasibility, we propose to highlight the portion of the video that activated the deepfake model the most. This solution resembles the idea of feature importance at input level preserving the overall faithfulness. If feasible, we can explore masking methods to produce saliency maps if necessary. In the case of the **FF++** model we can also provide the cropped face which is actually employed by the detection model.

Proposed solution
• **Interactive explanations** highlighting the **suspicious frames** of the video: these allow users to review them afterwards, making the model's decision more transparent. • **Logit scores:** values between 0 and 1 indicating the likelihood of each frame being deepfake. • **Face bounding box:** in the case of **FF++,** box localizing the face passed to the ResNet. • (Tentative) **Heatmap**: pixel matrix highlighting regions of the image influencing the decision.

3.4.3 Audio Deepfake Detection

Another possible source of false or manipulated information comes from the modification of audio tracks. This module uses wav2vec (Baevski, 2020), a powerful framework that maps raw audio into a vector representing its main features. Based on this feature extractor, a specific network can be trained on top for a given task. In the case of the implemented architecture[4], the features extracted from wav2vec are used to train a classifier that distinguishes between real and fake inputs.

At the current state, the final details of the implementation and the training dataset are not known. Assuming a general architecture for audio classification, we believe that **time localization** is a suitable approach to better understand the model: finding the sections of audio most relevant for the classification. In fact, the fake-score calculator from this module is based on the maximum score obtained throughout the audio sequence, divided into segments. The explanation can therefore be shown as a graphical and possibly interactive bar, highlighting the top-k areas of maximum activation so that the user can click on them to review them.

Proposed solution
• **Interactive explanations** highlighting the **suspicious segments** of the audio: these allow users to review them afterwards, making the model's decision more transparent. • **Logit scores:** values between 0 and 1 indicating the likelihood of each segment being deepfake.

3.4.4 Text Deepfake Detection

For text deepfake detection, the selected model[5] is a fine-tuned version of DeBERTa-v3-base, which performs binary classification of machine-generated fragments. At this stage, the module is still under development, and the final implementation details are not yet known. As such, the proposed strategy considers the current state of the model and will be refined if needed on T9.3.

[4] The fine-tuned model is available at: https://huggingface.co/mo-thecreator/Deepfake-audio-detection

[5] Available at: https://huggingface.co/OU-Advacheck/deberta-v3-base-daigenc-mgt1a

For transparency, we propose returning the classification logits as a proxy of the model's confidence in the prediction. Beyond that, a feasible exploratory approach focuses on input-level perturbations to understand how the model's output changes. For example, on a sentence classified as fake, **progressive padding**—progressively masking longer segments of text with a special token code—can be applied to determine the minimal portion of input necessary for the classification. This highlights which parts, or interactions of parts, of the text are most influential in the model's decision in a similar fashion to masking strategies—with the added benefit that padding is an in-distribution perturbation for most self-supervised Transformer models. This input-focused method not only provides interpretable insights for users, but also opens the door to exploratory socio-linguistic analyses, helping to study the differences between natural and machine-generated language. Additionally, robustness can be evaluated by replacing words with synonyms and observing the impact on the classification output.

Proposed solution
• **Logit score:** value between 0 and 1 indicating the likelihood of the text being deepfake. • (Tentative) **Text Saliency:** mapping of text segments to numerical scores indicating each segment's relative importance for the overall incoherence score. Equivalent to the proposed solution for Image-Caption Coherence in Section 3.3.

3.5 DISINFOSCORE MODULE

All the previous modules are finally integrated into a single score called the **Disinfoscore**. The aggregation of all the distinct modules' outputs is not trivial, since some may not be activated depending on which data modalities are present in the news piece.

The Disinfoscore aggregates the outputs of all previous modules into a single value that reflects the overall disinformation level of the input. In particular, the similarity, coherence, and deepfake detection modules are combined by averaging their scores. This approach is inherently interpretable, as the final value can be explained by presenting the individual contributions of each module. The intermediate values can be shown directly through textual descriptions to provide users with a transparent view of how the score is computed. For instance, if the image–caption coherence module returns a score of 0.96, this indicates a strong mismatch between the caption and the image, contributing significantly to the overall Disinfoscore. Thanks to its adaptability, ease of implementation, and straightforward integration into the architecture, the Disinfoscore remains both user-friendly and easily interpretable, ensuring transparency.

Proposed solution
• **Set of independent fake scores** from each activated detection module: ◦ **Numerical** value of each fake_score.

4 CONCLUSION

The AI4Debunk project aims to develop an innovative and adaptive disinformation detection system, leveraging the multidisciplinary expertise of its consortium. The proposed modular architecture is designed to handle diverse inputs and capture the various forms disinformation can take, while ensuring the highest levels of transparency and trustworthiness in the solutions it offers. By integrating insights from SSH partners, we have incorporated social science perspectives into our research, establishing a foundation for future advancements in disinformation detection with trustworthiness at their core.

At the same time, we have outlined specific strategies to enhance the trustworthiness of the solutions that can be implemented within the project's current modules. Looking ahead, during T9.3, the strategies described in Section 3 will be implemented and tested across the project's modules. This will allow us to assess their feasibility and gain insights into the reasoning behind the models. Future work will also include exploratory analysis such as masking, which will generate heatmaps identifying the most relevant parts of text or images. By combining technical innovation with trustworthy, user-centred approaches, AI4Debunk is paving the way for solutions that are not only effective in detecting disinformation but also robust and transparent.

Funded by
the European Union

REFERENCES

Barak, M., & Shahab, C. (2023). The conceptualization of critical thinking: Toward a culturally inclusive framework for technology-enhanced instruction in higher education. *Journal of Science Education and Technology, 32(6)*, 872-883. Retrieved from: https://doi.org/10.1007/s10956-022-09999-4

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, *48*(1), 207-219.

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., & Watrin, P. (2022). Is Attention explanation? An introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3889-3900).

Bordt, S., Raidl, E., & von Luxburg, U. (2025). Position: Rethinking Explainable Machine Learning as Applied Statistics. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., ... & Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, *2*.

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., & Wattenhofer, R. (02 2020). On Identifiability in Transformers.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2019). ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429*.

Harel, N., Obolski, U., & Gilad-Bachrach, R. (2022). Inherent inconsistencies of feature importance. *arXiv preprint arXiv:2206.08204*.

High-Level Expert Group on AI (2019). Ethics guidelines for Trustworthy AI. Retrieved from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Jacovi, A., & Goldberg, Y. (2020). Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (Long and Short Papers) (pp. 3543-3556).

Kadir, M. A., Mosavi, A., & Sonntag, D. (2023, July). Evaluation metrics for xai: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)* (pp. 000111-000124). IEEE.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning* (pp. 2668-2677). PMLR.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In *International conference on machine learning* (pp. 5338-5348). PMLR.

Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Madsen, A., Lakkaraju, H., Reddy, S., & Chandar, S. (2024). Interpretability needs a new paradigm. *arXiv preprint* arXiv:2405.05386.

Molnar, C. (2020). *Interpretable machine learning*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Serrano, S., & Smith, N. A. (2019, July). Is Attention Interpretable? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2931–2951). doi:10.18653/v1/P19-1282

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review*, *9*(11).

Wiegreffe, S., & Pinter, Y. (2019, November). Attention is not not Explanation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 11–20). doi:10.18653/v1/D19-1002

Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., ... & Mina, A. X. (2018). A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018* (pp. 603-612).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., & Verdoliva, L. (2023). Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20606-20615).

Wu, Y., AbdAlmageed, W., & Natarajan, P. (2019). Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9543-9552).

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the bar of ai-generated image detection with clip (2023). *arXiv preprint arXiv:2312.00195*.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11).

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2387-2395).

Funded by
the European Union

Review Sheet of Deliverable/ Milestone Report

D8.3 Participative Assistive AI-powered Tools for Supporting Trustworthy Online Activity of Citizens and Debunking Disinformation

Editor(s):	Álvaro Parafita, Alejandro Astruc, Eddie Conti, Axel Brando (BSC)
Responsible Partner:	BARCELONA SUPERCOMPUTING CENTER (BSC)
Status-Version:	Draft v0.1
Date:	31/10/2025
Distribution level (CO, PU):	Public
Reviewer (Name/Organization)	Jan Kragt - Stichting Innovative Power (IP)
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"

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		

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
… 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			
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			

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
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)				
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			Multiple deepfake detection systems.
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)				

Funded by the European Union

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
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