AI4TRUST

D7.4

Innovation, Exploitation and Sustainability Plan v2

PARTNERS

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

Project acronym	AI4TRUST
Project full title:	AI-based-technologies for trustworthy solutions against disinformation
Grant info:	ID 101070190-AI4TRUST
Funding:	HORIZON-CL4-2021-HUMAN-01-27 - AI to fight disinformation (RIA)
Version:	1.0
Status	Final Version
Dissemination level:	Sensitive — limited under the conditions of the Grant Agreement
Due date	28/02/2025
Delivery date (resubmission)	28/02/2025
Work Package:	WP7 - Communication, Dissemination and Exploitation
Lead partner for this deliverable:	Fincons Group AG (FINCONS)
Partner(s) contributing:	All partners
Main author(s):	Marco Giovanelli (FINCONS), Marcello Paolo Scipioni (FINCONS)
Contributor(s):	All partners

Summary of Modifications

VERSION	DATE	AUTHOR(S)	SUMMARY OF MAIN CHANGES
0.1	13/01/2025	Marco Giovanelli (FIN)	Preliminary notes on the document for the revision.
0.2	21/01/2025	Marco Giovanelli (FIN)	Updated version of section 2 and 3. Annex I added.
0.3	24/01/2025	Marco Giovanelli (FIN)	Revised version of section 4.
0.4	11/02/2025	Marco Giovanelli (FIN)	Restructured and revised section 5, 6 and 7. Annex II added.
0.5	14/02/2025	All partners	Conclusion of the collection of partners' updates about Individual Innovation and Exploitation per Asset. Update of Annex II.
0.6	17/02/2025	Marco Giovanelli (FIN)	Finalisation for the internal review.
0.7	24/02/2025	Lina Livdane (GDI) and Danilo Giampiccolo (FBK)	Internal review.
0.8	27/02/2025	Marco Giovanelli (FIN) and Marcello Paolo Scipioni (FIN)	Implementation of revisions and suggestions from internal review.
1.0	28/02/2025	Riccardo Gallotti and Serena Bressan (FBK)	Final review by Project Coordinator and Project Manager.

History of Changes from D7.2 to D7.4

This deliverable of the **"AI4TRUST - AI-based technologies for trustworthy solutions against disinformation"** project, titled **"D7.4 - Innovation, Exploitation, and Sustainability Plan v2"**, is a revised version of the previously submitted "D7.2 - Innovation, Exploitation, and Sustainability Plan v1", **incorporating the recommendations outlined in the "General Project Review Consolidated Report (HE)", dated 28 June 2024**, following the project's first Review Meeting. This deliverable is part of **Work Package 7 "Communication, Dissemination, and Exploitation"** (hereinafter referred to as WP7). The following **changes** have been implemented to address the recommendations:

Recommendation	Changes
No **new Key Performance Indicators (KPIs)** have been added or removed in the second version. However, the **connection between project KPIs and Key Exploitable Results (KERs)** remains **unclear and fragmented** across different sections of the document.	A new **Section 4.4** has been introduced to **analyse the connections** between the project's **Expected Results** and the **outcomes achieved** in **AI4TRUST**, providing **clear references** to the three main **Key Exploitable Results (KERs)**. These **KERs** are now **defined in Section 4.2** with a **more detailed and effective description**.
• While some **improvements** have been made in the revised version, there is still **no joint exploitation strategy** or a **clearly defined business model**. The lack of clarity regarding **what will be commercialized and how** raises concerns about the **sustainability of the project beyond its duration**. • The **market analysis** in the second version is **more detailed**; however, the **value proposition** remains **ambiguous**, particularly in terms of **what** the consortium intends to commercialize and **how**. • There is **no business plan**, either at the **individual** or **joint level**. Given the relevance of this aspect, a **business model canvas** may be beneficial. Additionally, the future of **AI4TRUST beyond the project's conclusion** remains **unclear**.	**Section 4** has been **significantly revised** to focus on the project's overall **Innovation and Exploitation Strategy**. Specifically, **Section 4.1** now provides a **clearer overview** of the **AI4TRUST Platform**, detailing its **implementation steps** and their connection to the project's **Key Exploitable Results (KERs)**, which are **thoroughly described** in **Section 4.2**. Furthermore, **Section 5** has been **restructured** to offer a **more comprehensive and effective** description of the project's **Sustainability Strategy**. In particular, **Section 5.1** now includes a **more detailed Sustainability Plan**, while the newly introduced **Section 5.3**, focusing on the **Business and Monetization Model** for the **AI4TRUST project**, outlines the **identified revenue streams** and provides an **in-depth analysis** of various **business model strategies**.

Recommendation	Changes
Regarding **individual assets**, some — such as **CNRS** — are identified as **knowledge-based**. Further **clarification** is needed on the **exploitation strategy** and the approach to **Intellectual Property Rights (IPR) management**.	**Section 6** has been **revised** to focus on **Intellectual Property Rights (IPR) Management**, offering a **comprehensive overview** of its **objectives, methodology, and strategies** for addressing **potential conflicts** and **intellectual property issues**. Additionally, it provides a detailed analysis of **effective IP protection measures** and **tools** relevant to the project.

In detail, the **overall changes** applied to the document are listed below, in order of appearance:

- **Section 1** has been updated to reflect the revised document structure;
- **Section 2** has been enriched with updates to the market analysis compared to the previous version;
- **Sections 2.1 and 2.2** have been revised and enhanced with insights from the workshops conducted with end-user partners;
- **Section 2.3** has been updated to reflect changes in the market, including newly introduced or discontinued platforms/tools over the past year, as well as the positioning of the AI4TRUST Platform in relation to them;
- **Section 2.4** now includes a more detailed description of the *Target Customers*, based on the outcomes of the workshops with end-user partners;
- **Section 2.5** has been expanded to provide details on the *Value Chain*, while the *Value Proposition* has been integrated into the newly introduced *Business and Monetisation* section (**Section 6.3**);
- **Section 3** has been revised and enhanced to include a description of the methodologies used to collect feedback from end-user partners;
- **Section 4** has undergone substantial revisions, focusing primarily on the project's overall *Innovation and Exploitation Strategy*;
- **Section 4.1** now provides a clearer overview of the AI4TRUST Platform, detailing its implementation steps and their relationship with the project's *Key Exploitable Results (KERs)*;
- The previous **Section 4.2**, *Stakeholders and Potential Benefits*, has been merged with **Section 2.4** to present a more comprehensive overview of the project's *Target Customers*;
- **Section 4.2** (formerly 4.3) has been extensively revised to offer a more detailed and effective description of the project's three main KERs;
- **Section 4.3** now focuses on *Individual Innovation and Exploitation per Asset*, with the specific details for each partner moved to **Annex II**;

- **Section 4.4** is a new section that analyse the connections between the project Expected Results and Results achieved in AI4TRUST;
- **Section 5** has been restructured to provide a clearer and more effective description of the *Sustainability Strategy*;
- **Section 5.1** now includes a more detailed *Sustainability Plan*;
- **Section 5.2** has been redefined to focus on the *Engagement Strategy*;
- **Section 5.3** introduces a new chapter on the *Business and Monetisation Model* for the AI4TRUST project, outlining identified revenue streams and analysing different business model strategies;
- The previous **Section 5.4** has now been merged into **Section 7**;
- **Section 6** has been restructured to focus on *IPR Management*, offering a detailed overview of its objectives, methodology, and approach to handling potential conflicts and intellectual property issues. It also outlines viable IP protection measures and tools applicable to the project;
- **Section 7** now provides a more comprehensive conclusion, offering further insights into the *Innovation, Exploitation, and Sustainability* aspects of the AI4TRUST project;
- **Section 8 (Annex I)** is a newly added section, reporting the outcomes of the workshops with end-user partners;
- **Section 9 (Annex II)** is a new section that consolidates updated details on *Individual Innovation and Exploitation per Asset* for each partner, aligned with their current needs.

Table of Contents

List of Acronyms

ACRONYMS	MEANING
AI	Artificial Intelligence
API	Application Programming Interface
CIB	Coordinated Inauthentic Behaviour
CSO	Civil Society Organisation
DWS	Disinformation Warning System
GDPR	General Data Protection Regulation
IIT	Institute of Informatics and Telecommunications
IP	Intellectual Property
IPR	Intellectual Property Rights
KYC	Know Your Customer
ML	Machine Learning
MVP	Minimum Viable Product
NGO	Non-Governmental Organisation

NLP	Natural Language Processing
SW	Software
UI	User Interface
WP	Work Package
USP	Unique selling point
GCN	Graph Convolutional Networks
GANs	Generative adversarial networks

List of Definitions

Term	Definition
Disinformation	"Disinformation is when false information is knowingly shared to cause harm."[1] "Disinformation is understood as verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm."[2]
Misinformation	"Misinformation is when false information is shared, but no harm is meant."[1]
Malinformation	"Mal-information is when genuine information is shared to cause harm, often by moving information designed to stay private into the public sphere."[1]
Foreground	Foreground means the tangible and intangible results which are generated within a given project, including pieces of information, materials and knowledge and whether or not they can be protected. It includes intellectual property rights (e.g., copyrights, industrial designs, patents, plant variety rights), similar forms of protection (e.g., rights for databases) and unprotected know-how (e.g., confidential material). Results generated outside a project are not foreground[3].
Joint ownership	Joint ownership refers to the situation where two or more legal entities share the ownership of the same asset and/or Key Exploitable Results. Joint ownership of results arises whenever (i) different participants have generated them jointly and (ii) it is not possible to ascertain the respective contribution of each participant, or to separate the results for the purpose of application, obtaining, or maintaining their protection[4].

[1] Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking. Council of Europe. https://rm.coe.int/information-disorder-report-november-2017/1680764666

[2] European Commission. (2018, April 26). Communication from the Commission to the European parliament, the Council, the European economic and social committee and the Committee of the regions: Tackling online disinformation: a European Approach (Report COM/2018/236). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236

[3] Europe - Glossary - European Commission (europa.eu)

[4] Europe - Glossary - European Commission (europa.eu)

Key Exploitable Result	A Key Exploitable Result (KER) is an identified main interesting result, which has been selected and prioritised due to its high potential to be "exploited" – meaning to make use and derive benefits- downstream the value chain of a product, process or solution, or act as an important input to policy, further research or education[5].
Primary target customers	Primary target customers are intended to be the main target groups who are most likely to be interested in the project's key exploitable results which may have the potential for commercial exploitation.
Secondary target customers	Secondary target customers are considered to be the secondary target groups who may be interested in the project's key exploitable results which may have the potential for commercial exploitation.

List of Figures

[5] Europe - Glossary - European Commission (europa.eu)

List of Tables

Executive Summary

The present document constitutes the deliverable **"D7.4 – Innovation, Exploitation and Sustainability Plan v2"** of the European project **"AI4TRUST – AI-based Technologies for Trustworthy Solutions Against Disinformation"** (hereafter also referred to as **AI4TRUST**). This is the **fourth deliverable (D)** of **Work Package (WP) 7 – "Communication, Dissemination and Exploitation"**, and it represents a **revised version** of **"D7.2 – Innovation, Exploitation and Sustainability Plan v1"**.

This version integrates the **recommendations** outlined in the **"General Project Review Consolidated Report (HE)"**, dated **28 June 2024**, following the project's **first Review Meeting**. It includes **enhanced analyses,** and an **updated strategy** aimed at **fostering innovation**, **maximising the exploitation** of research outcomes, and **ensuring the long-term sustainability** of **AI4TRUST solutions** within the **European context**.

For further details regarding the **revisions**, please refer to the section **"History of Changes from D7.2 to D7.4"** above.

1. Introduction

The primary objective of **Task 7.2 – Exploitation Strategy and Innovation Management (T7.2)** of AI4TRUST WP7 is to develop a **comprehensive strategy** encompassing **economic, strategic, and commercial analyses** of AI4TRUST's **exploitable results**. This task addresses **legal and IPR challenges**, identifies **exploitation opportunities**, and integrates **stakeholder feedback** to enhance **marketability**, ensuring a seamless transition from **research to market implementation**.

D7.4 presents a systematic plan for transitioning from project results to mission-oriented exploitation of the AI4TRUST Platform. The document serves as a revision of D7.2, which initially outlined the strategy for efficient innovation, scaling, and utilisation of project results. D7.4 integrates the **recommendations** outlined in the **"General Project Review Consolidated Report (HE)"**, dated **28 June 2024**, following the project's **first Review Meeting**, thereby refining the strategy for achieving sustainable exploitation of the project's outcomes also beyond its lifespan. For more information about the revision of D7.4, please see the "History of changes from D7.2 to D7.4" section above. This deliverable is organised with the following structure:

- The **second chapter** defines the **market positioning** of **AI4TRUST**, analysing the **feasibility and potential** of the project in **combating false information** through a **hybrid system** that integrates **advanced AI solutions** with **human cooperation**.
- The **third chapter** outlines the **methodology** for the **exploitation strategy**, detailing the **phases** for identifying **Key Exploitable Results (KERs)**, formulating a **comprehensive strategy**, and developing **individual plans** for **innovation and sustainability**.
- The **fourth chapter** provides a **high-level overview** of the **AI4TRUST Platform's functionalities**, emphasising its **objectives** in combating **false information** and supporting **media practitioners, policymakers, and researchers** through **advanced tools and analytics**.
- The **fifth chapter** discusses the **strategies for innovation and exploitation**, focusing on **stakeholder feedback** to enhance **marketability and sustainability**, while also detailing **individual plans** for each partner to **maximise the potential** of their assets.
- The **sixth chapter** presents the **preliminary sustainability plan** for **AI4TRUST**, concentrating on **long-term viability** through **engagement strategies, business models, and operational planning** to ensure the project's **continued effectiveness** in addressing **false information**.
- The **seventh chapter** focuses on the **strategy for managing Intellectual Property Rights (IPR)** within the **AI4TRUST project**, emphasising **transparent agreements** on **asset utilisation**, **systematic reviews**, and a **structured framework** for addressing **IP-related issues** among partners.

2. Market and Customer Analysis

This section is dedicated to defining and describing **the market in which AI4TRUST aims to position itself** by developing the proposed solutions and all the integrated services within the project. The goal of this analysis is to describe the **feasibility and potential of the project**, which develops a hybrid system where machines cooperate with humans, relying on advanced AI solutions against sophisticated false information creation techniques to support media professionals and policymakers. The **AI4TRUST Platform** will monitor various online social platforms almost in real-time, filtering out social noise and analysing multimodal content (text, audio, visual) in multiple languages (up to 70% coverage in the EU) with novel AI algorithms. It will cooperate in an automated manner with an international network of human fact-checkers and media professionals, who will be periodically triggered and will frequently provide validated data to update the algorithms. The proposal, based on a human-centred approach to technology development and aligned with European social, ethical, and legal values, will be integrated into the standard toolbox of analysts working on false information.

Disinformation is a delicate issue that has captured the interest of institutions and governmental bodies striving to combat it for the collective good. It refers to false or misleading information that is deliberately created and disseminated with the intent to deceive or manipulate. This can include fabricated news stories, doctored images, or misleading statistics that are spread to influence public opinion or behaviour. **Misinformation**, on the other hand, encompasses false or misleading information that is shared without malicious intent. This can occur when individuals unknowingly share incorrect facts or rumours, believing them to be true. Misinformation can spread rapidly, especially on social media platforms, where users may not verify the accuracy of the content before sharing. **Malinformation** involves the use of genuine information with the intent to cause harm or damage. This can include the strategic release of true information to harm an individual or group, such as exposing private information or using accurate data to mislead the public in a harmful way.

The **EU's commitment to fighting disinformation** began as early as 2015 with the development of a strategic communication action plan to counter disinformation campaigns. Over the years, the EU's contribution has continued through actions and funding aimed at developing corrective solutions and mitigating the spread of disinformation, exacerbated by the growing popularity of social networks[6] (see Figure 1).

[6] Special Report 09/2021: Disinformation affecting the EU: tackled but not tamed | European Court of Auditors

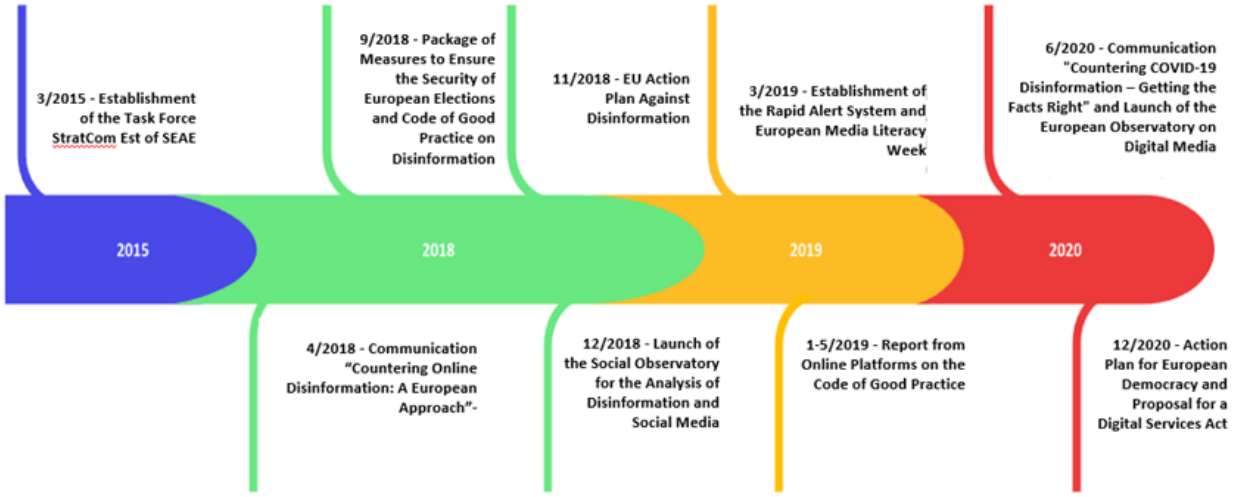

Figure 1: *Timeline of Major Initiatives Adopted by the EU Against Disinformation (Special Report 09/2021: Disinformation affecting the EU: tackled but not tamed, European Court of Auditors)*

To effectively **analyse disinformation** and develop **countermeasures**, it is essential to examine both the **informational tools** used by **users** and the ways in which **news publishers** have adapted. Over the past decade, **tech platforms and other intermediaries** have significantly influenced **news accessibility**. While **search engines** and **social media** serve different functions, **news access** has long been dominated by two major companies: **Google and Facebook (now Meta)**. The rising popularity of **digital audio and video** is introducing **new platforms**, further transforming the **media landscape**. These shifts represent a **"new normal"**, in which publishers must navigate an increasingly **complex platform environment**, characterised by **fragmented attention, low trust, and diminished open and representative participation**.

Every year, it is noticeable that direct access to apps and websites is becoming less important, while **social media is becoming more important due to its ubiquity and convenience** (Figure 2). At an aggregate level, we have reached a tipping point in recent years, with the preference for social media (30%) now far exceeding direct access to information (22%).[7] Substantial differences are also observed by age group. It is highlighted that **younger users** are less likely to access a news website or app directly and **more likely to use social media** or other intermediaries. The annual changes noticed in the comparison between direct and social access seem to be less due to older individuals changing their habits and more due to emerging behaviours of younger groups. The following chart for the UK shows that over 35s have changed their access preferences little over time, while the 18-24 group has become significantly less likely to use a news website or app (Figure 3)[8].

[7] https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023

[8] Digital News Report 2023 | Reuters Institute for the Study of Journalism

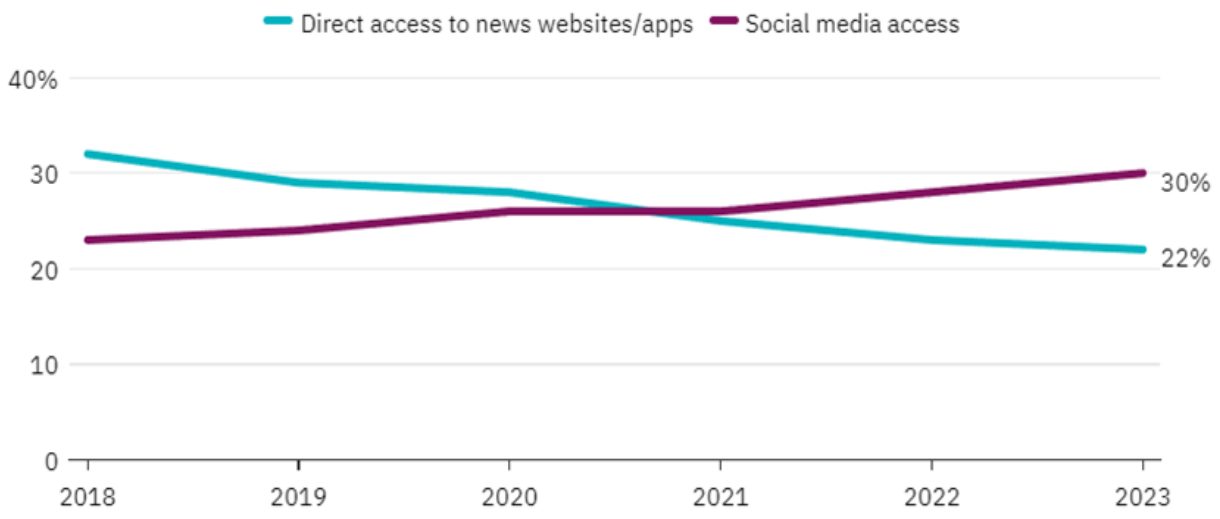

Figure 2: *Proportion of main online news retrieval channel (Digital News Report 2023, Reuters Institute)*

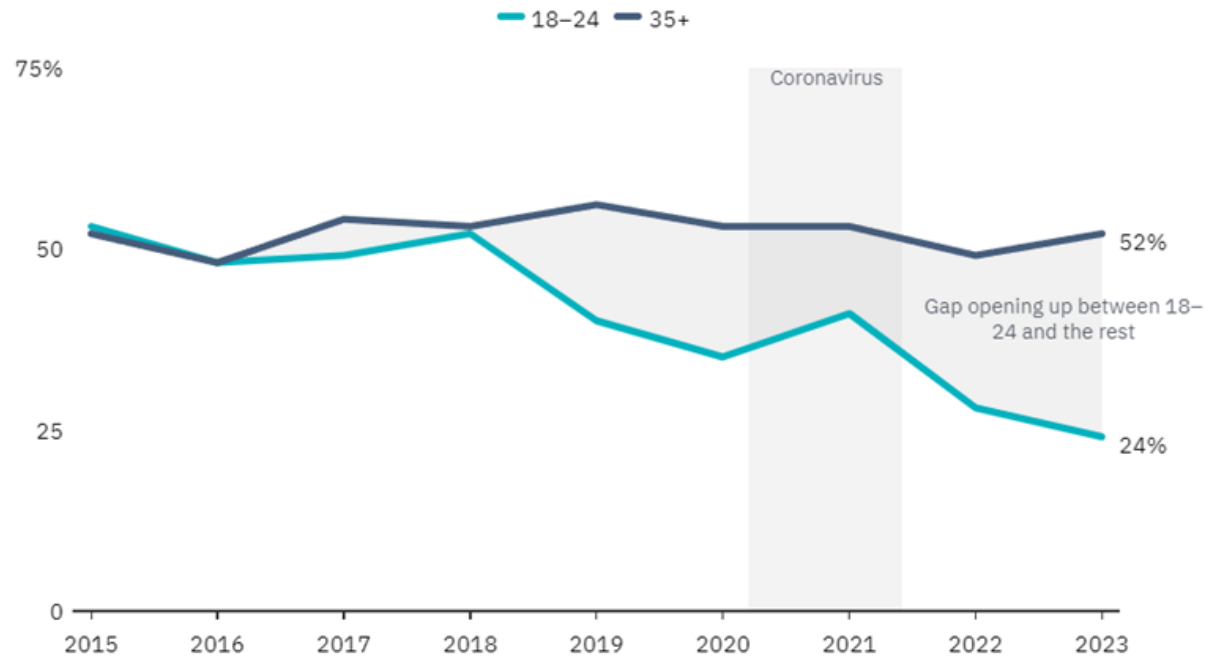

Figure 3: *Weekly proportion of users that accessed online news by going direct to a news website or app (Digital News Report 2023, Reuters Institute)*

The **illustrated trends** underscore the **emerging future direction** in the **sharing, dissemination, and utilisation** of news, as well as the necessary **adaptations** required by **information providers**. The growing **superficiality in news consumption**, the **expansion of social media and networks**,

and the **conflicts driven by economic interests** further **exacerbate misinformation**, heightening the risk of **disseminating inaccurate information** by **news providers**.

As we look toward 2025, **misinformation and disinformation** (hereinafter also "mis/disinformation") **are ranked as critical risks**, holding positions #4 and #5, respectively[9], in the recent **Global Risks Perception Survey 2024-2025** assessing global threats from over 900 experts across academia, business, government, international organisations and civil society. The increasing prevalence of disinformation is exacerbated by geopolitical tensions, such as Russia's invasion of Ukraine and ongoing conflicts in the Middle East and Sudan. These tensions contribute to **societal polarisation**, further amplifying the impact of false information on public perception and behaviour. **In 2027, mis/disinformation are projected to remain the top concern** among respondents, reflecting a growing awareness of the challenges posed by the rapid spread of misleading content.

The above-mentioned report also notes that the risks associated with mis/disinformation are **interconnected with other global threats**, such as state-based armed conflict and extreme weather events (Figure 4). As societies become more fragmented, the potential for false information to influence public opinion and exacerbate existing tensions grows, making it a critical area of concern for policymakers and citizens alike. The landscape of mis/disinformation is evolving rapidly, driven by technological advancements and geopolitical dynamics. Addressing these challenges requires a **multifaceted approach** that includes **enhancing media literacy, promoting transparency, and fostering collaboration among stakeholders** to combat the spread of these detrimental phenomena effectively.

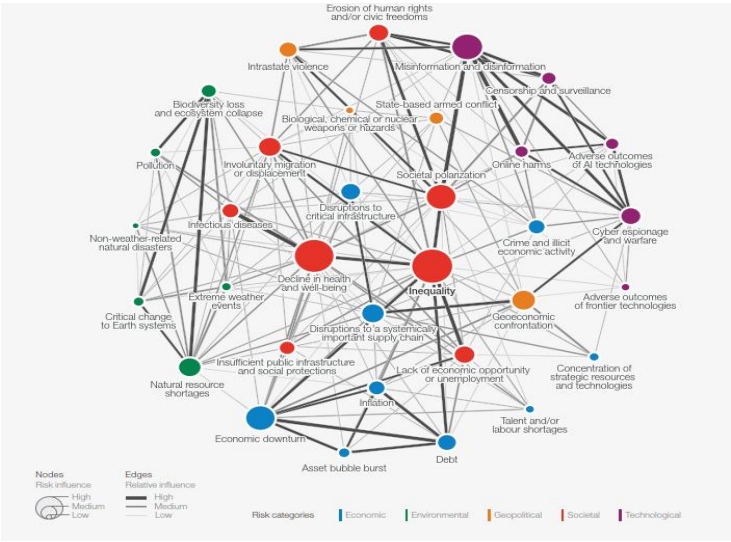

Figure 4: *Interconnections map of Global risks landscape (World Economic Forum Global Risks, Perception Survey 2024-2025)*

[9] World Economic Forum, Global Risks Report 2025

The **rise of digital platforms** and the **proliferation of AI-generated content** have significantly contributed to the increasing prevalence of **misinformation** and **disinformation**. With over **5.5 billion individuals** now connected online, **societal and political polarization** has intensified, leading to **algorithmic biases** that further complicate the **information landscape**. This polarization not only affects the **accuracy of content** but also amplifies the challenges in **distinguishing between true and false information**.

Misinformation and **disinformation** have become **pervasive**, with the volume of **misleading content** steadily rising. This surge is further exacerbated by the **fragmented media landscape**, making it increasingly difficult for **citizens, companies, and governments** to **identify** and **combat disinformation**. The advent of **generative AI** has lowered the **barriers to content production**, enabling various actors—ranging from **state agencies** to **individual users**—to **automate** and **expand disinformation campaigns**, thereby increasing their **reach and impact**.

In the ongoing battle against disinformation, **several European countries** are leading the charge through significant **investments and legislative measures**. Germany stands out as a pioneer with its Network Enforcement Act (NetzDG)[10], introduced in 2017, which set stringent compliance rules for online platforms to swiftly identify and remove disinformation. This law exemplifies Germany's proactive stance in combating false information, despite facing criticism regarding its implications for freedom of expression. The **European Union** as a whole has also made substantial strides with the implementation of the **Digital Services Act (DSA)**[11] and the **Digital Markets Act**[12], which establishes a comprehensive regulatory framework aimed at enhancing accountability and transparency among online platforms. This legislation mandates that EU member states designate authorities to enforce compliance, ensuring that disinformation is addressed effectively across the region.

Furthermore, the establishment of the **Central European Digital Media Observatory**[13] highlights the collaborative efforts among fact-checkers and academics to analyse and mitigate the impact of disinformation. The EU's East StratCom Task Force[14], particularly through its **EUvsDisinfo project**, focuses on countering disinformation campaigns, especially those emanating from external sources like Russia. Countries such as France and the Netherlands are also making notable contributions, with initiatives aimed at improving media literacy and fostering public awareness about disinformation. These nations are investing in educational programs and partnerships to empower citizens to critically evaluate the information they encounter online. Overall, Germany, along with the EU and countries like France and the Netherlands, are at the forefront of investing in measures

[10] Network Enforcement Act Regulatory Fining Guidelines

[11] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

[12] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

[13] European Digital Media Observatory (EDMO)

[14] EU's East StratCom Task Force

to combat mis/disinformation, reflecting a collective commitment to **safeguarding democratic processes and ensuring informed public discourse** in the digital age.

The **market for anti-disinformation technology based on advanced human-machine hybrid solutions** is an emerging segment that is rapidly gaining importance. These solutions are designed to detect, analyse, and counter the spread of mis/disinformation to improve the quality of service offered and the reliability of transmitted information. The **main trends** characterising the market context are:

- **Increase in Disinformation:** Social platforms and digital media have seen a rise in the spread of false information**,** often amplified by bots and algorithms.
- **Emerging Technologies**: Artificial Intelligence (AI) and Machine Learning (ML) are becoming crucial tools for analysing and recognizing disinformation patterns. AI-generated content refers to any text, audio, video, or image that is created by artificial intelligence systems. For instance, AI can generate text through natural language processing algorithms, create audio using voice synthesis technologies, produce videos through deepfake technology, and generate images using generative adversarial networks (GANs).
- **Fact-Checking**: This is the process of verifying the accuracy of information before it is published or disseminated. Fact-checking involves assessing claims against reliable sources and evidence to determine their truthfulness. It plays a critical role in maintaining the integrity of information in the media landscape.
- **AI-Based Detectors for AI-Generated Content:** These are specialised tools designed to identify and analyse content created by AI systems. They employ advanced algorithms to detect patterns and characteristics unique to AI-generated materials, such as inconsistencies in text, audio artifacts, or visual anomalies in images and videos. By utilising these detectors, organisations can more effectively combat the spread of false information and ensure that the content shared with the public is authentic and credible.
- **Multisectoral Collaboration**: Governments, non-governmental organisations, and the private sector are collaborating to develop solutions against disinformation.
- **Growth in Investments**: There is growing interest from investors in technologies that address disinformation, seen as crucial for the sustainability of modern democracies.

2.1. Market Needs and Opportunities

Media disinformation represents one of the most significant challenges of the **digital era**, with profound impacts on **society**, **politics**, and the **economy**. **Citizens**, **consumers**, **businesses**, and **institutions** are both authors and beneficiaries of **media information**, necessitating the **continuous evolution** of the sector to ensure **efficient service** and **high-quality standards**.

The **target market**, being new and rapidly growing, presents multiple opportunities for actors interested in developing **technological solutions** and **human-machine hybrid tools** to address the needs and opportunities the market presents. To effectively position **AI4TRUST solutions** within the sector outlined above, it is crucial to address **market needs** and meet **consumer demands**. From the analyses conducted, it is evident that **proper positioning** relies on **technological innovation**, **ease of use**, **reliability**, **accuracy**, and **strategic collaborations**. These combined elements enable the offering of **advanced and effective solutions** to combat **false information** and foster a more **reliable and secure information ecosystem**.

Market Needs

The **drivers of market growth** can be classified based on **service users** and **information providers**. The main aspects to consider include the proliferation of disinformation, characterised by the exponential increase in mis/disinformation transmitted through social media and digital platforms, creating an urgent demand for effective solutions to counter the phenomenon. The growing production of material that fuels disinformation has led to increased public awareness of the damage caused by unreliable information, driving the introduction of stricter regulations and the adoption of new anti-disinformation technologies. **Advances in AI and Machine Learning technologies positively stimulate the sector's growth** by enabling nearly real-time analysis and accurate detection of false or manipulated information, significantly supporting work currently done manually. It is important to highlight that businesses, governments, and institutions require tools to monitor and protect their online reputation against disinformation campaigns to ensure the integrity of their social status.

The market currently expresses specific needs that include:

- **Accuracy of Information**: Users constantly seek reliable sources and verified information to make informed decisions and guide their social, commercial, and political choices. For instance, citizens may change their habits, and companies may establish their business strategies based on the information circulating globally.
- **Reputation Protection**: Businesses and institutions need tools to monitor and protect their online reputation, filter inappropriate content for dissemination, and identify false or altered information that could impact their operations.
- **Consumer Education**: There is a growing need to educate the public on recognizing disinformation and techniques to avoid it. Providing tools capable of filtering false information is a significant social advantage, which must be accompanied by proper user education, enabling them to independently identify valid and altered information.
- **Regulatory Compliance**: With the increase in regulations on online content, platforms must adhere to increasingly stringent regulations regarding the dissemination of false information. Authorities have already initiated various measures to regulate digital security, privacy, and information, establishing norms and guidelines to protect all user categories.

- **Effective Detection Mechanisms**: There is a critical need for robust systems that can accurately detect and flag false information. This includes the development of sophisticated AI models capable of analysing vast amounts of data to identify potentially false information that can be reported to fact-checkers for further verification.
- **Nearly Real-time Analysis**: The ability to analyse and verify news in nearly real-time is essential to prevent the rapid spread of disinformation.
- **Scalability**: Solutions must be scalable to handle the enormous volume of content generated on digital platforms daily.
- **Integration with Existing Systems**: AI solutions need to be easily integrated with existing content management and social media platforms to ensure widespread adoption.

In recent times, the market has increasingly moved towards **multimodal integration**, aiming to develop more advanced solutions capable of integrating the analysis of **text**, **audio**, and **video** to enhance the **accuracy** and **completeness** of disinformation detection. As previously mentioned, **human-machine collaboration** is essential, and the combined use of **AI** and **human fact-checkers** is becoming the norm to ensure that data remains **accurate** and **up-to-date**.

Two additional factors driving the market recently include the **advanced automation** of **monitoring** and **analysis processes**, which is improving the **efficiency** and **scalability** of anti-disinformation solutions, and **linguistic expansion**, which allows for broader **language coverage** to address disinformation in various regions and languages.

On **January 9th, 2025**, a workshop was held with the project's **end-user partners** (i.e., fact-checkers, journalists and relevant researchers' part of the AI4TRUST consortium, as well as policymakers from the project's network of stakeholders) to gather their **feedback** on the **AI4TRUST Platform**. During the session, participants were divided into four groups based on the **primary target customers** identified: **researchers**, **journalists**, **fact-checkers**, and **policymakers**. This approach ensured that the development of the platform remained **user-centric**. Each group was tasked with completing four distinct canvases: the **Personas Canvas**, **Value Proposition Canvas**, **Ad Lib Value Proposition Template**, and **Prototype Canvas** (see chap. 3 "Methodology"). The workshop highlighted several **needs**, particularly in relation to the primary target customers identified for the **AI4TRUST solution** (discussed in more depth in section 2.4).

During the workshop, several key **market needs** emerged, reflecting the requirements of different target customers. **Fact-checkers** expressed a need for tools that enable **effective monitoring** of disinformation, **reliable identification** of **AI-generated content**, and **structured data presentation**, alongside transparency in the methodologies used for detection. **Journalists** highlighted the necessity for tools that **verify online news** and provide **contextual information** about content sources, as well as automated assessments to identify **AI-generated media**. **Policymakers**

emphasised the importance of having access to **credible sources of evidence** to inform their decisions. **Researchers** identified the need for **regulation-compliant data** for training and evaluation, **insights** to refine their training processes, and methods to collect **check-worthy data** from social media, along with **analytics** on the spread of disinformation. These identified needs underline the diverse requirements across stakeholder groups, emphasising the importance of a **comprehensive solution** that addresses these challenges.

Market Opportunities

Opportunities for AI in addressing media disinformation are extensive and varied. With the rise of social media and online platforms, misleading information has become a significant concern, threatening public safety, democracy, and civil debate. This situation creates a demand for effective solutions to identify and combat false information. **Advanced AI solutions** present numerous opportunities to combat sophisticated disinformation techniques, support media professionals, and enhance fact-checking activities. These opportunities are primarily driven by the capabilities of **machine learning, natural language processing (NLP), and reinforcement learning (RL)**.

Machine learning, a subset of AI, allows computers to learn from data and make intelligent judgments, thereby improving performance and producing more accurate results in various tasks. This is particularly crucial for detecting potentially false information to be reported to fact-checkers, as machine learning algorithms can analyse the inherent patterns and characteristics of textual material without relying on pre-labelled training data. This approach provides a proactive and scalable solution to the widespread issue of disinformation.

NLP enables computers to comprehend, interpret, and produce human language, which is essential for understanding and analysing textual data. In the context of false information detection, NLP can be used to clean and tokenize text, extract relevant features such as readability and word frequencies, and analyse the structure of information sources to spot false information. This capability is fundamental for applications like sentiment analysis tools, making NLP a critical technology in our data-driven society.

RL offers another advanced AI solution by training an agent to distinguish between authentic and false information through a reward and punishment system. This method can be further enhanced by improving feature extraction and utilising advanced NLP techniques. By continually refining these techniques, RL can significantly improve the accuracy and effectiveness of false information detection systems.

Utilising **large datasets** can aid in creating reliable models for categorising news items as authentic or fraudulent. These models can be evaluated using metrics such as recall, precision, and F1 score

to ensure their effectiveness. The availability of comprehensive datasets allows researchers and practitioners to **develop and refine models** that can identify potentially false information that can be reported to **fact-checkers**. Advanced AI solutions also support **media professionals and policymakers** by providing tools to **analyse and verify news** in nearly real-time, where machines cooperate with humans, helping to prevent the rapid spread of disinformation. These solutions assist in **defining rules and countermeasures** to combat false information effectively, thereby supporting the integrity of information distribution.

Furthermore, **AI can enhance fact-checking activities by supporting the verification process**, making it faster and more accurate. This helps media organisations and fact-checkers keep up with the high volume of information and ensures the integrity of the news being disseminated. Lastly, bringing **generative AI** knowledge to the next frontier involves developing more sophisticated AI models that can generate accurate and reliable content. This advancement supports the fight against disinformation and manipulation, ultimately defending EU citizens from the adverse effects of mis/disinformation[15].

Advanced AI solutions offer promising opportunities to combat disinformation, support media professionals, and enhance fact-checking activities by leveraging machine learning, NLP and reinforcement learning. **The market needs and opportunities for AI in media disinformation are significant.** With continued development, AI has the potential to improve human-computer interaction and unlock the enormous amount of unstructured text data that exists in the digital age.

In the last year, new opportunities have emerged in parallel with market needs. The **workshop we conducted on January 9th, 2025** (see above and chap. 3 "Methodology") has highlighted a series of opportunities mainly with respect to the primary target customers identified for the AI4TRUST solution. The market opportunities identified during the workshop highlight significant potential for growth and collaboration. Media outlets expressed a **strong interest in obtaining verified information**, indicating a willingness to invest in reliable sources to enhance their reporting. Additionally, the **availability of public funding** aimed at combating disinformation presents a viable avenue for supporting the platform's development and sustainability. The increasing volume of disinformation necessitates **robust analytical tools**, especially as social media platforms begin to remove existing monitoring tools, thereby creating a commercial space for AI4TRUST's monitoring capabilities. Furthermore, the rising number of fact-checking platforms underscores the **demand for effective solutions in this domain**, while journalists seek tools that enable them to focus on quality reporting rather than the challenges posed by mis/disinformation. Collectively, these

[15] Radha, J., Inmugesh, R., Kumar, J. R., Kumar, V. N., & Dhinakaran, R. (2025). Fake news detection with artificial intelligence, natural language processing and reinforcement learning. In *Challenges in Information, Communication and Computing Technology* (pp. 530-535). CRC Press.

insights suggest a **promising landscape for AI4TRUST to capitalise on emerging market needs** and secure its position as a leader in the fight against disinformation.

2.2. Pains and Gains

Advanced solutions in **Artificial Intelligence (AI)**, **Machine Learning (ML)**, **human-machine hybrid solutions**, and **digital technologies** in general address several critical challenges in combating false information. At the same time, they present numerous **growth** and **development opportunities**. One of the main challenges is the constant **evolution of false information creation techniques**, which use increasingly sophisticated methods, such as **deepfakes**, **image manipulation**, and **automatically generated content**, making their detection exceptionally difficult. This requires **AI solutions** to be **continuously updated** and improved to keep pace with these new techniques.

Technological barriers represent another significant obstacle. The accuracy of control and management algorithms heavily depends on the **quality** and **quantity** of available data. Without sufficiently representative and high-quality data, the algorithms can produce **inaccurate results**, compromising the effectiveness of the solutions. Moreover, the need to process **nearly real-time data** on a large scale poses considerable **technical challenges** in terms of **computing power** and **data management**. **Ethical and legal issues** constitute another critical area. The use of **AI** and **automated solutions** to monitor online content raises concerns regarding **user privacy** and the risk of **censorship**. Balancing the need to combat false information with protecting **users' rights** is a delicate challenge that requires **transparent solutions** that comply with **privacy regulations**.

Despite these challenges, there are significant opportunities for **human-machine hybrid technologies** based on **AI** and **ML** to combat false information. **Strategic collaborations** represent a promising avenue to enhance the effectiveness and adoption of these technologies. Partnering with **social media platforms**, **academic institutions**, and **governments** can facilitate access to **high-quality data**, improve **algorithm accuracy**, and increase the spread of these solutions. **Continuous innovation** is essential to maintaining a **competitive edge**. Investing in **research and development** allows for the introduction of **new features** and improvements in solution efficiency. This includes developing more advanced **algorithms** for **multimodal content analysis** (text, audio, video), expanding recognised **languages**, and integrating new technologies such as **machine learning**.

Finally, **emerging markets** offer vast growth potential. In many regions, false information is particularly pervasive, and the demand for effective solutions is increasing. Expanding presence in these markets can not only help reduce the spread of false information but also open new

commercial opportunities for companies developing anti-disinformation technologies. **Digital technological solutions**, specifically those based on the use of **Artificial Intelligence** to combat false information, must confront significant challenges related to the evolution of false information creation techniques, technological barriers, and ethical issues. However, there are ample **growth opportunities** through **strategic collaborations**, **continuous innovation**, and **expansion into emerging markets**. Addressing these challenges with innovative and collaborative approaches is crucial for **long-term success** in this rapidly evolving sector. **Artificial Intelligence (AI)** has been increasingly utilised for detecting **false information** and mitigating its spread, offering significant advantages but also presenting certain challenges. The following are the key **"pains and gains"** that have been identified.

Gains

For what concerns the "Gains" side, AI models have demonstrated **high accuracy** in detecting potentially false information to be reported to fact-checkers. This high level of precision helps in quickly identifying and flagging false information before it spreads widely. AI can process large volumes of data rapidly and accurately, making it an essential tool in the digital age where vast amounts of textual data need to be analysed. AI systems, particularly those using **reinforcement learning (RL)**, can learn from their mistakes and improve over time. By using a reward and punishment system, these systems can better distinguish between real and false information, enhancing their effectiveness with continuous use.

AI, especially **Natural Language Processing (NLP)**, can **handle unstructured text data** prevalent in news articles. This capability is crucial for understanding and analysing the vast amounts of data generated daily on digital platforms[16]. Researchers are developing models to track and **analyse cross-platform information** flows, examining how content mutates or adapts as it moves between different social media environments. This helps in understanding the spread and impact of false information across various platforms. AI can utilise **advanced techniques** like deep Q-learning algorithms, Dynamic GCN (Graph Convolutional Networks), and FakeNewsTracker for improved precision in false information detection. These techniques enhance the **accuracy and reliability of AI systems** in identifying false information. AI helps in **understanding the social impact** of false information, including their influence on political events and democracy, the role of social bots, and

[16] Radha, J., Inmugesh, R., Kumar, J. R., Kumar, V. N., & Dhinakaran, R. (2025). Fake news detection with artificial intelligence, natural language processing and reinforcement learning. In *Challenges in Information, Communication and Computing Technology* (pp. 530-535). CRC Press

the real-world consequences. This understanding is crucial for developing effective countermeasures[17].

The **workshop we conducted on January 9th, 2025** (see above and chap. 3 "Methodology") has highlighted new "gains" that have emerged in the last year, especially regarding potential primary target customers. The **gains** emerged during the workshop highlight significant benefits for various target customers. **Fact-checkers** will experience enhanced capabilities in monitoring false information across social media platforms, supported by clear methodologies and AI-based technologies for identifying tampered content. **Journalists** will benefit from the ability to verify information quickly and accurately, leading to an expanded database of verified facts that can improve the quality of their reporting. **Policymakers** will gain access to aggregated data on false information trends, enabling them to make informed decisions. Additionally, **researchers** will have access to updated datasets and regulation-compliant data for training AI methods, along with indices of verified and manipulated content. Overall, these gains emphasise the **comprehensive support AI4TRUST provides in combating false information and enhancing the integrity of information** dissemination across multiple languages and platforms.

Pains

For what concerns the “Pains” side, the effectiveness of AI in detecting false information largely **depends on the quality and diversity of the training data**. If the training data is biased or limited, the AI system may not perform well in real-world scenarios. This limitation affects the generalisability and reliability of AI models. **Language is complex** and nuanced, with elements like sarcasm, irony, and cultural references posing challenges for AI systems. These complexities can lead to misclassification of news items, reducing the accuracy of AI models.

AI models **need to be continuously trained** with new data to keep up with the evolving nature of false information. This requirement demands significant computational resources and expertise, making it a challenging and resource-intensive process[18]. Moreover, **ethical and privacy concerns** related to data collection and usage in false information detection need to be addressed. Ensuring data security and adhering to privacy regulations is crucial when developing and deploying AI models. Balancing these concerns while maintaining the effectiveness of the models remains a significant challenge. Finally, AI models may be vulnerable to **adversarial attacks** and manipulation, especially those that mimic human-like behaviour or use sophisticated rewriting

[17] D. Plikynas, I. Rizgelienė and G. Korvel, *"Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact through Machine Learning,"* in IEEE Access, doi: 10.1109/ACCESS.2025.3530688. (pp. 23-38)

[18] Radha, J., Inmugesh, R., Kumar, J. R., Kumar, V. N., & Dhinakaran, R. (2025). Fake news detection with artificial intelligence, natural language processing and reinforcement learning. In *Challenges in Information, Communication and Computing Technology* (pp. 530-535). CRC Press

styles. Ensuring the resilience of AI models against such attacks is essential for maintaining their reliability and effectiveness[19].

During the **workshop we conducted on January 9th, 2025** (see above and chap. 3 "Methodology") several new "pains" emerged. These **pains** bring to light the challenges faced by fact-checkers, journalists, policymakers, and researchers. **Fact-checkers** expressed concerns about social media platforms removing monitoring tools, which hinders their ability to analyse the increasing volumes of disinformation. Additionally, the reliability of AI tools used for fact-checking was questioned, as many are perceived as untrustworthy. **Journalists** noted the difficulty and time-consuming nature of verifying content, particularly when it is AI-generated, with the risk that false information can go viral before verification occurs. **Policymakers** highlighted the challenge of identifying effective policies in a rapidly evolving misinformation landscape. **Researchers** pointed out issues such as scattered and outdated datasets, increasing regulatory complexities, and barriers to accessing data from social networks, all of which complicate their efforts to analyse and address misinformation effectively. **These pains underscore the urgent need for a robust solution** that addresses these challenges comprehensively.

In conclusion, although AI presents promising solutions for detecting false information and mitigating its spread, **it is not without its inherent challenges**. Research and development are then focusing on improving the precision and effectiveness of these systems, addressing ethical and privacy concerns, and ensuring the resilience of AI models against adversarial attacks[20]. On the one hand, while the integration of advanced AI technologies presents significant gains in the fight against disinformation; on the other hand, it is imperative to address the accompanying challenges. **Continuous innovation, ethical/legal considerations, and strategic collaborations** will be essential in enhancing the effectiveness of these solutions. By navigating these complexities, **stakeholders can better leverage AI's potential** to create a more informed and resilient digital landscape.

2.3. AI4TRUST Market Context and Competitor Analysis

The **considered market**, being emerging, of interest to various stakeholders, and rich in opportunities, is **highly competitive and populated by numerous actors** offering different types of solutions. The **commercial proposal of AI4TRUST**, based on advanced AI solutions against sophisticated false information creation techniques to support the potential target customers

[19] D. Plikynas, I. Rizgelienė and G. Korvel, *"Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact through Machine Learning,"* in IEEE Access, doi: 10.1109/ACCESS.2025.3530688. (pp. 23-38)

[20] Ngueajio, Mikel, et al. "*Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques: A Survey of Explainable AI Techniques."* ACM Computing Surveys.

identified in section 2.4, is enforced by a **comparative analysis conducted in the digital sector and social media management**. The initial analysis identified **11 tools that offer services comparable** to those provided by the AI4TRUST project. The analysis that was previously done in the former version of this document (D7.2) is further enhanced by the identification of an **additional 14 tools**, as illustrated in **Table 1**.

Tool	Features	Type of service
Truly Media[21]	**Collaborative platform for verifying user-generated content.** ● Tools for scrutinising text, images, videos for identification of false information ● Examination of metadata, source details, timestamps, geolocation, to establish content origin and context ● Shared workspaces ● Monitoring and verification of content from popular social media platforms ● Automated ad-hoc verification workflows ● Multilingual support ● Resources for training and guidance to enhance their media verification skills and optimise the tool's functionality ● API Integration	Subscription fees
CrowdTangle[22] **(Discontinued after 14/08/24)**[23]	**Social media monitoring tool used to track performance and engagement of the content through various social media platforms.** ● Nearly Real-time social media monitoring based on keywords, hashtags, accounts, etc. ● Monitoring of trending content and viral posts on multiple platforms (Facebook, Instagram, Twitter, Reddit) ● Aggregated custom dashboard ● Historical data (past trends, change of interests over time) ● Statistics about engagement metrics such as likes, shares, comments, etc.	N/A

[21] Truly Media

[22] CrowdTangle | Content Discovery and Social Monitoring Made Easy

[23] https://transparency.meta.com/researchtools/other-datasets/crowdtangle

Tool	Features	Type of service
Tweetdeck[24]	**Social media management tool for Twitter.** • Tracking of specific feeds, hashtags, mentions, etc. • Nearly Real-time updates • Advanced search and filtering for tracking trends	Subscription fees
Telemetrio[25]	**Telegram chats rating based on subscribers and other statistics.**	Freemium with basic features and subscription tiers to access advanced functionalities
NewsWhip Spike[26]	**Media monitoring and analytics tool.** • Nearly Real-time trend analysis • Cross-platform social media tracking based on topics, keywords, etc... • Predictive insights about future popularity of content and emerging trends • Audience sentiment analysis • Historical data	Subscription-based model
Check by Meedan[27]	**Tool for collaborative verification of online content.** • Nearly Real-time verification within multiple users • Multiple social media platforms integration • Verification workflows and structured check lists • Report generation • Geolocation and timeline analysis for context access • Access to training resources for users to improve verification skills	Open-source software[28]
Google Fact Check tools[29]	**Web search engine for fact checked content** • Debunked stories and images search • ClaimReview markup tool for adding	Free services

[24] Now integrated in Twitter Pro
[25] Analytics Service for Businesses on Telegram - Telemetrio
[26] Predict Content or Stories that Matter Ahead of Time - NewsWhip Spike
[27] Connect with your community on messaging apps with Check tiplines
[28] https://github.com/meedan
[29] Fact Check Tools Recents

Tool	Features	Type of service
Google Trends[30]	**Web tool for tracking and analysing search query popularity and trends.** • Nearly real-time topic tracking • Topic comparison • Filtering by language, geolocation, and time	Free services
Google API Cloud Vision[31]	**Tool for extracting insights from image, videos and documents.** • Image labelling • Face recognition • Points of interest recognition • Optical Character Recognition (OCR) • Tagging of explicit content • Document categorisation • Object tracking • Activity recognition	Pay-as-you-go pricing model
Amazon Rekognition[32]	**Tool for verification of images and videos.** • Facial verification • Face similarity • Face attribute recognition • Content moderation • Custom object detection • Text detection • Object labelling • Video segment detection • Celebrity recognition	Pay-as-you-go pricing model, where users are charged for the specific services they utilise. No upfront commitments or minimum fees
Geolocation[33]	**Tool for geolocation of IP address.** • IP address Information extraction • 2D map visualisation • Nearly Real-time open-source API	Subscription fees for access to the geolocation services and APIs

[30] Google Trends
[31] Vision AI: strumenti di AI visiva e per immagini | Google Cloud
[32] Image Recognition Software, ML Image & Video Analysis - Amazon Rekognition - AWS
[33] Geolocate the Location of an IP Address | Geolocation

Tool	Features	Type of service
NewsGuard (*)[34]	**NewsGuard provides content ratings based on journalistic criteria to combat disinformation by evaluating news sources for credibility. It uses a team of journalists to assess news outlets and individual articles, labelling them for trustworthiness.** ● **Journalist Ratings**: News Guard's human-based assessments of news sources and individual articles. ● **Browser Extension**: Available as a browser plugin, it flags potentially unreliable sources as users browse the web. ● **Content Warnings**: Flags misleading or false information to users. ● **Trust Scores**: Assigns scores to news outlets and articles, showing their reliability. ● **Fact-checking integration**: Works with several major fact-checking organisations for content validation.	Subscription model, offering enterprise licensing to businesses and organisations

[34] NewsGuard - Transparent Reliability Ratings for News and Information Sources

Tool	Features	Type of service

The Factual (*)[35]	**The Factual is a mobile app and browser extension used to score and rank news content based on various quality metrics to help users identify high-quality and credible news sources. It aims to guide readers toward higher-quality information and away from low-quality information.**
	● Scores news content on a 0-100 scale based on several factors:
	  ● **Extent and Quality of Sources:** Evaluates the diversity and reliability of the sources cited in the news content.
	  ● **Expertise of the Journalist:** Assesses the qualifications and expertise of the journalist who authored the content.
	  ● **Opinionated Nature of Language:** Analyses the language used to determine the level of bias or opinionated language.
	  ● **Historical Reputation of the Site:** Considers the historical credibility and reputation of the news site.
	● **Utilises machine learning and AI** to automate the evaluation and scoring process.
	● **Identifies higher quality sources** and sources from different sides of the political spectrum to provide a balanced view
	● **Available as a mobile app and browser extension** for easy access and nearly real-time news quality assessment.
		Subscription model

[35] The Factual: Your Personal News Quality Evaluator - AI Tools | AI Tools

Tool	Features	Type of service
Logically (*)[36]	**Logically combines AI-driven analysis with human expertise to provide comprehensive fact and image verification services, ensuring users receive accurate and reliable information.** • **Fact and Image Verification**: Provides services to verify the authenticity of facts and images. • **AI-Powered Analysis**: Utilises AI to analyse claims, opinions, and events. • **Automated Search Assistant**: Employs AI as part of its automated search assistant feature to help users find accurate information. • **Human Fact-Checkers**: Relies on human fact-checkers to assist in verifying information. • **Nearly Real-Time Monitoring**: Monitors more than one million web domains and social media platforms in nearly real-time. • **Mobile App and Browser Extension**: Available as a free mobile app and browser extension for easy access to its services.	Subscription fees

[36] Logically

Tool	Features	Type of service
Full Fact (*)[37]	**Full Fact is a media company that specializes in fact-checking news stories and public statements. It uses a combination of human expertise and artificial intelligence to identify and debunk false information. The organisation aims to provide accurate and reliable information to the public, helping to combat misinformation and improve the quality of public discourse.** • **Fact-Checking:** Conducts thorough fact-checking on news stories, public statements, and other forms of media to verify their accuracy. • **Automated Tools:** Utilises AI tools to assist fact-checkers in identifying the most important and check-worthy information of the day. • **AI Development:** Working on an algorithm to detect when someone knowingly repeats false information. • **Educational Resources:** Provides resources and guides to educate the public on how to identify misinformation and understand the fact-checking process. • **Collaborations:** Partners with media organisations, tech companies, and other stakeholders to enhance fact-checking efforts and improve information accuracy. • **Misinformation Alerts:** Offers alerts and reports on trending misinformation and disinformation topics. • **Public Reports:** Publishes detailed reports on misinformation trends and the results of their fact-checking efforts.	Grants and donations. Partnerships with media organisations and other stakeholders to support its fact-checking initiatives and operations

[37] Full Fact AI – Full Fact

Tool	Features	Type of service
Hoaxy (*)[38] (discontinued after August 14, 2024)	**Hoaxy is a platform developed by Indiana University to track the spread of misinformation on social media platforms and visualize how false information spreads.** • **Misinformation Tracker**: Analyses how specific disinformation stories spread across social media platforms. • **Visualizing Spread:** Provides interactive visualizations showing how false information travels, including its sources and patterns. • **Fact-checking Integration**: Allows users to see fact-checking responses to viral misinformation. • **Trend Monitoring**: Tracks emerging disinformation trends across various social media networks.	Free access with premium features and research partnerships.
Fabula AI (*)[39] (discontinued following its acquisition by X, formerly Twitter)[40]	**Fabula AI is a startup that focuses on using deep learning algorithms to detect and analyse the spread of disinformation on the internet.** • **Graph Deep Learning:** Utilises advanced graph deep learning techniques to identify unique patterns in the spread of disinformation. • **Disinformation Detection:** Differentiates between the spread of false information and veritable stories. • **Algorithm Development:** Contributes to the development of algorithms aimed at combating false information. • **Integration with X (ex-Twitter):** Enhances X's capabilities in identifying and mitigating the spread of disinformation on its platform.	Research grants and collaborations

[38] Hoaxy
[39] FABULA AI
[40] https://blog.x.com/en_us/topics/company/2019/Twitter-acquires-Fabula-AI

Tool	Features	Type of service
Grover (*)[41]	**Detects and generates false information using AI to combat misinformation.** • **AI-Generated Content Detection:** Identifies AI-generated misinformation by mimicking the language of specific publications. • **Fake Content Generation:** Capable of generating false information to understand and detect it more effectively. • **Language Modelling:** Adapts to the language and style of various publications for accurate detection. • **Research-Based:** Developed by researchers at the University of Washington with a focus on combating misinformation.	Research grants and collaborations
Sensity AI (*)[42]	**Detects deepfakes and assesses the severity of visual threats.** • **Deepfake Detection:** Utilises video forensics and computer vision to identify deepfake images and videos. • **Detection API:** Provides an API for integrating deepfake detection into other applications. • **Visual Threat Assessment:** Evaluates the severity of visual threats posed by deepfakes. • **Nearly Real-Time Analysis:** Offers nearly real-time analysis of images and videos to determine their authenticity.	Subscription model and enterprise solutions

[41] Originality AI Plagiarism and Fact Checker - Publish With Integrity

[42] Sensity AI: Best Deepfake Detection Software in 2025

Tool	Features	Type of service
Claimbuster (*)[43]	**Instant fact-checking of user-provided text and monitoring of political debates.** • **Instant Fact-Checking:** Uses Google Fact-Check Explorer API to verify the veracity of user-provided text. • **Search Results Aggregation:** Gathers search results similar to the written claim and provides truthfulness determinations. • **Political Debate Monitoring:** Utilises AI to highlight claims in political debates that are deemed fact-check worthy.	Subscription model for enterprise solutions
Adverif.ai (*)[44]	**Adverif.ai is an AI-powered service for detecting fake advertisements and inappropriate content.** • **FakeRank Algorithm:** Uses AI to identify potentially harmful content, including fake advertisements and inappropriate material. • **Human Reviewers:** Pairs AI detection with human reviewers to enhance accuracy and reliability. • **Malware Detection:** Identifies links that may contain malware. • **Content Moderation:** Flags potentially explicit or inappropriate thumbnails and content. • **Notifications:** Alerts users about detected harmful content.	Subscription model for enterprise solutions

[43] ClaimBuster

[44] Zefr | Walled Garden Brand Suitability

Tool	Features	Type of service
Alto Analytics (*)[45]	**Provides actionable insights about disinformation and deepfakes using artificial intelligence.** • **Alto Analyzer:** Maps and clusters digital conversations around specific subjects. • **Multilingual Analysis:** Analyses online conversations in more than 50 languages. • **Targeting Insights:** Helps users understand where to target advertising or communications. • **Alto Insights:** Uses machine learning to provide visual data reports that are actionable and accessible.	Subscription Model
Blackbird AI (*)[46]	**Detects and analyses adversarial and deceptive content using AI-powered algorithms.** • **Deception Detection:** Identifies conversations with signs of adversarial conflict. • **Sentiment Analysis:** Analyses and detects negative perceptions about organisations. • **Nearly Real-Time Monitoring:** Monitors social media and other platforms for deceptive content. • **Customizable Alerts:** Provides alerts for specific types of adversarial content. • **Dashboard:** Offers a dashboard for tracking and analysing deceptive content.	Subscription model for enterprise solutions.

[45] Alto Intelligence

[46] Blackbird.AI | Narrative & Risk Intelligence Platform

Tool	Features	Type of service
Defudger (*)[47]	**Defudger is an AI-powered solution for the authentication of visual content, detecting manipulation in videos and images.** • **Manipulation Detection**: Identifies edited images and deepfake videos. • **Blockchain Authentication**: Uses blockchain technology to authenticate original visual content. • **Content Database**: Maintains a database of visual content validated as authentic using blockchain technology, preventing duplicates or altered content from being passed as authentic.	Subscription model for enterprise solution
Bot Sentinel (*)[48]	**Identifies and calls out social media accounts that are bots on Twitter using AI and machine learning.** • **Bot Detection:** Uses AI and machine learning to identify bot accounts on Twitter. • **Behaviour Analysis:** Analyses behaviours prohibited by Twitter to classify accounts. • **Accuracy:** Claims a 95% accuracy rate in classifying bots. • **User Dashboard:** Provides a dashboard for users to monitor and report suspicious accounts. • **Reports and Alerts:** Generates reports and alerts for identified bot activities.	Donations

Table 1: *Comparative analysis results (The tools added to the previous analysis are indicated with an asterisk "*")*

The **solutions currently available on the market** include:

- **Monitoring and Analysis Platforms**: Systems designed to monitor social media platforms in nearly real-time and analyse content to identify false information. Solutions include continuous nearly real-time monitoring, data analysis with advanced algorithms, and visualisation tools to quickly identify false information trends. These platforms are typically ideal for media professionals, policymakers, and businesses requiring constant and detailed

[47] defudger.com

[48] Bot Sentinel - Dashboard

monitoring. They can assist journalists in fact-checking news, policymakers in managing disinformation crises, and businesses in protecting their online reputation.

- **Fact-Checking Tools**: Solutions designed to assist human fact-checkers in identifying and validating false information. These tools are characterised by AI algorithms for natural language processing, databases of verified facts, and integration with news platforms. They are essential for journalists and newsrooms needing to quickly verify information before publication. They are also useful for governmental entities and non-governmental organisations working to combat false information on a large scale.
- **Educational Tools**: Designed to educate the public on recognising false information and on the importance of digital literacy, these tools include interactive educational content, teaching materials, and resources to increase users' awareness and critical thinking skills. They are generally aimed primarily at end consumers who can benefit most from learning the skills necessary to navigate the contemporary information landscape.

AI4TRUST distinguishes itself within the competitive landscape of false information detection and fact-checking tools by addressing several critical limitations prevalent among existing platforms. Unlike many competitors, AI4TRUST prioritises **transparency**, ensuring that content contributed by fact-checkers remains their **intellectual property (IP)** and is not utilised for **training purposes** without appropriate **compensation**. This approach fosters **trust** and encourages greater user participation. Moreover, AI4TRUST offers a **comprehensive suite of tools** that extends beyond traditional **user-generated content verification**, incorporating **advanced AI-driven analytics** to detect **deepfakes** and **false information** across multiple media formats. Its **multi-modal analysis capability**, encompassing **text, images, and videos**, sets it apart from competitors that often lack such an integrated approach.

Additionally, AI4TRUST addresses the widespread challenges of **unsupported languages** and the growing complexity of **regulatory frameworks** governing **data collection** — key concerns for **researchers** and **policymakers**. By providing **multilingual support** and a **robust compliance framework**, AI4TRUST enhances its **usability** and **effectiveness** across diverse contexts, positioning itself as a **leading solution** for stakeholders seeking **reliable and comprehensive** tools in the fight against **false information**.

2.4. Target Customers

The market for **advanced Artificial Intelligence (AI) solutions** designed to combat **false information** can be effectively **segmented** using various criteria, enabling a more precise identification of **customer needs** and a targeted approach to addressing them. The **market segmentation** for the **AI4TRUST** solution is strategically structured to facilitate its **commercial exploitation** by aligning with the specific requirements of different **potential target customers**.

As illustrated in **Figure 5**, the potential **target customers** can be categorised into two main groups: **1) Primary target customers; 2) Secondary target customers.**

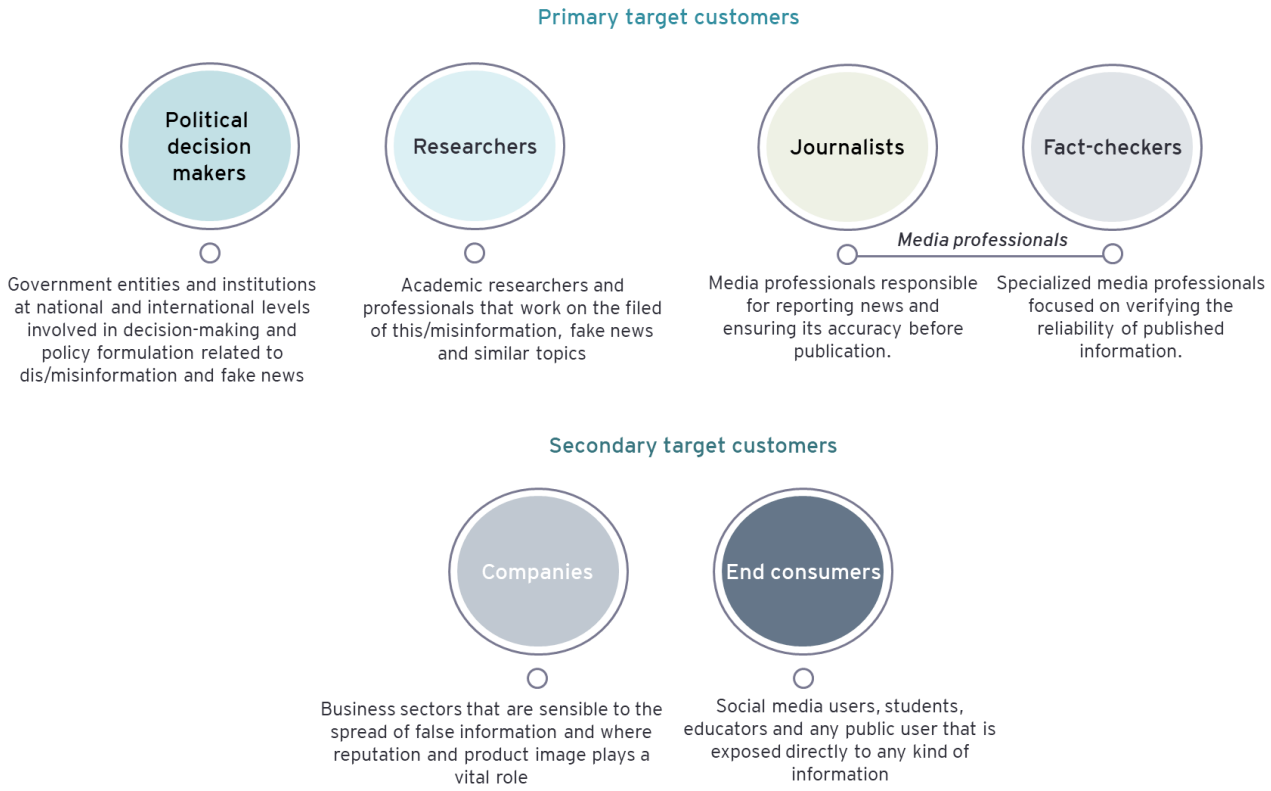

Figure 5: *Target customers*

The distinction between primary and secondary target customers within the AI4TRUST framework is based on the direct impact and involvement each group has in addressing false information. **Primary target customers**, such as **fact-checkers, journalists, researchers, and policymakers**, are directly engaged in the core activities of monitoring, analysing, and formulating responses to false information. Their work is fundamental to the platform's mission of enhancing fact-checking and developing effective countermeasures. **Secondary target customers**, including **companies and end consumers**, are indirectly affected by the outcomes of these efforts. Companies may utilise the insights generated by the platform to protect their brand reputation and navigate public perception, while end consumers benefit from the overall improvement in information quality and trustworthiness in the media landscape.

The following **primary target customers** have been identified:

<u>Media Professionals</u>

Media Professionals encompass **all individuals operating within the media landscape**, responsible for ensuring the dissemination of accurate information to the public. Their primary needs revolve around **content verification, false information detection, and the ability to monitor**

news items effectively. The **AI4TRUST Platform** directly addresses these needs by offering advanced tools that enable **AI-driven content analysis**. Specifically, the platform facilitates the **detection of false information signals** across **text, images, and audio**, equipping media professionals with the necessary resources to assess the **credibility of information** encountered in their work. This is particularly critical in an era where **false information spreads rapidly** across multiple channels.

Furthermore, the platform's **monitoring dashboard** enhances users' ability to **track and evaluate risks** associated with specific news items, ensuring that media professionals can **uphold journalistic integrity** and deliver **reliable information** to their audiences. Given that the **reputation and credibility** of media organisations depend on their ability to **provide accurate and timely news**, these professionals require **rapid and precise fact-checking solutions**. The **fast-paced** nature of their work necessitates tools capable of processing **large volumes of data in real time**, ensuring that content is **accurately verified before dissemination**. This includes the capacity to **analyse texts, images, videos**, and other forms of **multimedia content** in **multiple languages** to identify any **falsehoods or manipulations**. Additionally, media professionals require quick access to **reliable sources and verified fact databases** to support their **verification processes**.

Advanced AI solutions offer significant value by improving the **accuracy and efficiency** of information verification. Leveraging **machine learning algorithms** and **natural language processing technologies**, these solutions can swiftly analyse content and detect **potential misinformation**. By integrating such tools into their **daily workflows**, newsrooms can **reduce the risk of publishing false information**, thereby safeguarding their **reputation** and maintaining **public trust**. Moreover, AI-powered solutions can **automate parts of the verification process**, allowing journalists to focus on **quality journalism** rather than **repetitive tasks**. With access to **real-time reports** and **in-depth analyses**, media professionals can make **informed decisions more efficiently**, enhancing the **timeliness of publications** without compromising **accuracy**.

Media Professionals can be divided into:

- **Fact-checkers:** they represent a **specialised subset** of media professionals whose **primary role** is to **verify the reliability** of published information. Their core needs revolve around **access to reliable tools** that facilitate the **rapid validation of content**.The **AI4TRUST Platform** addresses these needs through advanced **automated data collection** from **diverse sources**, streamlining the **information-gathering process** for verification. Additionally, the platform provides **text analysis functionalities** to assess the **check-worthiness** of content and retrieve **previously fact-checked claims**, offering fact-checkers essential resources to **substantiate their evaluations**. Moreover, the **human validation dashboard** enables **manual content review**, complemented by **automated insights**, thereby enhancing the **accuracy and reliability** of fact-checking processes. This **hybrid approach**, combining **AI-driven analysis** with **human expertise**, ensures **more effective verification** and strengthens the fight against **false information**.

- **Journalists**: The media professionals segment includes **journalists working in newsrooms and press agencies**, operating in a highly dynamic environment characterised by a **continuous flow of information** and **constant pressure** to publish content swiftly. In this context, having **reliable verification tools** is crucial to ensuring the **accuracy** of information before publication. While journalists, like **fact-checkers**, require tools for **information verification**, their needs extend beyond fact-checking to encompass the broader **news reporting process**. Their **primary requirement** is access to **already verified news**, allowing them to confidently report information with full awareness of their **responsibility** for its accuracy and implications. The **AI4TRUST Platform** supports journalists by providing **comprehensive news analysis**, including the **detection of sensational content** and **deepfakes**. Its **advanced capabilities** in **visual and audio content analysis** help ensure that the materials used in reporting are **authentic and credible**. Furthermore, the **monitoring dashboard** enables journalists to stay informed about **emerging trends** and **potential false information**, allowing them to produce stories that are both **accurate and relevant** to current events.

<u>Policymakers</u>

The policymakers segment comprises **government entities and institutions at national and international levels**, responsible for **monitoring and countering disinformation**. These entities operate in a **complex and high-stakes environment**, where the **spread of false information** can have serious implications for **national security, political stability, and public trust**. The ability to **identify and respond swiftly** to disinformation is crucial for **maintaining public order and protecting citizens**. Government entities require **advanced tools** to monitor **information dissemination online** and make **informed decisions** on **regulation and crisis management**. They need solutions capable of **analysing vast amounts of data** from diverse sources, including **social media platforms, news websites, and blogs**, to swiftly identify **potentially harmful information**. Additionally, they must **assess the impact** of disinformation and develop **strategic interventions**. **Timely access to updated analyses and detailed reports** is fundamental for ensuring the **effectiveness** of their actions.

Hybrid AI and Machine Learning-based solutions provide significant value to **policymakers** by **enhancing monitoring and analytical capabilities**. These technologies employ **tailored algorithms** to collect and process data **in near real-time**, offering a **comprehensive and up-to-date** overview of the **information landscape**. By analysing **detailed reports**, decision-makers can gain a deeper understanding of **disinformation dynamics**, identify **sources of false information**, and evaluate **their spread and impact**.

The **AI4TRUST Platform** supports **policymakers** by streamlining the **information verification process**, providing access to **reliable intelligence**, and offering **innovative AI-driven tools** to **combat mis-, dis-, and malinformation**. Moreover, the Platform facilitates knowledge-sharing by enabling **comparative insights into best practices** from **European organisations** tackling disinformation. The **Collective Analysis of Social Media Actors and Items** feature within AI4TRUST empowers **policymakers** to: 1) **Evaluate social networks** through **Social Network Analysis**; 2) **Assess the Reliability State of Social Media**; 3) **Detect Coordinated Inauthentic Behaviours**; 3) **Analyse language-specific disinformation trends** via the **Infodemic Observatory**;

4) **Generate semi-automated reports with mitigation guidelines** using the **Recommendation Tool.** By leveraging these capabilities, **policymakers** can implement **more effective policies**, introduce **targeted regulations**, and coordinate **rapid, informed responses** during crises. Furthermore, the **continuous monitoring** enabled by AI4TRUST strengthens **policy adaptability** and enhances overall **resilience against disinformation campaigns**.

Researchers

The researchers segment comprises a **broad network of academic and related professionals** engaged in fields associated with **mis-, dis-, and malinformation**. This group plays a **critical role** in the broader effort to **counter the spread of misleading content**. Researchers operate within an **academic ecosystem** where **interdisciplinary collaboration** and access to a **shared pool of knowledge** are essential.

The **continuous cycle** of **evaluation, publication, and peer review** in **academic journals and conferences** ensures the **advancement of knowledge** in this field. Furthermore, the ability to access **large-scale datasets** from diverse sources is instrumental in **enhancing research quality** and fostering the **development of cutting-edge solutions** to mitigate the spread of disinformation. The **AI4TRUST Platform** provides **significant benefits** to researchers by offering:

- **Access to human-validated results** to support **robust and reliable** research findings;
- **A comprehensive dataset of collected news items** sourced from **social listening data streams;**
- **An interdisciplinary environment** that encourages **collaboration with industry partners** and other academic institutions;
- **A shared pool of knowledge**, fostering the **co-development of innovative methodologies.**

By leveraging these resources, researchers can **experiment with new approaches**, **develop advanced technological solutions**, and generate **impactful research outcomes**. The AI4TRUST Platform ultimately facilitates **the dissemination of findings** in **high-impact academic journals and conferences**, reinforcing **future research projects and funding opportunities**.

The following **secondary target customers** have been identified:

Companies

The **business segment** includes industries such as **healthcare, finance, and technology**, which are particularly **susceptible to reputational damage** resulting from the spread of **false information**. Operating in **highly competitive and regulated markets**, these companies depend on **customer and stakeholder trust** for their **long-term success**. The proliferation of **disinformation** poses significant risks, including **reputational harm, financial losses, and regulatory non-compliance**. To effectively mitigate these risks, businesses require **advanced tools** capable of **real-time brand monitoring** and **rapid response** to emerging threats. This necessitates solutions that can process **large volumes of data** from diverse online sources — including **social media, forums, blogs, and news websites** — to detect early signals of **disinformation** and enable swift intervention. The **AI4TRUST Platform** provides **substantial value** to companies by offering:

- **Automated data collection tools** to monitor **consumer opinions** and **brand-related discussions;**
- **Sentiment analysis and content classification** to gain **insights into users' emotional responses;**
- **Real-time alerts** on potential reputational threats, allowing businesses to **intervene proactively;**
- **Detection of mis-, dis-, and malinformation** to **counteract harmful narratives** before they escalate;
- **Reverse video search and deepfake detection**, assisting in **debunking manipulated content** and uncovering attempts to **discredit brands.**

By leveraging these **AI-driven solutions**, businesses can not only **protect** but also **enhance** their **corporate reputation**. Furthermore, **predictive analytics** can help companies **anticipate potential crises**, affording them the time to **develop effective communication strategies** and implement **preventive measures**. The **AI4TRUST Platform** ultimately **empowers organisations** to navigate the **complex digital landscape** with **greater confidence and resilience** against the challenges of **disinformation**.

End Consumers

The end consumer segment encompasses a diverse range of users, including **social media users, students, educators, and the general public**. In an era defined by an **incessant flow of digital information**, these groups are **highly exposed** to **disinformation and manipulated content**. The ability to **distinguish between accurate and false information** has become an **essential skill** for navigating today's complex **information landscape**. End consumers require **accessible and user-friendly tools** to **enhance their digital literacy**, specifically their ability to **critically evaluate online content** and **identify false information**. These tools must provide **support in recognising, verifying, and contextualising information** in an intuitive manner. The **AI4TRUST Platform** delivers **substantial value** to end consumers by offering:

- **Automated fact-checking tools** to verify the authenticity of **online content;**
- **Deepfake detection and document intelligence features**, assisting in **identifying manipulated media;**
- **Real-time alerts** on potential false information to **increase user awareness;**
- **Guidance on assessing source credibility**, empowering individuals to **differentiate between reliable and unreliable sources;**
- **Insights on misinformation trends**, allowing users to stay **informed about evolving disinformation tactics.**

By integrating **AI-driven solutions** into **everyday digital interactions**, end consumers can become **more resilient** to the **spread of false information**. Furthermore, the **AI4TRUST Platform** supports **public awareness initiatives** by **analysing disinformation patterns** and **contributing to digital literacy campaigns**. Ultimately, these technologies **equip users with the skills and knowledge** necessary to **critically assess online information**, fostering a more **informed and vigilant society**.

The **interconnections between the various target customers** — both primary and secondary — in the context of the AI4TRUST solution are crucial for creating a cohesive ecosystem aimed at combating false information and enhancing the integrity of information dissemination.

On the first level, **Media Professionals** (fact-checkers and journalists) serve as the frontline defenders against false information. Their work directly impacts the quality of information that reaches the public. Fact-checkers validate news and provide accurate information, while journalists report on current events and trends. The insights generated by fact-checkers can inform journalists' narratives, ensuring that the content they produce is not only engaging but also factually reliable. This collaboration is essential, as it creates a feedback loop where journalists rely on the rigorous validation processes of fact-checkers to enhance their reporting, while fact-checkers can draw on the real-world implications of the news they assess.

Researchers play a vital role in understanding the broader implications of false information and developing methodologies to combat it. They analyse data collected through the AI4TRUST Platform to identify patterns in mis/disinformation, assess the effectiveness of various countermeasures, and contribute to the development of new tools and techniques. Their findings can inform both media professionals and policymakers, providing evidence-based insights that guide strategies for addressing mis/disinformation. This interconnection fosters a collaborative environment where research informs practice, and media professionals can leverage academic insights to enhance their work.

Policymakers rely on accurate information to craft policies and regulations that address the challenges posed by mis/disinformation. They benefit from the analytics and insights provided by the AI4TRUST Platform, which help them understand the landscape of false information and its impact on society. By collaborating with media professionals and researchers, policymakers can develop informed strategies that not only address immediate concerns but also promote long-term solutions for enhancing information integrity. This relationship is symbiotic, as effective policies can empower media professionals and researchers to operate more effectively in their respective roles.

On the secondary level, **companies** are influenced by the work of media professionals and policymakers. They require accurate information to manage their brand reputation and respond to mis/disinformation that may affect their public image. By utilising the insights generated by the AI4TRUST Platform, companies can monitor their online presence, understand consumer sentiment, and engage proactively with their audience. This connection highlights the importance of media professionals in shaping public information, as their reporting can significantly impact how companies are perceived.

End Consumers are the ultimate beneficiaries of the interconnected efforts of all target customers. They rely on media professionals to provide accurate information and on researchers to develop tools that enhance the quality of content available to them. Policymakers play a role in ensuring that the regulatory environment supports transparency and accountability in information dissemination. The AI4TRUST Platform empowers end consumers by equipping them with the

knowledge and tools to critically evaluate the information they encounter, fostering a more informed public.

The interconnections among the target customers within the AI4TRUST framework create a **collaborative ecosystem** where media professionals, researchers, policymakers, companies, and end consumers work together to combat mis/disinformation. Each user type contributes to and benefits from the collective efforts to **enhance information integrity**, ultimately leading to a more informed society. This synergy is essential for addressing the complexities of false information in today's digital landscape.

2.5. Value Chain

The concept of the **value chain** refers to the series of activities and processes that organisations engage in to deliver a product or service to the market, ultimately creating value for customers and stakeholders. In the context of the AI4TRUST solution, the value chain encompasses the collaborative efforts of various target customers — media professionals, researchers, policymakers, companies, and end consumers — each contributing to the overarching goal of combating false information and enhancing the integrity of information dissemination. By leveraging the capabilities of the AI4TRUST Platform, these target customers engage in a sequence of activities that not only improve their individual outputs but also collectively elevate the quality of information available to the public.

Media Professionals, including fact-checkers and journalists, are at the forefront of the AI4TRUST value chain. **Fact-checkers** utilise the Platform to access advanced analytical tools that allow them to validate news and assess the credibility of information. By leveraging the AI-driven data analysis methods available through AI4TRUST, fact-checkers can efficiently identify false information signals in text, images, and audio. This capability not only streamlines their verification processes but also enhances the accuracy of the information they provide to the public. As a result, the work of fact-checkers contributes to a more informed society, as they ensure that only verified information reaches consumers. The value added here is significant: by utilising AI4TRUST, fact-checkers improve the reliability of news content, thereby fostering trust in media sources.

Journalists, on the other hand, benefit from the insights generated by fact-checkers and the analytical capabilities of the AI4TRUST Platform. They can access validated information and utilise tools that help them identify sensational content and deepfakes. This empowers journalists to produce high-quality reporting that is both engaging and factually accurate and reliable. The integration of AI4TRUST into their workflow enhances their ability to monitor emerging trends and false information, allowing them to craft stories that resonate with the public while maintaining

journalistic integrity. The added value for journalists lies in their ability to deliver trustworthy news, which in turn strengthens their credibility and the overall quality of journalism.

Researchers contribute to the value chain by analysing data collected through the AI4TRUST Platform to identify patterns and trends in false information. Their work informs the development of new methodologies and tools that enhance the Platform's capabilities. By collaborating with media professionals, researchers ensure that the insights they generate are relevant and actionable, ultimately improving the effectiveness of the AI4TRUST solution. The value added by researchers is evident in their ability to provide evidence-based recommendations that guide the strategies of media professionals and policymakers, thereby enhancing the overall impact of the Platform.

Policymakers (Policymakers) rely on the analytics and insights provided by AI4TRUST to craft policies that address the challenges posed by mis/disinformation. By understanding the landscape of false information and its societal implications, they can develop informed strategies that promote transparency and accountability in information dissemination. The collaboration between policymakers and media professionals ensures that policies are grounded in real-world data, which enhances their effectiveness. The added value for policymakers is the ability to implement policies that not only mitigate mis/disinformation but also empower media professionals and researchers to operate more effectively in their roles.

Companies engage with the AI4TRUST Platform to monitor their brand reputation and respond to mis/disinformation that may affect their public image. By utilising the monitoring tools and analytics provided by AI4TRUST, companies can track discussions related to their brand across various media channels, allowing them to identify potential threats and engage proactively with their audience. This capability enhances their ability to manage their reputation and build consumer trust. The value added for companies lies in their capacity to navigate the complexities of information dissemination, ensuring that they can respond to challenges effectively and maintain a positive public image.

End Consumers are the ultimate beneficiaries of the interconnected efforts of all target customers within the AI4TRUST value chain. They rely on the work of media professionals to access accurate information and on researchers to develop tools that enhance the quality of content available to them. The insights generated by AI4TRUST empower end consumers to critically evaluate the information they encounter, fostering a more informed public. The added value for end consumers is the assurance that they are receiving reliable information, which enables them to make informed decisions in their daily lives.

During the **workshop we conducted on January 9th, 2025** (see above and chap. 3 "Methodology"), stakeholders provided valuable insights into the impact of the AI4TRUST solution on the value

chain for the potential primary target customer groups identified (fact-checkers, journalists, policymakers, and researchers).

Fact-checkers emphasised the **critical importance of monitoring social networks** for **false information**, enabling them to make **informed decisions** on which content to investigate further. They highlighted the need for **specialised tools** that facilitate the **identification of tampered objects**, allowing for the **efficient assessment of media authenticity**. Additionally, fact-checkers stressed the necessity of being able to **fact-check content and review results directly within the AI4TRUST Platform**, thereby **streamlining their workflow** and enhancing their ability to **provide accurate information to the public**.

Journalists articulated the need for a solution that enables them to **verify information quickly and accurately**, particularly in the context of **rapidly spreading viral content**. Furthermore, they **underscored the importance of correctly interpreting verification results**, ensuring that findings can be effectively communicated to their audiences while **maintaining journalistic integrity**.

Policymakers conveyed the necessity of being able to **anticipate the evolving landscape of misinformation, malinformation, and disinformation**. They highlighted the need for **timely insights** that can **inform policy decisions** and enable them to **respond effectively** to emerging threats within the **information ecosystem**.

Researchers, including **mathematicians and sociologists**, emphasised the importance of **accessing regulation-compliant data** for the **training, evaluation, and fine-tuning** of AI methodologies. They expressed a need for access to **indices of verified and manipulated content**, as well as **indexed fact-checked and labelled data**. **Sociologists** specifically noted the value of obtaining **analytics on the spread of false information**, which would **enhance their understanding of the information landscape** and **support their research efforts**.

In summary, **AI4TRUST establishes a robust value chain**, connecting diverse **target customers**, each contributing to the **overarching goal** of **combating false information** and **enhancing information integrity**. The **collaborative efforts** of media professionals, researchers, policymakers, companies, and end consumers culminate in a **more informed society**, where **trust in information sources is restored**, and the **negative impacts of false information** are mitigated.

The **AI4TRUST project** not only **enhances the capabilities of individual stakeholders** but also **strengthens the overall information ecosystem**, fostering a **resilient and transparent** digital environment. By **promoting accountability and trust** in information sources, AI4TRUST ultimately contributes to **a more reliable and well-informed public discourse**.

3. Methodology

The overall objective of Task "T7.2 – Exploitation strategy and innovation management" within WP7 of AI4TRUST is to guarantee a **comprehensive and strategic approach to innovation, exploitation, and sustainability**. To achieve this objective, the **methodology** employed during the initial phase of the activity has outlined four main phases:

- **Overall overview and key exploitable results of the AI4TRUST Project.** In this activity, the **main functionalities** of the AI4TRUST Platform are identified and grouped to obtain a clear overview of the Platform capabilities. This has involved multiple partners in the definition of the specific functionalities and in their harmonisation, in order to provide the end-users with a consistent and effective solution for their specific needs. This activity has been fundamental for sharpening the goal and the vision of the AI4TRUST Platform, but also for the **mapping between the described functionalities and the key exploitable results**, that have been identified and analysed in the subsequent activities. The results of this activity can be found in the "Overview of AI4TRUST Platform" section (sec. 4) and "Key Exploitable Results" section (sec. 4.1).
- **Initial Innovation, Exploitation and Sustainability Strategy.** This activity gives a preliminary overview of the **initial innovation, exploitation and sustainability plan**. The objective is to formulate a **comprehensive macro strategy** that integrates innovation, exploitation, and sustainability efforts across all assets and partners, taking into account the different segments that stakeholders belong to, to ensure exploring various market segments and needs and different possible business models. An engagement strategy has been devised to collect feedback from internal and external stakeholders, fostering **synergies and opportunities for collaboration,** and promoting the results obtained within the project, enhancing its expected impacts. Moreover, an Intellectual Property Rights (IPR) management plan is presented to ensure proper management of intellectual property rights related with the development of Platform components. The envisioned strategy is presented in the "Innovation and Exploitation Strategy" (sec. 5), "Sustainability Strategy" (sec. 6) and "IPR Management" (sec. 7) sections.
- **Individual Innovation, Exploitation, and Sustainability Plan for Each Asset of Each Partner.** This activity aims to **classify the individual assets based on relevance to the AI4TRUST project** with potential impact on innovation, exploitation, and sustainability. The work carried out was based on a comprehensive evaluation process through the **consultation of all consortium partners** for discerning and documenting the specific assets of interest for each of them trough **semi-structured interviews**. This detailed assessment aimed to identify the diverse resources and capabilities of each entity involved in the project. Thanks to those interviews, the activity has highlighted not only **tangible assets** but also **intangible knowledge**, expertise, and unique contributions that can significantly impact the success of the overall

exploitation strategy approach, by underlying strategic advantages that can be leveraged for optimal project outcomes. This **interview-based assessment** represents a dynamic tool that will follow the project lifecycle and will enable a continuous and adaptive optimisation of resources and capabilities for the overall success of project outcomes. The following Table 2 represents the **interview structure and information requested to partners**. **Assets** identified by interviews with consortium partners have been classified in **four main categories** (Figure 6).

QUESTIONS
1. **Exploitable assets and results**: Describe what assets you intend to exploit (whether this involves specific components, tools, knowledge, methodologies, skills, etc.). ASSET 1: • Description • Target Groups / beneficiaries • Innovation, Exploitation & Sustainability Plan: (explain innovation outcomes from the identified asset in terms of results or knowledge that could be exploited and describe the timeline plan you have for using the asset. Provide concrete actions) ASSET N: • Description • Target Groups / beneficiaries • Innovation, Exploitation & Sustainability Plan: (explain innovation outcomes from the identified asset in terms of results or knowledge that could be exploited and describe the timeline plan you have for using the asset. Provide concrete actions)
2. **Partner expertise toward assets innovation**: Explain in detail how your skills align with the previously mentioned assets and articulate how your expertise will drive the innovation of project outcomes associated with those assets.
3. **Rationale**: Explanation of why you are interested on those assets (the added value they provide), how do you plan to exploit them (academically or industrially: e.g., provide as commercial solution, certification services, standardisation, consultancy, further R&D, positioning).
4. **Opportunity which appeared/appears**: your participation is the result of the real need of your customers (for industrial partners) or internal needs (for user partners). For academic partners, mention if AI4TRUST is in line with other projects (continuation) and reuse of know-how. Are there any other opportunities in the pipeline when the project finishes?

QUESTIONS	
	5. **Your Value Proposition towards Joint Exploitation of AI4TRUST outcomes**: what do you expect from project partners, what benefits will you deliver to the rest of the consortium, what components/interest do you share with other partners.

Table 2: *Interview Structure*

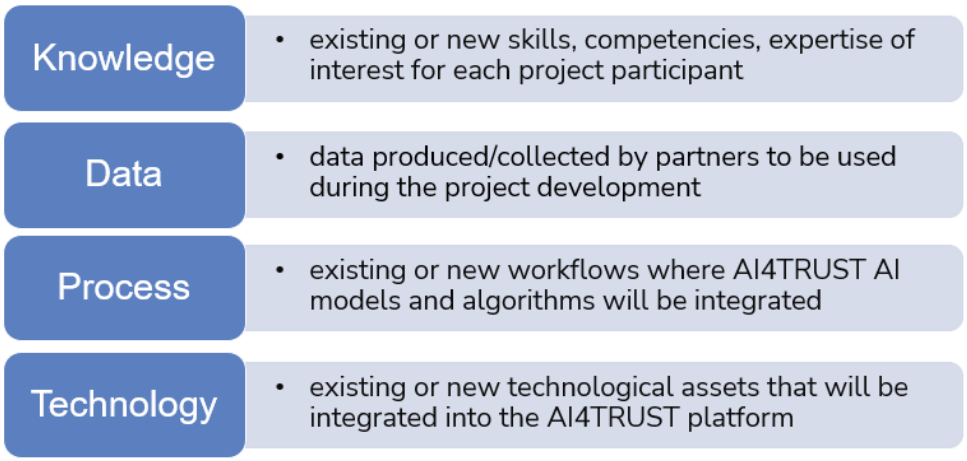

Figure 6: *Assets Macro Categories*

After the identification of the individual assets, each partner was asked for **customised plans for innovation, exploitation, and sustainability for each identified asset**, to maximise the potential of each asset within the project consortium resources. This outlined strategies for enhancing innovation, maximising exploitation potential, and ensuring long-term sustainability for each asset. The purpose is to understand the unique characteristics of each asset, involving not only identifying the technological advancements but also recognizing the innovative potential inherent in each asset. The potential exploitation of assets needs to involve strategically capitalising on the identified assets **to maximise their potential values**. This entails **drawing a pathway and a go-to market strategy**, underlining **opportunities**, and establishing **strategic partnerships and business models**, starting from the exploitation strategy of each asset. Finally, the integration of sustainability into the plans is **crucial for long-term viability**. Sustainability considerations go beyond the immediate financial gains and require a **holistic approach** ensuring that the project outcomes will contribute to a sustainable and ethically responsible future. The results of this activity can be found in the "Individual Innovation and Exploitation per Asset" section (sec. 5.1 and Annex II).

To complement these activities, **on the 9th of January 2025 a workshop with the end-user partners of the project** was organised to collect their point of view about the AI4TRUST Platform. During the session, partners have been grouped according to the four different Target Customers

identified, ensuring a user-centric approach to the Platform's development. Each group has been tasked with filling out four distinct canvases:

- **Persona Canvas**: it is a canvas for bringing customer segments to life by giving them a name, role, and identifiable characteristics, making it easier to empathize with their needs and experiences. It helps create a visual representation of the persona, depicting their emotions, appearance, and context. By exploring key aspects such as needs, hopes, and fears, the canvas helps identify what truly matters to the persona and what influences their decisions. It also examines both positive and negative trends in their life, highlighting opportunities that could enhance their experiences and challenges that may create obstacles. Through this structured approach, the Persona Canvas (Figure 7) facilitates the creation of shared mental models, enabling more effective communication about different customer types and tailor solutions that align with their expectations.

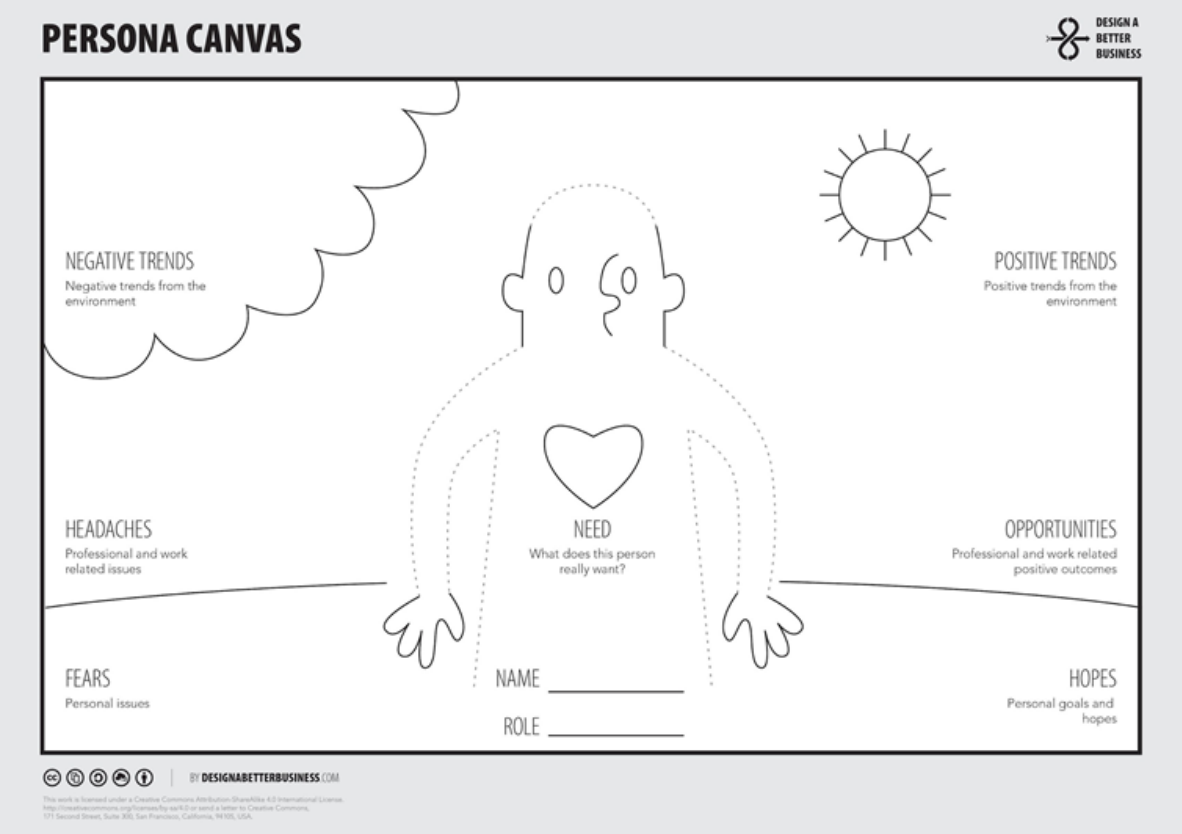

Figure 7: *The Persona Canvas (from DESIGNABETTERBUSINESS.COM)*

- **Value Proposition Canvas**: it is a canvas designed to help deeply understand the customers by mapping their needs, challenges, and desired outcomes against the products and services being offered. It focuses on identifying the jobs-to-be-done, whether functional, social, or emotional, that customers are trying to accomplish in their work or personal lives. By analysing pains, such as obstacles and frustrations preventing task completion, and

gains, which include desired benefits and positive outcomes, it helps refine the value propositions to better align with customer expectations. The canvas also explores pain relievers — specific ways a product or service can alleviate customer struggles — and gain creators, which highlight how offerings can enhance positive experiences. Finally, it defines the products and services that effectively address customer needs, ensuring a well-rounded and compelling value proposition. By systematically aligning these elements, the Value Proposition Canvas helps develop solutions that truly resonate with the target audience (Figure 8).

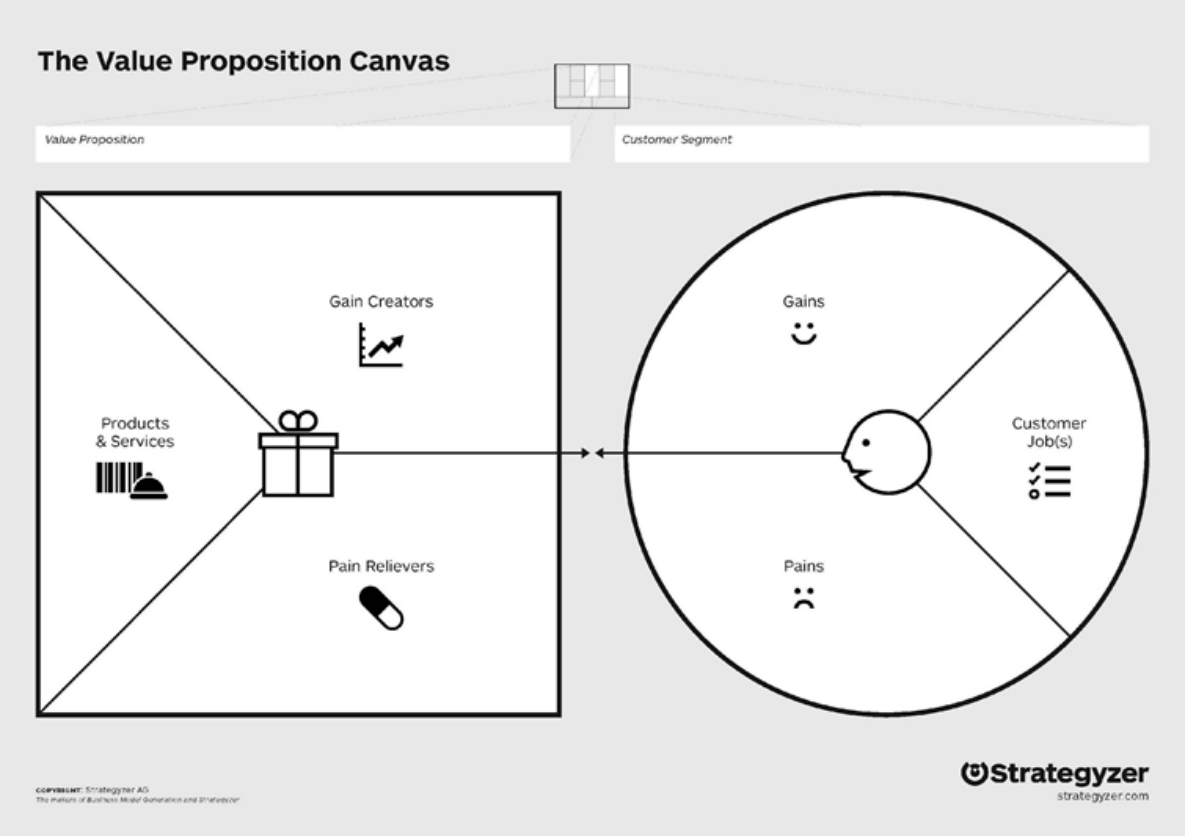

Figure 8: *The Value Proposition Canvas (from DESIGNABETTERBUSINESS.COM)*

- **Ad Lib Value Proposition Template**: it is a structured template designed to quickly craft and refine different variations of a value proposition by synthesizing key insights from the Value Proposition Canvas (Figure 9). It helps clearly articulate how to create value by filling in predefined blanks that capture essential elements such as customer needs, pains, gains, and the solutions offered. By forcing a concise and precise formulation of the value proposition, this template enables the exploration of multiple strategic directions efficiently, testing and iterating on different approaches. By generating alternative versions, it helps compare and refine the messaging, ensuring communicating the value in the most compelling and customer-centric way.

Figure 9: *The Ad-Lib Value Proposition Canvas (from Strategyzer AG)*

- **Prototype Canvas**: it is a canvas designed to transform value propositions into tangible concepts by outlining and testing key aspects of a product or service before full development. It focuses on defining the customer's job-to-be-done, ensuring that the prototype addresses a relevant need or problem. By identifying customer benefits, the canvas clarifies what value the product delivers and what will make users genuinely satisfied. The customer promise captures the core commitment of the product, aligning it with expectations set in the Value Proposition Canvas. Key functionalities are mapped in the key features section, prioritising essential elements while avoiding unnecessary complexity. Finally, the steps section outlines the minimum actions a customer must take to achieve their goal, refining the user experience to ensure efficiency and ease of use (Figure 10).

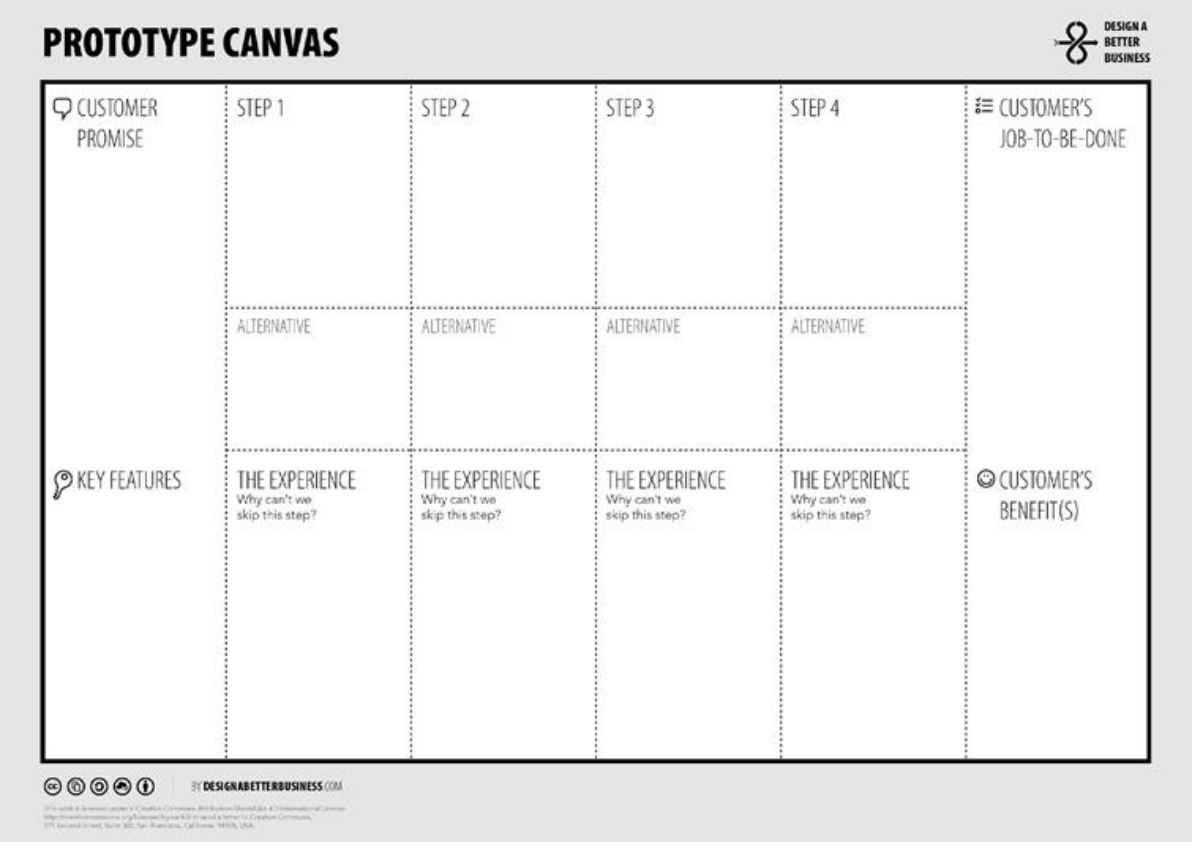

Figure 10: *The Prototype Canvas (from DESIGNABETTERBUSINESS.COM)*

The outcomes of the workshop are documented in the Annex I of this deliverable, while their synthesis is included in the relevant sections below.

4. Innovation and Exploitation Strategy

This section describes the **strategies for innovation and exploitation** adopted by the AI4TRUST project. The defined innovation and exploitation strategy is based on the internal and external stakeholder feedback implementation to enhance marketability and sustainability. This is achieved by adapting different approaches for professionals, researchers, and policymakers to address unique needs and barriers.

This activity involves **the integration of relevant feedback from stakeholders** to enhance the **exploitation and marketability of the results achieved** through the AI4TRUST project to ensure sustainability of the results over the time. This process includes conducting both **internal and external interviews** to gather insights, opinions, and suggestions from key stakeholders such as professionals, researchers, and policymakers. Internal interviews may involve discussions with project team members, researchers, and developers involved in the AI4TRUST project. These interviews aim to gather perspectives on the strengths, weaknesses, opportunities, and challenges associated with the developed tools, technologies, and strategies within the project. External interviews, on the other hand, involve engaging with external stakeholders such as professionals in the media industry, fact-checking organisations, policymakers, and other relevant entities.

The goal is to obtain external viewpoints on the practicality, effectiveness, and market demand for the outcomes of the AI4TRUST project. By integrating feedback coming from internal and external interviews, it will be possible to identify **barriers** that may fight the successful exploitation and adoption of the project results, as well as the **facilitators** that can enhance results marketability. The **outcomes of interviews** conducted with internal and external stakeholders, coupled with the identification of barriers and facilitators, marked the **initial stage in driving the exploitation and innovation plan** at the project's overarching level. It will be crucial to keep monitoring this activity throughout the project duration to enable **necessary adjustments, enhancements, or customisations** based on real-world feedback received. This **iterative process** ultimately enhances the likelihood of successful exploitation and ensures the long-term sustainability of project outcomes.

This way, the AI4TRUST project has designed a tailored approach to cater to the unique needs of these varied target customers groups, guaranteeing that the results are optimally advantageous for each segment:

- For **professionals involved in the media industry, fact-checking organisations, and journalism**, the AI4TRUST project focuses on providing tangible tools and solutions that **enhance their capabilities in combating mis/disinformation**. This includes the development and deployment of advanced AI technologies, such as debunking tools and

content verification methods, which can be seamlessly **integrated into their existing workflows**. Moreover, **APIs and plugins** will be provided by the AI4TRUST project to facilitate easy integration into partners' news production pipelines, allowing professionals to incorporate AI-driven solutions without significant adjustments to their established processes. The aim is to empower professionals to navigate the challenges posed by false information effectively, thereby elevating the overall quality and reliability of their work and offering a more dependable and fact-checked news service to a worldwide audience.

- For **researchers**, the project prioritises the advancement of **research, educational, and programs**. For instance, through the integration of AI4TRUST results into academic curricula, the project endeavours to furnish upcoming researchers with state-of-the-art knowledge and capabilities in the realm of false information detection. This methodology guarantees a steady stream of proficient researchers that contribute to the sustained endeavours in countering false information. Furthermore, the partnership with academic partners promotes **interdisciplinary research**, nurturing a more profound comprehension of the collaborative and evolving mechanisms involved in generating and disseminating problematic online content. Moreover, the academic partners will materialise the gained scientific knowledge and technological experience through **new scientific papers** in high-impact venues for the project. Finally, academic institutions play a pivotal role in disseminating knowledge and **training future professionals and researchers**. Simultaneously, the project aims to enhance the **technological transfer** potential for local companies. This ensures that the knowledge produced by the project is not only shared through **educational channels** but also implemented in **real-world contexts**.
- For **policymakers**, acknowledging the influential role played in shaping the **regulatory framework concerning mis/disinformation**, the project strives to offer invaluable **insights and recommendations**. This entails consolidating research discoveries and policy implications, rendering them easily understandable and relevant for policymakers. Through active involvement with policymakers, the project endeavours to aid in the formulation of well-informed and impactful policies aimed at tackling the issues presented by dis/mis/malinformation in the digital space.

Additionally, the **market analysis** and the **interaction with business experts** will help foster **concrete exploitation opportunities** by identifying specific business needs and gaps, and discussing how AI4TRUST technologies can deliver value to specific customer segments. The AI4TRUST project, in fact, represents an important opportunity for the consortium partners to explore commercial exploitation paths. The AI4TRUST exploitation strategy aims to **develop an MVP starting from the identified Key Exploitable Results** (KERs, see Section 4.2) that constitute the Platform itself.

For each identified KER, **specific business models** will be examined and tailored to facilitate the adoption of a commercial strategy. Engaging with both internal and external experts is crucial for

identifying business models for commercialisation. Iterative refinements and evaluations will ensure that the proposed KERs are applicable for the targeted markets, with a clear value proposition crafted for various customer segments, leveraging the key features developed throughout the project. The **viable business models** could be based on different revenue streams and will be further assessed and evaluated within the Sustainability strategy section to identify feasible monetization avenues.

4.1. Overview of AI4TRUST Platform

The project aims to distinguish itself in the market for digital disinformation through **several key factors**. Firstly, it focuses on **Technological Innovation**, committing to provide cutting-edge AI solutions for identifying and mitigating disinformation by filtering social noise and analysing multimodal content. The goal is to **develop state-of-the-art tools** that can quickly analyse large volumes of data and accurately identify false and manipulated content **using advanced algorithms and machine learning technologies**.

Additionally, the project emphasises **ease of use**, ensuring that the tools are intuitive and accessible to a wide range of potential target customers. A priority is placed on **reliability and accuracy**, with a commitment to high precision in detecting false information through sophisticated algorithms and credible verification sources. **Collaboration with experts in fact-checking and information sciences** is integral to maintaining the reliability of the solutions and adhering to the highest standards. Lastly, the AI4TRUST Project is dedicated to forming **strategic partnerships** to effectively combat false information and develop efficient tools.

To this end, **the goal of AI4TRUST** is to tackle mis/disinformation with **human validated content items** and **state-of-the-art tools** for monitoring, debunking and obtaining analytics. This is offered through a cross-country joint effort to build a platform that is trustworthy, ethical, respectful of data ownership and of user's privacy in adherence with EU values and guidelines. In particular, the AI4TRUST Platform aims to empower:

(i) media practitioners in debunking and monitoring of news items;
(ii) policymakers with analytics and insights;
(iii) researchers in data mining and development of AI models.

This goal is well coupled with **the AI4TRUST vision** of extending and improving the fact-checking activities, supporting the definition of rules and countermeasures, and bringing generative AI knowledge to the next frontier, with the final overall goal of **defending EU citizens from false information and manipulation**.

To this end, the AI4TRUST Platform has been designed as part of an **incremental development process**, with **three distinct versions planned**. Each version integrates components developed during the corresponding phase of the three planned R&D iterations (Figure 11).

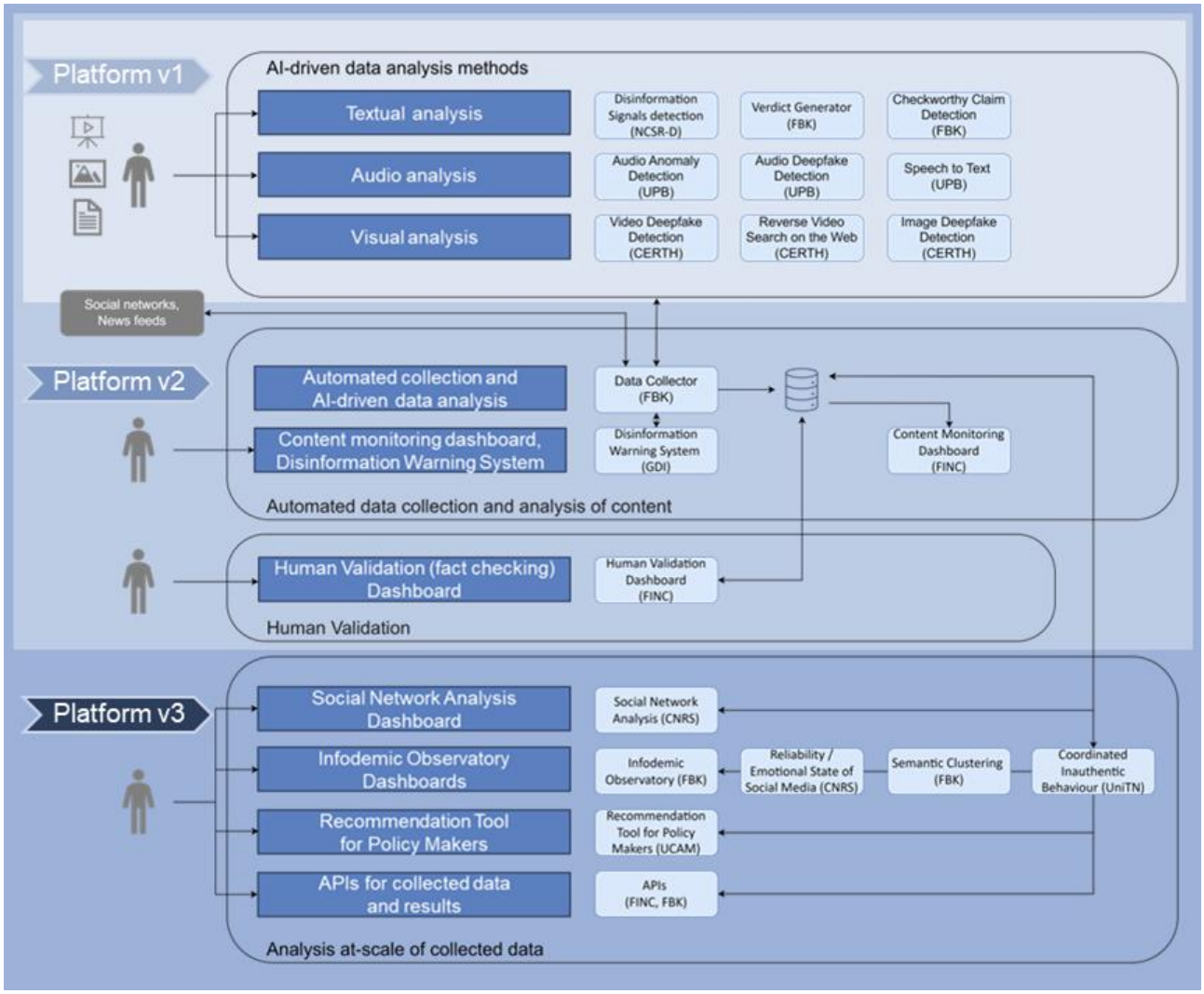

Figure 11: *AI4TRUST Roadmap*

In the **first version** (v1) (Mil. 2, Month 21, see **D5.5**[49]), the AI4TRUST Platform focuses on analysing individual content items using AI-driven data analysis methods developed during the first R&D iteration. The objective is to empower users to leverage these analysis methods for detecting disinformation. This includes analysing visual content to identify "deepfakes" or reused videos on the web, examining audio for deepfake speech or converting speech to text for transcription, and evaluating text for disinformation indicators such as hate speech, offensive or sensational language, or assessing whether the text warrants further verification. This set of tools can be directly utilised

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

and assessed by end-users, while also serving as foundational components for more advanced automated analysis services in subsequent iterations.

The **second version** (v2) of the AI4TRUST Platform (Mil. 4, M25, see **D5.6**[50]) is enriched with the automated collection and analysis of content from social media and news feeds, incorporating tools developed during the second R&D iteration. Content items related to three key topics (i.e., public health, climate change, and migrants) are automatically gathered and processed through a pipeline that integrates AI-driven analysis methods from the previous phase, along with new techniques such as detecting sensational content, identifying visual-text misalignment, and detecting video anomalies. This automated analysis is enhanced by the Disinformation Warning System (DWS), which flags potentially misinformative content. A monitoring dashboard provides users with access to the automatically collected content, alongside the results from the analysis pipeline. Additionally, a human validation dashboard enables fact-checkers to manually review the content, supported by insights from the automated tools, and validate the findings.

The **final version** (v3) of the AI4TRUST Platform (Mil. 6, M38, see **D5.7**[51]) will build on the data collected and processed in the previous stages, introducing advanced methods for large-scale analysis of the aggregated content. This iteration will incorporate updates to previously developed tools along with new tools created during the final R&D iteration. Notably, it will feature the "Social Networks Stack", which will provide users with aggregated contextual insights derived from historical data. Social network analysis will be utilised to examine the spread of disinformation, while advanced data analytics, such as source reliability assessments and infodemic trend monitoring, will be addressed within the infodemic observatory. These features will be complemented by policy-oriented services designed to monitor disinformation risks and support policymakers in crafting more effective strategies. The results will be integrated into dashboards, offering users graphs and visualizations that provide insights and trends across the overall dataset.

The following section provides an **overview of the functionalities offered by the AI4TRUST Platform, the Key Exploitable Assets** grouped into functional areas and finally **the identification of the KERs** in relation to the 3 initial versions (v1, v2, v3) aimed at the exploitation strategy of the AI4TRUST Platform.

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

4.2. Key Exploitable Results

The **AI4TRUST Platform** envisages **multiple assets** which can be grouped in the following functional areas (see Figure 6):

- **Analysis and Monitoring of Single News Items**, for the assets responsible for the textual, visual and speech analysis of single news items and the subsequent automatic ranking according to the manipulation risk evaluated by the Disinformation Warning System;
- **Data Acquisition and Normalisation**, for the assets responsible for the automatic acquisition of single news items from different sources and the collection of the respective human validations;
- **Collective Analysis of Social Media Actors and Items**, for the assets responsible for higher-level analysis across the multiple news items collected.

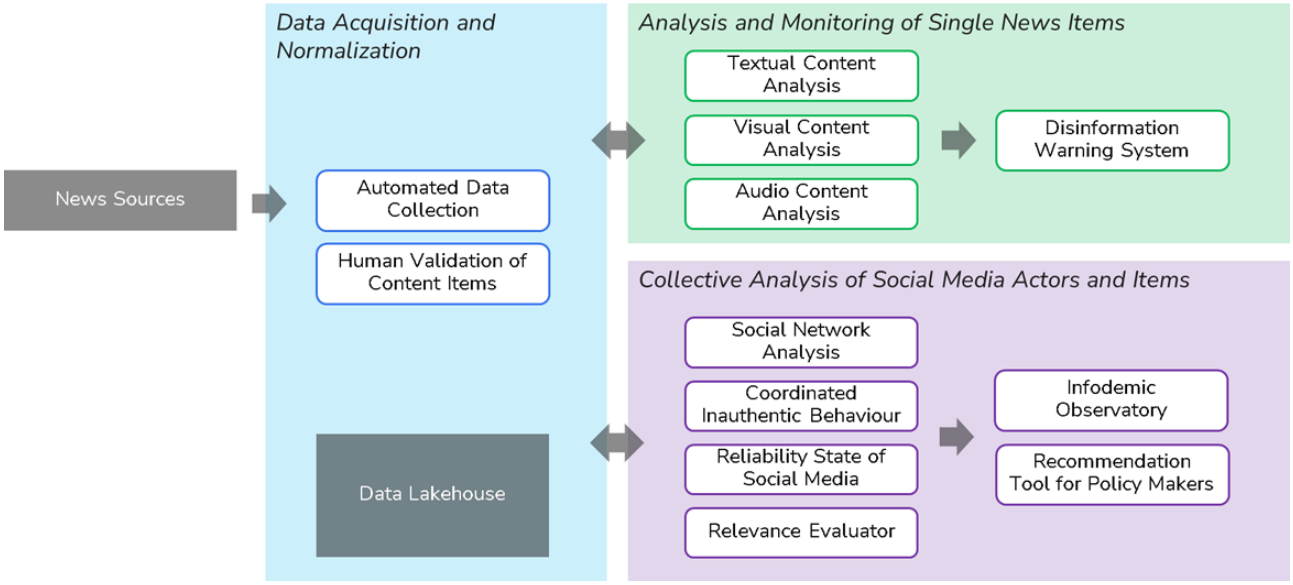

Figure 12: *Functional Architecture*

For what concerns the **Data Acquisition and Normalisation**, the main functionalities are:

- **Automated Data Collection**, which obtains single news items from the envisaged social media (i.e., YouTube and Telegram);
- **Human Validation of Content Items**, which enable fact-checkers to provide manual characterisation of analysed single news items.

Data acquisition assets are key to the whole Platform, as **all tools in the AI4TRUST Platform are data-driven**. As a matter of fact, analysis tools focusing on single media items and aggregated analysis tools all depend on the data that have been collected by data acquisition tools.

The automated data collection tool is configured in AI4TRUST to collect relevant data within **three well-defined topics (i.e., public health, climate change, migrants)**, as the amount of data in input to the Platform has to be limited due to computational and storage constraints. However, this filtering is parametric and can be tuned according to specific market needs, which makes the AI4TRUST Platform versatile to be **potentially used in the future for topics and in contexts different from those identified within the project**.

Automated data collection plays a crucial role for media professionals, policymakers and businesses which require an analysis over large amounts of data from diverse sources, formats, and languages. **Human validation** ensures the reliability and accuracy of these data sources, enabling journalists, newsrooms and other media users to sustain their verification processes effectively.

In relation to the **Analysis and Monitoring of Single News Items**, the main functionalities are:

- **Textual, Visual and Speech analysis**, which consist of a series of modules that enable the user to:
    - **detect disinformation signals**, segmenting the given input text with associated labels (i.e., “hate speech”, “clickbait”, “offensive”) and providing a confidence score;
    - **detect the check-worthiness of a claim**, labelling the given input text as “check-worthy” (i.e., factual and verifiable text that appears to be false, may be of public interest or of impact to the public, or may cause harm to the society, entities, groups, or individuals) and providing a confidence score;
    - **retrieve previously fact-checked claims**, searching the given input text amongst archived fact-checked claims and providing a similarity score;
    - **detect image deepfakes**, reporting the probability scores of different manipulation techniques for the given image;
    - **detect video deepfakes**, reporting the probability scores of different manipulation techniques for the given video;
    - **reverse search videos on the web**, extracting the most relevant frames from the given video, automatically performing image-based search on the web and returning a list of visually matching online sources and similar videos;
    - **detect audio deepfakes**, reporting the probability scores of manipulations for the given audio;
    - **detect audio anomalies**, providing a list of splicing points with the relative confidence scores;
    - **transcribe speech to text**, enabling textual analysis on the given audio content.
- **Disinformation Warning System**, which automatically labels single news items with a disinformation risk score, based on the results of the Textual, Visual and Speech analysis.

All these functionalities allow the **evaluation of single content items in a multimodal w**ay, acquiring automated indications regarding specific characteristics. This is especially useful for fact checkers and media professionals, who need to daily assess large quantities of content items in a context in which disinformation is growing in quantity and becoming more and more sophisticated, given the availability of AI-based tools that allow producing increasingly convincing disinformation content.

Monitoring features, based on the **Disinformation Warning System**, are aimed at assisting policymakers, media experts and consumers in overseeing online information dissemination, which involves understanding disinformation patterns, assessing their social impact, and identifying their sources. **Companies** can benefit from monitoring as well. By tracking data related to brands and products, businesses can deepen their understanding of social interactions, allowing them to promptly mitigate potential reputational-related threats.

Collectively, **textual, visual and speech analysis** can be implied by **all customer segments**, giving means to rapidly validating content accuracy.

As regards the **Collective Analysis of Social Media Actors and Items**, the main functionalities are:

- **Social Network analysis**, which evaluates how much hierarchical, inequal or biassed the social network is;
- **Coordinated Inauthentic Behaviour**, which allows to explore coordination dynamics around same/similar pieces of content across different digital spaces by checking simultaneously the status of specific pieces of contents within one Platform and the status of same/similar content for how they are present on different platforms;
- **Reliability State of Social Media**, which provides the end-user with synthetical quantitative indices of the disinformation risk presented in given areas of social media, to bring attention to areas of interest where the risk of disinformation flowing is higher;
- **Relevance Evaluator**, which establishes content relevance on the basis of three criteria: language, author's virality and post's virality;
- **Infodemic observatory**, a map-based tool that provides disinformation statistics and indices, such as the number of unreliable news circulating for a given country, the social media volume or the risk index in a given country, regarding a certain topic, and in a given timeframe;
- **Recommendation tool for policymakers**, which draws on different functionalities of the AI4TRUST Platform, namely the detection and analysis of disinformation signals (e.g., claim validity and Social Network Analysis), to link aggregated inputs from the AI4TRUST Platform with a classification of their severity level and a guide towards mitigation measures. The hybrid character is conferred by both the human-centred design of the tool and the human supervision of the recommendation inputs, particularly when dealing with disinformation classified as a systemic risk.

The evaluation and **analysis of items and social media actors** are directly related to the needs of political decision makers and companies. Through these processes, coupled with monitoring, users can acquire analytical capabilities, access source reliability, and anticipate potential threats.

For what concerns the identification of **Key Exploitable Results (KERs)**, starting with the **first version** (v1), the Platform focused on analysing individual content items using AI-driven methods to detect false information, laying the groundwork for the Toolbox. In the **second version** (v2), the emphasis shifted to automated data collection and human validation, leading to the establishment of the Monitoring and Human Validation functional area. The **final version** (v3) will introduce advanced analytics capabilities, culminating in the Analytics (collective analysis). Each version added core functionalities that enhanced the Platform's overall effectiveness, while the identification of the functional areas served to group the various assets associated with each KER. The progression of the AI4TRUST Platform from its initial versions (v1, v2, v3) to the identification of the three KERs illustrates **a logical and strategic development process**.

Table 3 identified KERs that are strictly linked to the 3 distinct versions initially planned and already detailed above. **For each KER the relative Key Exploitable Assets have been identified** (grouped through functional areas). Although 3 KERs have been identified, it is worth mentioning the essential precondition to take into consideration regarding the functionality and interdependencies of the KERs.

KER#1, referred to as the Toolbox, operates independently, providing essential tools for the analysis of single news items through various AI-driven methods. In contrast, **KER#2** (Monitoring and Human Validation) requires automatically collected data that undergoes processing through some textual, visual, and audio analysis assets from KER#1 before being utilised by the Disinformation Warning System. This action enhances the human validation process by allowing fact-checkers to review and provide feedback on content that has been flagged for potential false information. Similarly, **KER#3** (Analytics) could rely on the outputs from KER#1 assets, as it utilises the automated data collection and analysis to generate insights, trends, and analytics regarding the spread of false information. However, KER#3 has the added flexibility to obtain such pre-processed data from other resources beyond KER#1. This means that KER#3 can analyse pre-processed data collected from various external sources without necessarily having to depend on KER#1 assets.

#	KER(s)	Description
1	**Toolbox (analysis of single news items)**	The Toolbox encompasses advanced analytical tools for the analysis and monitoring of individual news items, utilising AI-driven methods to detect false information across textual, visual, and audio content.
2	**Monitoring and Human Validation**	Monitoring and Human Validation focuses on the automated collection of data and the human validation of content, ensuring the reliability and accuracy of information through a combination of automated tools and manual review processes.
3	**Analytics (collective analysis)**	Analytics provides comprehensive insights into the dynamics of false information by analysing aggregated data from social media and other sources, enabling users to identify trends, assess disinformation risks, and inform strategic decision-making.

Table 3: *Key Exploitable Results (KERs) mapping*

In the following three tables (Tables 4, 5, and 6), the mapped **Key Exploitable Assets** have been **associated with each of the identified Key Exploitable Results**. Each asset is described in detail with its "Individual Innovation, Exploitation & Sustainability Plan" in the following sections (from sec. 5.3 to sec. 5.19), according to the involved partner.

Functionality	Key Exploitable Asset	Partner	Type of Asset	Individual Innovation, Exploitation & Sustainability Plan
Textual Content Analysis	AI model for check worthiness of textual claims	FBK	Foreground	See sec. 10.1
Textual Content Analysis	AI model for retrieval of previously fact-checked claims	FBK	Foreground	See sec. 10.1
Textual Content Analysis	Document intelligence – Technology	NCRS-D	Foreground	See sec. 10.4

Functionality	Key Exploitable Asset	Partner	Type of Asset	Individual Innovation, Exploitation & Sustainability Plan
Textual/Visual Content Analysis (backend only)	AI model for visual-text misalignment detection	CERTH	Foreground	See sec. 10.2
Visual Content Analysis	Tool for reverse video search on the Web	CERTH	Background	See sec. 10.2
Visual/Speech Content Analysis	Deepfake image/video detection	CERTH	Background	See sec. 10.2
Visual Content Analysis (backend only)	AI models for sensational content detection	CERTH	Foreground	See sec. 10.2
Visual Content Analysis (backend only)	Tool for video anomaly detection	UNITN	Foreground	See sec. 10.3
Speech Content Analysis	Speech-to-text technology and web service	POLITEHNICA (formerly UPB)	Foreground	See sec. 10.6
Speech Content Analysis	Audio deepfake detection technology and web service	POLITEHNICA (formerly UPB)	Foreground	See sec. 10.6
Speech Content Analysis	Audio anomaly detection technology and web service	POLITEHNICA (formerly UPB)	Foreground	See sec. 10.6

Table 4: *Key Exploitable Assets mapping for KER#1*

Functionality	Key Exploitable Asset	Partner	Type of Asset	Individual Innovation, Exploitation & Sustainability Plan
Automated Data Collection	Social listening data stream	FBK	Foreground	See sec. 10.1
Disinformation warning system	Disinformation warning system (DWS)	GDI	Foreground	See sec. 10.8
Human Validation of Content Items	Human Validation Tool	FINCONS	Foreground	See sec. 10.17

Table 5: *Key Exploitable Assets mapping for KER#2*

Functionality	Key Exploitable Asset	Partner	Type of Asset	Individual Innovation, Exploitation & Sustainability Plan
Social Network Analysis	Social Network Analysis Tool	CNRS	Foreground	See sec. 10.5
Coordinated Inauthentic Behaviour	Coordinated Inauthentic Behaviour Tool	UNITN	Foreground	See sec. 10.5
Reliability State of Social Media	Reliability State of Social Media Tool	CNRS	Foreground	See sec. 10.5
Relevance Evaluator	Relevance Evaluator	CNRS	Foreground	See sec. 10.5
Infodemic Observatory	Infodemic Observatory Tool	FBK	Foreground	See sec. 10.5
Recommendation Tool for Policymakers	Hybrid Recommendation Tool	UCAM	Foreground	See sec. 10.16

Table 6: *Key Exploitable Assets mapping for KER#3*

The **workshop** held with AI4TRUST stakeholders on January 9, 2025 (see chapter 3 "Methodology") highlighted **how KERs and the related assets could meet the specific needs of different potential target customers**.

For **fact-checkers**, the Platform offers a comprehensive monitoring dashboard (based on DWS), which facilitates the identification of news items across various languages that are worthy to be

analysed. Additionally, it enables the creation of a database containing verified AI-tampered content, enhancing the efficiency of their verification processes. **Journalists** benefit from access to a database of human-validated information, which is integrated into the monitoring dashboard, allowing them to quickly verify content. **Policymakers** are equipped with tools that aggregate information and data collected from the AI4TRUST Platform, enabling them to analyse trends related to misinformation, malinformation, and disinformation effectively. **Researchers**, including mathematicians and sociologists, gain access to a Platform API that allows them to retrieve fact-checked data from existing archives. They could also obtain regulation-compliant data for training and fine-tuning AI methods, as well as indices of verified and manipulated content. The Platform streamlines the filtering, collection, and standardization of check-worthy multilingual data from multiple channels, providing indexed fact-checked and labelled data. Furthermore, sociologists can leverage the Platform to generate informative visuals and analytics regarding the spread of false information on the web, as well as access check-worthy social media data. Overall, these KERs reflect the Platform's commitment to **enhancing the capabilities of its diverse target customer base** in the fight against false information.

4.3. Individual Innovation and Exploitation per Asset

This section explores the specific **approaches and methodologies** adopted by each **partner** to maximise the potential of their individual assets. By strategically leveraging their **unique strengths and expertise**, the partners aim to create **tailored plans** that not only foster **innovation** but also ensure the **sustainable exploitation routes** of the **AI4TRUST Key Exploitable Assets**. This strategic alignment underscores the importance of a **cohesive approach** to both **innovation and exploitation**, facilitating collaboration and synergy among consortium members. Through this approach, the partners work collectively to achieve the **optimal valorisation** of the project results, ensuring long-term success and impact.

Annex II (sec. 10) provides in detail the individual exploitation plans for each partner involved in the AI4TRUST project, highlighting how they intend to utilise their unique assets and expertise to maximise the project's potential value. These individual plans are integral to the overall strategy, ensuring that each partner's contributions are aligned with the project's goals and objectives. Each partner of the AI4TRUST project has a different role in the valorisation of the project results, based on the asset owned and the mission. **Three groups** have been defined according to the role:

- **Scientific partners** ensure knowledge generation and capacity of providing technologies beyond the state-of-the-art.
- **Business partners** develop and commercially exploit technological solutions based on project results.

- **End-user partners** benefit from the adoption of project results in their daily operations and duties.

The **assignment of each partner to these categories** is reported in the table below (Table 7). As shown, different partners have multiple roles within the project, highlighting the opportunity to **exploit the AI4TRUST Platform from different perspectives**, not only the business one.

Scientific partners	Business partners	End-user partners
• FBK • CERTH • UNITN • NCSR - D • CNRS • POLITEHNICA (formerly UPB) • UCAM	• FIN • GDI • MALDITA • SAHER	• GDI • DEMAGOG • MALDITA • ELLINIKA • EURACTIV • SKYTG24 • ADB • EMS

Table 7: *Partners' roles*

4.4. Connection between project Expected Results and Results achieved in AI4TRUST

This section examines the critical connection between the **expected results** outlined in the **Description of Action (DoA)** of the **AI4TRUST** project and the results that have been achieved throughout its implementation. The following table (**Table 8**) provides a detailed comparison of the **specific expected results** alongside the corresponding **results achieved** in **AI4TRUST**, including their **achievement status**. This comparison highlights the progress made and underscores the alignment between the project's goals and the outcomes attained during its course.

#	Expected Result	Results achieved in AI4TRUST	Status
R1.1	A digital platform for gathering and processing social media and news aggregators data.	KER #2: Automated Data Collection, Disinformation Warning System, Monitoring Dashboard	Achieved
R1.2	Novel datasets (> 1Tbyte) of news and social media texts and interactions, images and audio-visual contents.	• Data collection in 8 languages for news media and YouTube • Data collection for Telegram in 8 languages	In progress (YouTube & News: active, Telegram: in progress)

#	Expected Result	Results achieved in AI4TRUST	Status
R2.1	Textual analysis: A large-scale dataset for training text-based methods and a multi-faceted arsenal of text-based methods (in > 8 languages) for combating disinformation.	• A large-scale dataset for training text-based verdict generation in 8 languages, 200 examples per language. Each example is composed by CLAIM+VERDICT+ARTICLE. Verdicts were written by AI4TRUST fact-checkers • KER #1: AI model for check worthiness of textual claims (EN, IT, ES), Document intelligence – Technology (EN, EL, RO, IT, ES, PL, DE, FR)	Achieved
R2.2	Audio analysis: An advanced multilingual (in > 5 languages) speech-to-text technology; Methods for generating and detecting deep fakes in audio content (15% reduced Equal Error Rate (EER) scores than State of the Art).	KER #1: • Speech-to-text technology (based on Transformer models) and web service (EN, RO, IT, ES, PL, DE, FR) • Two methods for audio deepfake detection and a web service. The first method was evaluated on a benchmark of 4 scientific datasets[52] with an average EER of 4.9%, which is 51% lower than the state of the art. The second method (Interspeech 2025 submission) achieves a 55% relative reduction in EER on the In-the-Wild dataset and a 63% reduction on our newly proposed real-world deepfakes dataset (AI4T) • Several methods for enabling fast speaker adaptation for state-of-the-art speech deepfakes generators (e.g. VITS, GradTTS, xTTS, Parler-TTS and Matcha-TTS), targeting specific voices. • A technology for audio anomaly detection (identifying splicing points and fake segments) and web service	In progress

[52] Interspeech 2024 paper: https://www.isca-archive.org/interspeech_2024/pascu24_interspeech.pdf

#	Expected Result	Results achieved in AI4TRUST	Status
R2.3	Visual analysis: Methods for detecting video re-use on the Web and in closed collections; Methods for detecting deep fakes in images/videos (15% higher ROC-AUC / Accuracy scores than State of the Art).	KER #1: • A method for video keyframe selection and keyframe-based reverse video search on the Web based on the GoogleLens technology • A method for deepfake image detection based on a pipeline used to identify the optimal augmentation strategy (greedy & genetic algorithm-based search) • A service for deepfake video detection, integrating a SoA-performing model identified through a state-of-the-art comparative analysis	In progress
R2.4	Multimodal analysis: A set of indicators (>3) for evaluating disinformation; a method for detecting sensational content (>80% acc.); tools for detecting anomalies/misalignments between sources from different modalities.	KER #1: • Two multimodal deepfake video detection methods; one comparing speech-to-text and lip-reading data and another one focusing on silent parts • Two methods for sensational content detection; one that shows >90% accuracy on the detection of visually disturbing content, and another one detecting various types of sensational actions/events using language-based descriptions of them • A method for assessing the misalignment between the visual content and an associated text description based on learned multimodal embeddings • A method for detecting abnormal events in videos based using VLM for captioning and visual-text alignment and LLMs for reasoning on the captions	Achieved

#	Expected Result	Results achieved in AI4TRUST	Status
R2.5	A disinformation warning system integrating individual AI tools and providing reports on websites and single pieces of content, labelling them as verified or fake/manipulated.	KER #2: The Disinformation Warning System (DWS) that combines the outputs of individual AI tools for disinformation signal detection, check-worthy claim detection, sensational content detection, video anomaly detection, visual-text misalignment detection, and domain disinformation detection to flag content with a high risk of disinformation. The DWS has been already integrated in the Monitoring Dashboard of the AI4TRUST platform	Achieved
R3.1	A toolkit for designing AI information verification systems for the intended user community that addresses the socio-cultural and linguistic factors at play in the generation and dissemination of misinformation and disinformation.	KER #1: conception, design, development, and deployment of an "SNA stack" for disinformation detection and mapping at scale, focusing on social media platforms. The SNA stack comprises the following tools: relevance evaluator, reliability state of social media, emotional state of social media, priority evaluator, social network visualisation, coordinated behaviour, and infodemic and recommendation tools KER #2: Provide on-the-fly contextual analysis of the dissemination of disinformation through the identification of coordinated (link-sharing) behaviour on social media; the visualisation of networks according to relevance and reliability scores; the aggregation of statistical outputs and posterior translation into standard recommendations for end users	In Progress

#	Expected Result	Results achieved in AI4TRUST	Status
R4.1	The AI4TRUST Platform.	• KER #1: Toolbox for the analysis of single news items • KER #2: Monitoring dashboard for automatically collected and analysed news items; Human Validation dashboard for fact-checking news items	Partially achieved
R5.1	A comprehensive methodology for validating AI4TRUST solutions.	• AI4TRUST platform quality model based on the ISO/IEC 25010 model • Platform-specific KPIs for quality assessment • Specialised evaluation workshop with evaluation scenarios targeting particular modules	Achieved
R5.2	A thorough evaluation and assessment for the first (>50 users) and second (>10.000 users) pilot session.	49 media professionals in total participated in the first piloting session. The number is expected to be increased in view of the next piloting phases.	First piloting session: almost achieved Second and third piloting sessions: not yet performed

Table 8: *Expected and achieved results in AI4TRUST*

5. Sustainability Strategy

This section outlines the strategies for **sustainability** and **engagement** adopted by the **AI4TRUST** project, alongside a preliminary analysis of the potential **business models**. The **sustainability plan** for the **AI4TRUST** project adopts a long-term perspective, analysing both potential threats and opportunities, assessing associated costs, and proposing integrated solutions to ensure a **lasting impact**. This plan is further reinforced by a comprehensive **engagement strategy** that prioritises continuous **knowledge dissemination**, **partnership building**, and **collaboration** across multiple projects. By fostering strong connections with **stakeholders** and promoting shared initiatives, the project ensures that its outcomes remain **relevant**, **scalable**, and **impactful** beyond its initial scope, contributing to sustained success in combating disinformation.

5.1. Sustainability Plan Outline

The development of a **Sustainability Plan** for the AI4TRUST solution is essential to ensuring its **long-term viability** and **effectiveness** in combating mis/disinformation beyond the duration of the project. The primary objective of this plan is to identify key **operational**, **financial**, and **market strategies** that will facilitate ongoing development, user engagement, and resource allocation. By addressing both potential risks and opportunities, the Sustainability Plan seeks to establish a robust framework that maximises the **AI4TRUST Platform's impact** and ensures its continued relevance within the evolving information landscape. The plan will encompass the following **key activities**:

- **Engagement Strategy**: This strategy aims to actively engage stakeholders, including **end-users**, **policymakers**, and **industry partners**, to raise awareness of the AI4TRUST solution, its innovative nature, its potential, and its functionalities. The goal is to foster **collaboration**, enhance **visibility**, and ensure that the research findings are effectively communicated and utilised, ultimately maximising the project's impact and **sustainability**.
- **Business and Monetisation Model**: This section will provide a strategic framework outlining how the Platform will generate **revenue** while delivering value to its potential customers. The model is designed to ensure long-term **viability** and impact in combating mis/disinformation. It will include the following elements:
    - **Business Model Canvas**: This will outline the key components that define the value proposition, customer segments, revenue streams, and operational structure. The Business Model Canvas serves as a strategic tool to visualise and refine the business approach, ensuring the long-term impact of the AI4TRUST solution.
        - **Value Proposition**: This component will articulate the unique benefits and advantages that the AI4TRUST solution offers to its target users and stakeholders. It will address how the solution meets specific needs, enhances existing solutions, and contributes to **societal**, **economic**, and **technological** advancements. By clearly defining the value propositions, the AI4TRUST solution will effectively communicate its significance, fostering engagement and ensuring that the results are utilised and sustained beyond the project's duration.
    - **Operations Plan**: This section will provide an overview of the administrative and operational planning necessary for the commercial exploitation of the AI4TRUST solution.

- **Cost Categories**: A comprehensive breakdown of the costs associated with the AI4TRUST solution will be conducted, categorising them into:
    - **Technological Minimum Viable Costs**: This includes costs related to basic maintenance and operational requirements of the Platform, such as server hosting, data storage, maintenance, and technical support, along with resources needed for continuous data collection and analysis.
    - **Development and Enhancement Costs**: These include the costs associated with the initial development of the Platform, as well as ongoing upgrades, feature additions, and improvements (e.g., software development, UI/UX design, algorithm enhancement, performance optimisation). This ensures that the solution remains **effective** and **competitive** beyond the project phase.
- This categorisation will facilitate **effective budgeting**, resource allocation, and financial planning, ensuring the long-term sustainability and viability of the project.
- **Personnel**: A breakdown of the **personnel** costs and types of expertise required for the development, operation, and maintenance of the Platform.
- **Key Partners**: This will identify the essential partnerships and external resources necessary to enhance the functionality and effectiveness of the Platform (e.g., assessment services, external data sets, IT infrastructure, legal advice).

- **Revenue Streams**: This section will identify potential sources of revenue for the AI4TRUST Platform (e.g., **Subscription Fees**, **Custom Assessment Services**, **Data Licensing**, **Freemium Model**, **Partnerships and Collaborations**). Identifying revenue streams is crucial for establishing financial sustainability and ensuring the Platform's long-term viability. This will help attract investors and allocate resources efficiently, enabling the Platform to continue its mission of combating mis/disinformation.
- **Pricing Policy**: A potential pricing structure will be estimated to ensure **profitability**. This will guide customers in understanding the value of different subscription plans and services, thus helping the Platform position itself effectively in the market.

- **Risk Analysis**: A comprehensive identification and assessment of potential risks related to the commercial exploitation of the AI4TRUST solution, particularly in regard to post-project sustainability, will be conducted. Risks may include general **economic conditions**, changes in **regulatory environments**, and industry-specific challenges, such as **security** and **privacy**

concerns. Specific product-related risks, such as shifts in public perception regarding mis/disinformation or the emergence of new technologies that could either enhance or challenge the Platform's effectiveness, will also be evaluated. Mitigation strategies for each identified risk will be proposed to enhance the Platform's **resilience**.

Furthermore, the plan will explore future integration challenges that may arise post-project, including:

- **Data Volume**: The volume of data to be processed, which will depend on the selected sources, topics, and filtering criteria.
- **Computational Power**: The computational requirements linked to the amount of data and types of analyses, which will need to be optimised according to specific needs.
- **Integration with Other Data Sources**: The need for integration with additional data sources to extend the Platform's monitoring capabilities, dependent on **API availability** and changing **legal requirements**.
- **System Modularity**: The ability to deploy customised versions of the Platform or utilise specific components as needed.
- **Remuneration of Involved Parties**: Careful consideration will be given to remuneration agreements, ensuring a balance that addresses stakeholders' needs while allowing for sufficient **flexibility** in exploitation.
- **Interoperability with Third-Party Applications**: Analysis of technical aspects, such as the use of common data models, authentication and authorisation processes, and monetisation approaches, will be conducted to ensure **compatibility**.

These integration challenges will be carefully analysed, with potential mitigations and improvements being proposed and, where feasible, implemented during the AI4TRUST project, thereby increasing the Platform's **impact** and paving the way for future development.

5.2. Engagement Strategy

The long-term engagement of AI4TRUST implies a constant and targeted sharing of dissemination knowledge and sensibilisation through the **networks and communication channels** defined within the project, together with the support from related projects such as the **Horizon Europe project AICODE**[53] (featuring FBK, CERTH, and EURACTIV as partners), the **CERV project HATEDEMICS**[54] (led by FBK with involvement from MALDITA, DMGG, and SAHER), and series of

[53] AI-CODE

[54] Hatedemics

forthcoming project proposals set for submission under the **Horizon Europe Clusters 2**[55]**, 3**[56]**, and 4**[57] funding calls. **Partners can contribute effectively to the overall engagement by:**

- **Sharing knowledge and insights** from AI4TRUST with its existing company network, which includes end-users, industry professionals, and academic institutions from Europe and beyond. This activity will be done proactively by forwarding constantly the information on the outputs and deliverables as they evolve over the duration of the project.
- **Leveraging the advantages of engaging with the above networks** to establish new partnerships and collaborations with companies that share a mutual interest and expertise in the AI4TRUST domain, fostering opportunities for future research and operational benefits.

The objective is to create new opportunities for the future project's active involvement in **various research and innovation projects**, as well as **other forms of collaboration** with research and development institutions and industrial partners in Europe and globally. This goal is broken down into three specific short-term objectives:

- to build a **network of potential academic and research partners** (European universities and research centres). The potential academic partners will, in the first phase, be selected according to the strongest performing fields of research within each institute.
- to establish a **network of potential industrial partners.**
- to create a **network of potential professional partners and policymakers** (including end-users)

With this perspective, the project provides a unique opportunity to **engage with other European and international organisations** and discuss their practices aimed at reducing the presence of false information in the digital space.

In addition, it is emphasised that the project's potential for growth in engagement, dissemination and awareness raising is amplified by its participation in the **AI against Disinformation Cluster**, a group of **sister projects under Cluster 4** aimed at growing the networking and expansion possibilities of the European mission against disinformation. From February 2025 to February 2026, AI4TRUST will serve as the secretariat of that Cluster. For instance, projects like **TITAN**[58] and **Vera.ai**[59] can provide valuable insights into the innovative approaches and the unique value of the AI4TRUST outcomes. In this context, it is important to highlight the **emerging synergy between projects** addressing similar issues, which we hope will be expanded to create a solid network of collaborators, facilitating knowledge-sharing, dissemination activities and results among

[55] Cluster 2: Culture, Creativity and Inclusive society - European Commission
[56] Cluster 3: Civil security for society - European Commission
[57] Cluster 4: Digital, Industry and Space - European Commission
[58] TITAN Project
[59] Vera.ai

interested entities. This is the central objective of **"Meet the Future of AI"**, a Horizon Europe Policy and Innovation initiative organised by **"sister projects"** such as **AI4TRUST**[60], **AI4Media**[61], **AI4Debunk**[62], **AI-CODE**[63], and the previously mentioned **TITAN** and **Vera.ai** projects. This initiative has already resulted in a dedicated **event "Meet the Future of AI: Countering Sophisticated & Advanced Disinformation"**[64] held in Brussels in June 2023 and the publication of a **White Paper titled "Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities"**[65] in February 2024.

In addition, as part of AI4TRUST's engagement efforts, a **Community of Practice** will be established by M38. This community will build upon the project's accumulation of knowledge and practical experiences to maximise its anticipated impacts. **FINCONS**, as the task leader, with the support of **EURACTIV** as WP7 Leader, will spearhead this initiative by identifying stakeholders who are interested in AI4TRUST's topics and outputs, both during and beyond the project duration. By leveraging the **AI4TRUST Europe-wide network**, as well as the extensive networks of the **Consortium partners**, FINCONS will facilitate the engagement of **media professionals, fact-checkers, researchers, policymakers, IT companies, NGOs/CSOs**, and their respective networks across multiple countries. The aim will be to collect **feedback** on AI4TRUST based on these stakeholders' experiences and to gain insights into their potential interests in the Platform and/or its components.

5.3. Business and Monetisation Model

As the **AI4TRUST** solution aims to combat **false information** and enhance **information integrity**, identifying **sustainable revenue streams** is crucial for its long-term viability and impact. The following potential revenue streams should be considered, each of which is aligned with the Platform's core mission and the needs of its target customers. The table below (**Table 9**) presents the potential revenue streams that could be evaluated for the **AI4TRUST** solution.

[60] https://ai4trust.eu/
[61] https://www.ai4media.eu/
[62] https://ai4debunk.eu/
[63] https://aicode-project.eu/
[64] Meet the Future of AI programme
[65] Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities

Revenue Stream	Description	Target Customers
Subscription fees	Annual/monthly access fees to access premium features and tools tailored to user needs.	Media professionals, researchers, and policymakers.
Consulting services	Customized insights and strategies for Companies based on data interpretation.	Companies needing data insights.
Partnerships	Collaborations with media companies and governmental bodies for joint initiatives.	Media companies, governmental organisations.
Pay-per-use	Access to specific AI-driven tools on a pay-per-use basis, allowing flexibility for users.	Media professionals, researchers.
License Fees	Recurring fees for organisations integrating AI4Trust technology into their systems.	Media companies, governmental institutions.
Commissions on transactions	Commissions on transactions facilitated through the platform, such as verified content sales.	Organisations needing validation services.

Table 9: *Potential revenue streams*

Subscription fees can be structured for different user tiers, allowing media professionals, researchers, and policymakers to access advanced tools and analytics tailored to their specific needs. This model ensures a steady income while providing target customers with valuable resources to combat mis/disinformation effectively.

Consulting services can be offered to help Companies interpret data and develop strategies based on insights generated by the Platform, creating additional value and fostering long-term relationships with target customers.

Partnerships with media companies and governmental organisations can lead to joint initiatives and funding opportunities, enhancing the Platform's reach and impact. These revenue streams are strategically aligned with the core mission of AI4TRUST to enhance information integrity and combat mis/disinformation, ensuring the sustainability and growth of the Platform in a rapidly evolving digital landscape.

Pay-per-use could be implemented by allowing potential target customers, such as media professionals and researchers, to access specific features of the AI4TRUST platform on a pay-per-use basis. For instance, users could pay for each instance of automated fact-checking or for

accessing advanced analytics tools. This approach provides flexibility for customers who may not require continuous access but need the tools for specific projects or tasks. Examples of services that use a pay-per-use strategy are cloud providers, such as Amazon AWS or Microsoft Azure platforms.

License fees could be adopted. A licensing model where organisations, such as media companies or governmental institutions, pay a recurring fee to access the Platform and its functionalities. This could include tiered licensing options based on the number of users, or the extent of features utilised, similar to how entertainment platforms like Netflix or Disney Plus operate. This model ensures a steady revenue stream while providing users with comprehensive access to the Platform's capabilities.

Commissions on transactions could be implemented. A commission-based model for specific transactions facilitated through its services. For example, AI4Trust could provide a marketplace for verified content or connect fact-checkers with organisations needing validation services; a commission could be charged on each transaction. This model is akin to concert ticket booking platforms that apply a pre-sales commission to ticket prices, allowing AI4Trust to benefit from the transactions occurring within its ecosystem.

In sum, the **AI4TRUST Platform** has been meticulously designed to address the pressing challenges posed by **false information** through a comprehensive suite of tools and functionalities. To effectively leverage the identified **Key Exploitable Results (KERs)**, we have developed three viable **Business Model Canvases**, each corresponding to one of the Platform's primary functional areas (Figures 13, 14, and 15): **Toolbox**, **Monitoring and Human Validation**, and **Analytics**. Each canvas outlines a distinct **value proposition** tailored to meet the specific needs of our target customers, including **media professionals**, **policymakers**, **researchers**, and **companies**.

The **Toolbox** focuses on providing advanced analytical tools that enable media professionals to efficiently verify the authenticity of news items. The **Monitoring and Human Validation** area underscores the importance of both automated data collection and human oversight, ensuring the reliability of information sources for media professionals and researchers alike. Lastly, the **Analytics** functional area offers comprehensive insights into **social media dynamics**, equipping policymakers and companies with the tools to understand and respond effectively to the evolving landscape of **mis/disinformation**.

Such business model strategies are widely used in contemporary products and services. By aligning the business models with the KERs identified throughout the project, **sustainable revenue streams** can be created, delivering significant value to the identified target customers. This strategic approach not only enhances the effectiveness of the **AI4TRUST Platform** but also fosters a

collaborative ecosystem that empowers all stakeholders in the ongoing fight against false information.

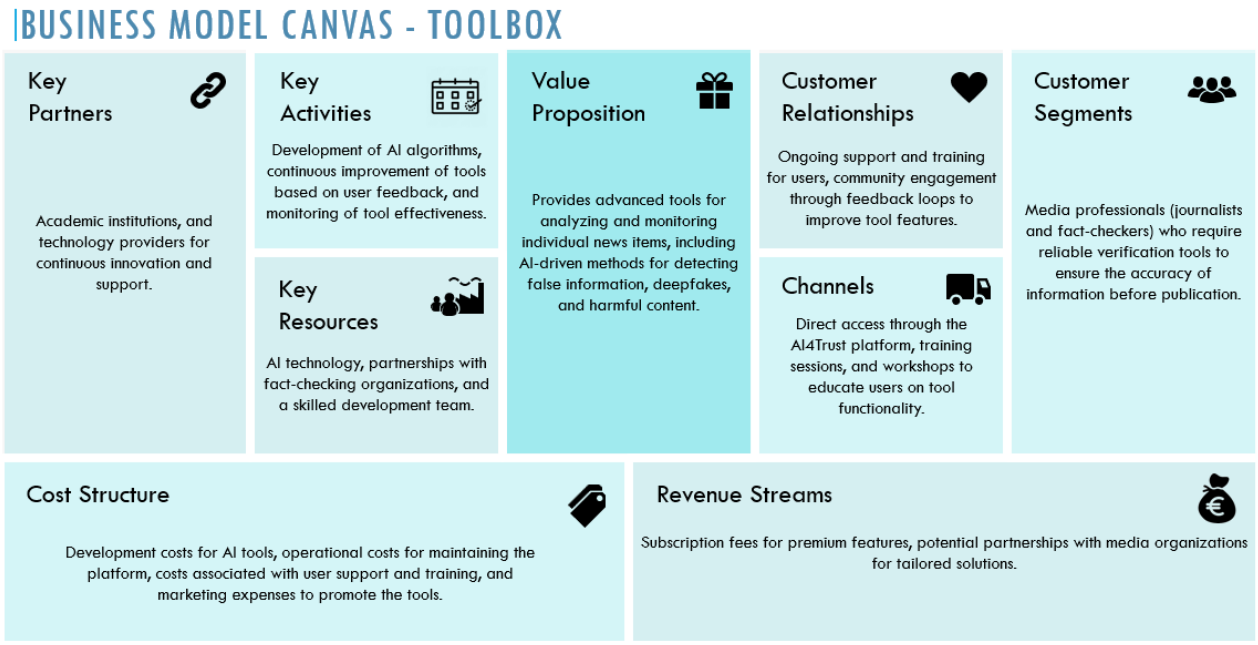

Figure 13: *Business Model Canvas - Toolbox*

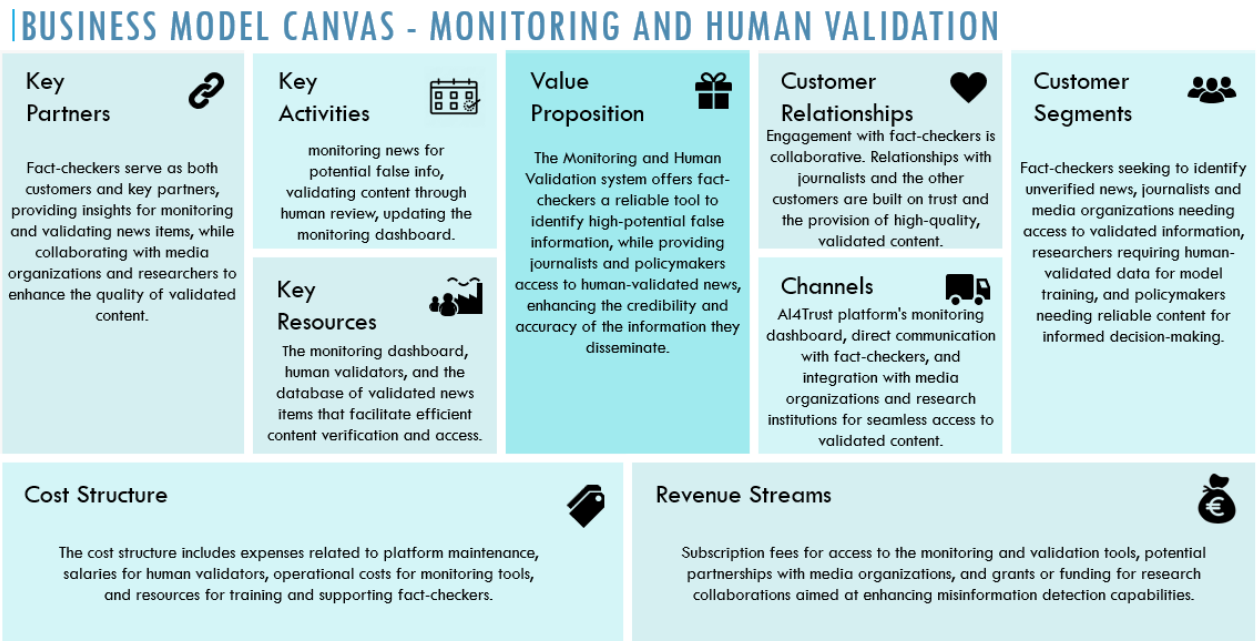

Figure 14: *Business Model Canvas - Monitoring and Human Validation*

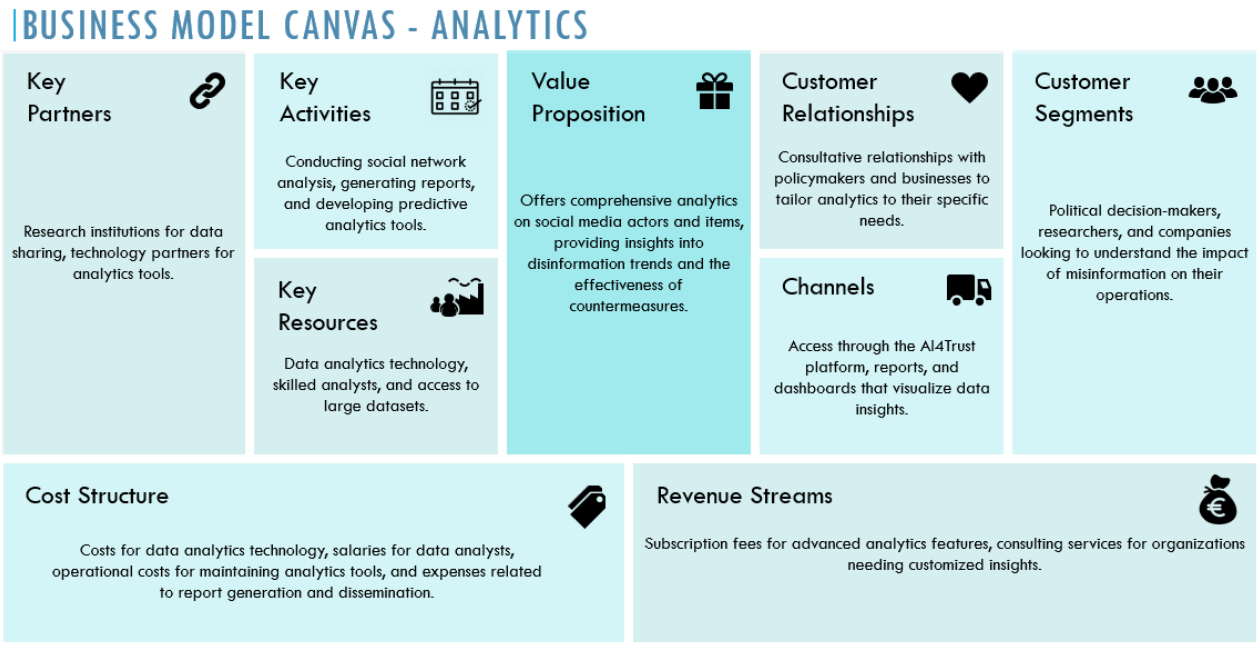

Figure 15: *Business Model Canvas - Analytics*

Value Proposition

The **value proposition** of the **AI4TRUST** project centers on delivering advanced solutions to combat **disinformation** and foster a reliable **information ecosystem** (Figure 16). By leveraging **Artificial Intelligence** algorithms, the Platform can swiftly identify and flag potentially false information, enabling prompt user responses. Its **nearly real-time detection** and verification capabilities are crucial for addressing the rapid spread of online false information. Furthermore, the use of **Machine Learning** techniques facilitates the prediction of false information trends, empowering digital platforms to implement preventive measures before false information becomes widespread.

The technology's flexibility and scalability allow organisations to customise the solution according to their specific needs, with the potential for expansion over time. Overall, **AI4TRUST** provides a comprehensive approach that combines **nearly real-time detection**, **predictive analysis**, and **scalability** to effectively tackle the challenges posed by false information in the digital age. This comprehensive approach not only enhances the **reliability of information** but also fosters a more **informed public**, setting the stage for the detailed value propositions of its **Key Exploitable Results (KERs)**: **Toolbox**, **Monitoring and Human Validation**, and **Analytics**.

KER#1) Toolbox Value Proposition: The value proposition of the Toolbox lies in its provision of advanced analytical tools designed to empower media professionals, particularly journalists and fact-checkers, in their quest for accurate information. By leveraging state-of-the-art AI-driven methods, the Toolbox enables users to detect false information across various media formats,

including text, images, and audio. This functionality is crucial in a fast-paced news environment where the rapid spread of false information can undermine public trust. The Toolbox not only enhances the efficiency of the verification process but also improves the quality of journalism by ensuring that only credible information is disseminated. Additionally, features such as deepfake detection and content analysis provide journalists with the necessary resources to uphold their ethical responsibilities, ultimately contributing to a more informed public.

***KER#2) Monitoring and Human Validation Value Proposition**:* The Monitoring and Human Validation offers a robust framework for ensuring the reliability and accuracy of information sources. Its value proposition is cantered on the integration of automated data collection with human validation processes, allowing for a comprehensive approach to fact-checking. This dual mechanism enhances the credibility of the information being analysed, as human validators can provide nuanced insights that automated systems may overlook. For fact-checkers and researchers, this functionality is invaluable, as it streamlines the verification process and ensures that the data used for analysis is trustworthy. Furthermore, the ability to monitor social media and news feeds in nearly real-time equips users with the tools needed to respond swiftly to emerging false information, thereby enhancing their overall effectiveness in combating mis/disinformation.

KER#3) Analytics Value Proposition: The Analytics delivers significant value by providing in-depth insights into the dynamics of false information across social media platforms and other digital spaces. Its value proposition lies in the ability to conduct comprehensive analyses of social network behaviours and identify trends in disinformation spread. For policymakers and researchers, these analytics are crucial for understanding the broader implications of false information on society and for developing informed strategies to counteract its effects. By offering tools such as the Infodemic Observatory and Social Network Analysis, the Analytics area empowers users to make data-driven decisions, craft effective communication strategies, and enhance their overall resilience against false information campaigns. This capability not only supports immediate response efforts but also contributes to long-term policy development and public awareness initiatives.

<u>Unique selling point</u>

The **unique selling point (USP)** of the **AI4TRUST** solution lies in its comprehensive, multi-faceted approach to combating **false information** through a combination of advanced **AI-driven tools**, **human validation processes**, and **nearly real-time analytics**. Unlike other platforms, **AI4TRUST** not only provides an **automated system** for monitoring **mal/mis/disinformation**, but it also integrates **human expertise** to ensure the reliability and accuracy of verification. This dual approach enhances the **credibility** of the verification process and empowers **media professionals**, **researchers**, and **policymakers** to make informed decisions based on **trustworthy data**. By offering a **user-friendly Platform** that caters to the specific needs of diverse target customers, **AI4TRUST**

could be positioned as an innovative solution in the market for **anti-disinformation technology**, based on advanced **human-machine hybrid solutions**.

For **fact-checkers**, the **USP** includes trustworthy **AI-based identification** of tampered objects, the ability to monitor social networks, and the creation of a database of **verified AI-tampered content**. These features enhance productivity by allowing fact-checkers to efficiently decide what to investigate. For **journalists**, the unique selling point is the availability of a database of **human-validated information** through the monitoring dashboard, along with **AI-based tools** that quickly verify content across **images**, **audio**, and **video**. This capability accelerates their work by replacing time-consuming self-research and ensures proper interpretation of the results obtained. **Policymakers** benefit from tools that aggregate information and data collected from the **AI4TRUST Platform** on **false information trends**. This enables them to anticipate rapidly evolving false information contexts and obtain **actionable insights** through reports and briefs, facilitating the creation of effective policies. Lastly, for **researchers**, the Platform offers an **API** to retrieve fact-checked data from existing archives, access to **regulation-compliant data** for training and evaluating AI methods, and indices of **verified** and **manipulated content**. This enables researchers to obtain large volumes of data for their studies and facilitates the publication and dissemination of **scientific results**.

Figure 16: *Value proposition Statements*

6. IPR Management

The **AI4TRUST** project adopts a comprehensive strategy for **Intellectual Property Rights (IPR)** management, ensuring the **transparent**, **ethical**, and **responsible** use of project outcomes. This is achieved through mutual agreements on results utilisation and systematic **IP reviews**, fostering a fair and structured approach to intellectual property. Recognising the importance of addressing legal and **IPR-related** matters at the project level, the **IPR management plan** is designed to mitigate potential risks while facilitating **equitable access** to project results. By establishing a framework for **legal considerations** and **IPR tracking**, the project ensures a transparent and ethical approach to innovation and knowledge dissemination.

For all results generated throughout the project, it is essential to reach a joint agreement on **terms and conditions** for the use of the results generated. In particular, further collaboration activities will be carried out with all partners to list and describe **IPR** to avoid conflicts, while simultaneously supporting individual or joint exploitation plans for all consortium members.

To this end, a **systematic review** of all generated results will be conducted, clarifying ownership, the software components implementing the related functionalities, and the dependencies from other components. This action will be carried out with a **collaborative approach**, ensuring that all partner members have a clear understanding of the **IP generated** throughout the project and of the terms and conditions for the use and access of project results.

As the consortium progresses in its efforts to maximise the impact of the **AI4TRUST** project, it acknowledges the importance of addressing joint ownership of results. In the coming months after the submission of **D7.4**, the current consolidation of Platform components will allow further evaluation of the concrete intentions of consortium partners regarding the exploitation of these joint results beyond the project lifetime. This **collaborative approach** will ensure that all partners are aligned in their efforts to leverage the outcomes of the project effectively, fostering a shared commitment to the **sustainable use** of the innovations developed throughout **AI4TRUST**.

IPR(s) Objectives

In this sense, the **AI4TRUST consortium** will place significant importance on the management of **legal** and **IPR** issues within the project framework, aiming to facilitate the effective exploitation and sustainability of its results. The consortium will oversee and address any **IPR** matters, ensuring that innovative ideas generated during the project are thoroughly evaluated, with appropriate management of the project's **foreground IP**. In this regard, the general objectives of **AI4TRUST's IPR Management Strategy** include:

- **Increasing awareness** among project partners regarding terms and issues related to **IP** and its protection, such as **background (BG) IP** and **results IP**, along with access rights and joint ownership.
- **Identifying the assets** produced throughout the project's lifecycle and the corresponding **IP flow**, particularly concerning jointly developed project results.
- Establishing a **framework** for:
    1. Clarifying access needs and rights, as well as ownership and exploitation claims;
    2. Identifying potential **IP conflicts** within the consortium and externally, while avoiding **IP infringement**;
    3. Making decisions regarding the **IP protection** of each **Key Exploitable Result** (KER) identified and the joint ownership of results;
    4. Implementing those decisions appropriately based on the selected **exploitation route** and ownership status.

With these considerations in mind, the following outlines the methodology that will be employed for **IPR Management** in the context of **AI4TRUST**.

IPR(s) Methodology

It is essential to highlight that the **IPR(s) Methodology** will play a crucial role in the process of recognising and addressing the primary **IPR** issues of the project. In particular, the **IPR(s) Methodology** will be implemented in a sequential manner as outlined below:

1. **Step 1**: Identifying **Ownership** of **Key Exploitable Results**.
2. **Step 2**: Selecting the appropriate **IP protection** tools for the identified **Key Exploitable Result(s)**.
3. **Step 3**: Identifying **Key Exploitable Results IPR(s) Ownership** distribution among partners.

The subsequent subsections will detail the step-by-step application of the **IPR(s) Methodology** to our project:

a. Identifying Ownership of Key Exploitable Results

In the initial phase of the **IPR(s) Methodology**, the **KERs** and their respective ownership will be determined. The primary objective of this phase is to ascertain **IP ownership** and **exploitation claims**, while also proactively identifying potential conflicts for each **KER**. It is essential to recognise that the ownership of certain **KERs** may be claimed by two or more partners, as they have jointly produced these results, making it impossible to separate the results.

b. Selecting the appropriate IP protection tools for the identified Key Exploitable Result(s)

The second step of the **IPR(s) Methodology** focuses on deciding on issues pertaining to **IP protection** to initiate the first steps towards **IP protection**. The **KERs** identified will be reported and

further detailed in the section "Key Exploitable Results" with a brief description of the main contributing partners, and the proposed **IP protection tools**. It is important to note that all considerations regarding joint ownership for a specific **KER** identified will be agreed upon by the consortium's partners.

c. Identifying Key Exploitable Results Ownership Distribution among Partners

In the third step of the **IPR(s) Management** methodology for the **KER(s)** of the AI4TRUST project, the consortium will assess the most appropriate procedure for determining the distribution of ownership among partners for each identified KER. Given that these key exploitable results are the product of collaborative efforts among multiple partners, they will inherently possess **joint ownership**. The consortium will engage in discussions to establish clear guidelines and procedures that outline the respective contributions each partner could claim, ensuring that **ownership rights** are fairly allocated. This **collaborative approach** will facilitate **transparency** and foster a **cooperative environment** as the partners work together to maximise the impact and **commercialisation potential** of the project outcomes.

d. Potential Conflicts and IP Issues

The **IPR Methodology** will serve as a robust tool for structuring and monitoring the **IPR management** process, while also assisting in the prevention of possible **IP conflicts** concerning ownership and exploitation claims. Project partners will confirm a final list of project **KER(s)** and ascertain their ownership and exploitation claims, while simultaneously organising the management of **IPR(s)** in accordance with the primary rights and obligations outlined in the **Grant Agreement** and the **Consortium Agreement** of **AI4TRUST**. All aspects of **IPR management** will be governed in accordance with the provisions established in the **Grant Agreement** and the **Consortium Agreement**, including the resolution of any arising issues.

e. IP Protection Measures and Tools

Various instruments exist for safeguarding **intellectual property (IP)**. Choosing the right **IP protection measures** depends on the unique attributes of the **IP** in question and the strategic objectives of the **IP** owner. The following list includes some of the tools and measures that could be employed to ensure effective **IP protection** of the **KER(s)** results identified:

- **Trademarks**
- **Copyright**
- **Industrial designs**
- **Patents**
- **Utility models**
- **Trade secrets**
- **Licensing Agreements**

- **Assignment Agreements**
- **Non-Disclosure Agreements (NDAs)**
- **Collaborative Agreements**
- **IP Registration**
- **Databases**
- **Digital Rights Management (DRM)**
- **Open-Source Licensing**

Further details about each of these protection tools will be outlined in the forthcoming deliverable **D7.6**[66].

7. Conclusions

The **AI4TRUST project** represents a significant step forward in the fight against disinformation. The initial version of this deliverable (**D7.2**) laid the groundwork for a strategy aimed at efficient innovation, scaling, and utilisation of the **AI4TRUST** outcomes. The feedback received from external experts during the **RP1 review** has been instrumental in shaping the revisions presented in this deliverable **D7.4**, ensuring that our strategies are robust, actionable, and aligned with best practices in the field. **D7.4 not only revises the initial strategies but also introduces a systematic plan that bridges the gap between project results and mission-oriented exploitation of the AI4TRUST Platform**.

This deliverable outlines the **sustainable outcomes** and **preliminary exploitation paths** for the identified assets, setting the stage for the final plan that will guide the innovation, exploitation, and sustainability of AI4TRUST offerings beyond the project's lifespan. Throughout the document, we have established a comprehensive understanding of the **market landscape**, identifying the needs and opportunities that AI4TRUST aims to address. The methodology employed has allowed for a structured approach to defining the project's goals, functionalities, and **key exploitable results**, which are essential for stakeholders.

The overview of the **AI4TRUST Platform** highlights its potential to empower **media professionals**, **policymakers**, and **researchers** with advanced tools for monitoring and combating false information. Individual plans for **innovation**, **exploitation**, and **sustainability** have been developed

[66] D7.6 - Innovation, Exploitation and Sustainability Plan v3 (due by M38 - February 2026, https://ai4trust.eu/public-deliverables/)

for each consortium partner, ensuring that their unique contributions align with the project's overarching objectives.

By integrating insights from various stakeholders and incorporating the **recommendations outlined in the "General Project Review Consolidated Report (HE)", dated 28 June 2024**, following the project's first Review Meeting, a **tailored approach that addresses the unique needs of potential customers** has been crafted, thereby enhancing the overall impact of the AI4TRUST solution. Furthermore, the **engagement strategy** outlined in this document fosters collaboration among stakeholders, enhancing the visibility and impact of the AI4TRUST project. Finally, the **Intellectual Property Rights (IPR)** management plan is crucial for ensuring that the innovations developed are protected and utilised effectively.

Looking ahead, **D7.6 "Innovation, Exploitation and Sustainability Plan v3"**, due at **M38**, will further solidify our commitment to **sustainability** and **exploitation**. This final plan will encapsulate the lessons learned throughout the project lifecycle and provide a comprehensive framework for the ongoing development and commercialisation of the **AI4TRUST Platform**. It will focus on ensuring that the innovations developed are not only viable in the short term but also sustainable in the long run, thus providing a final **IPR Management strategy** outlining the partners' intentions on how to proceed with joint ownership of results for the **AI4TRUST solution** after the project's conclusion. In fact, in the coming months after the submission of **D7.4**, the current consolidation of **Platform components** will allow proceeding, with a discussion among partners on **joint ownership of results**.

In conclusion, the framework established for innovation, **exploitation**, and **sustainability strategy** provides a **forward-thinking perspective** that will guide the AI4TRUST project in maximising its impact. By addressing the complexities of disinformation in a rapidly evolving digital landscape, AI4TRUST is poised to make a meaningful contribution to the integrity of information dissemination and the protection of public discourse.

8. Annex I

This annex provides details regarding the **workshop conducted on the 9th of January 2025** with the end-user partners of the **AI4TRUST** project. The workshop focused on **four key canvases**: the **Personas Canvas**, the **Value Proposition Canvas**, the **Ad-lib Value Proposition Template**, and the **Prototype Canvas**. It was specifically designed to address the needs of four **Target Customers**: the **Researcher**, the **Fact-checker**, the **Journalist**, and the **Policymaker**. The outcomes of the workshop are integrated into the relevant sections of this deliverable, while further methodological details can be found in **"Methodologies"** (chap. 3).

8.1. Personas Canvas

This section presents the feedback collected from the **end-user partners** related to the **Personas Canvas**. A separate canvas for each **Target Customer** is provided (Figures 17, 18, 19, and 20).

Researchers

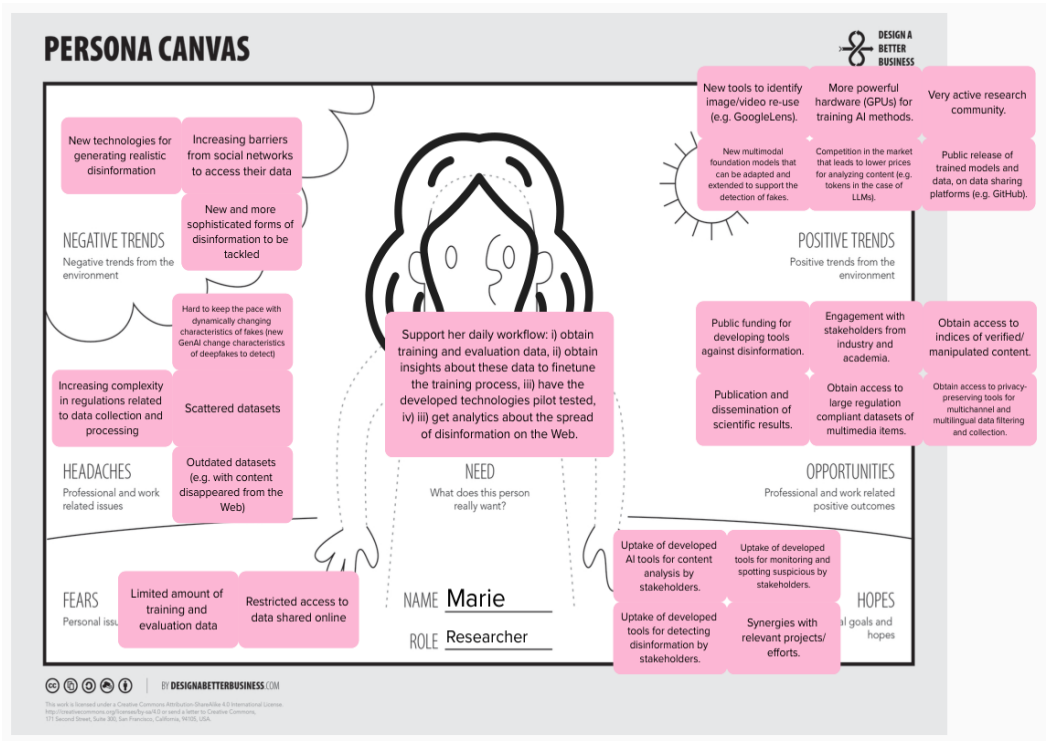

Figure 17: *Persona Canvas for the "Researcher" Target Customer*

Fact-checkers

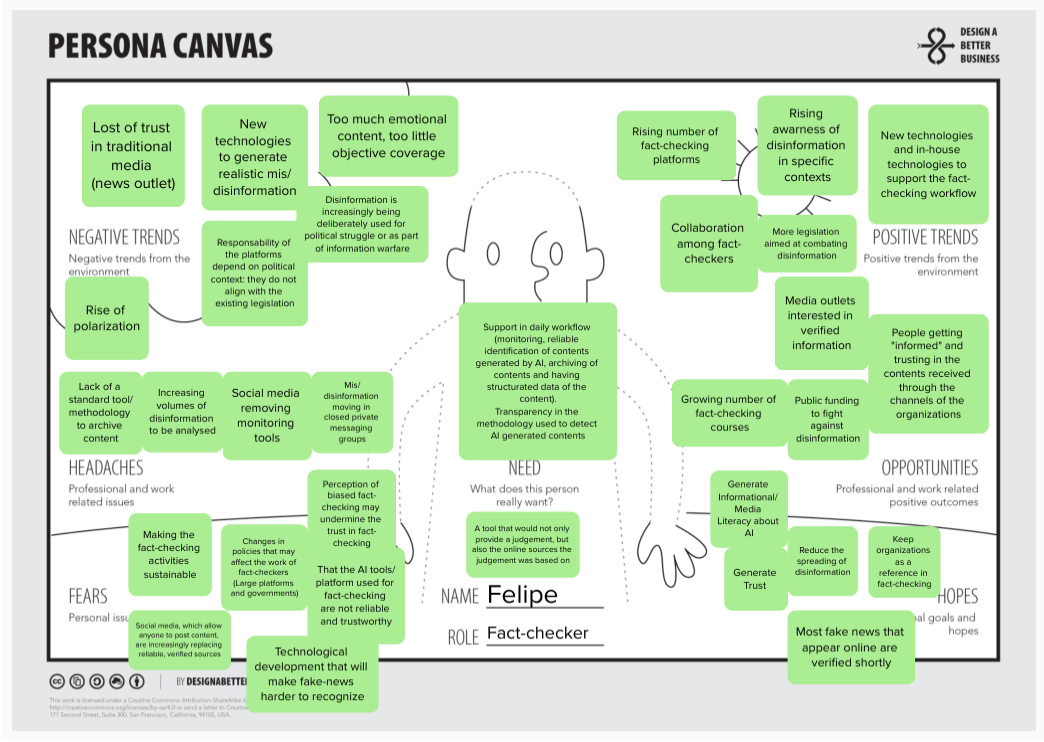

Figure 18: *Persona Canvas for the "Fact-checker" Target Customer*

Journalists

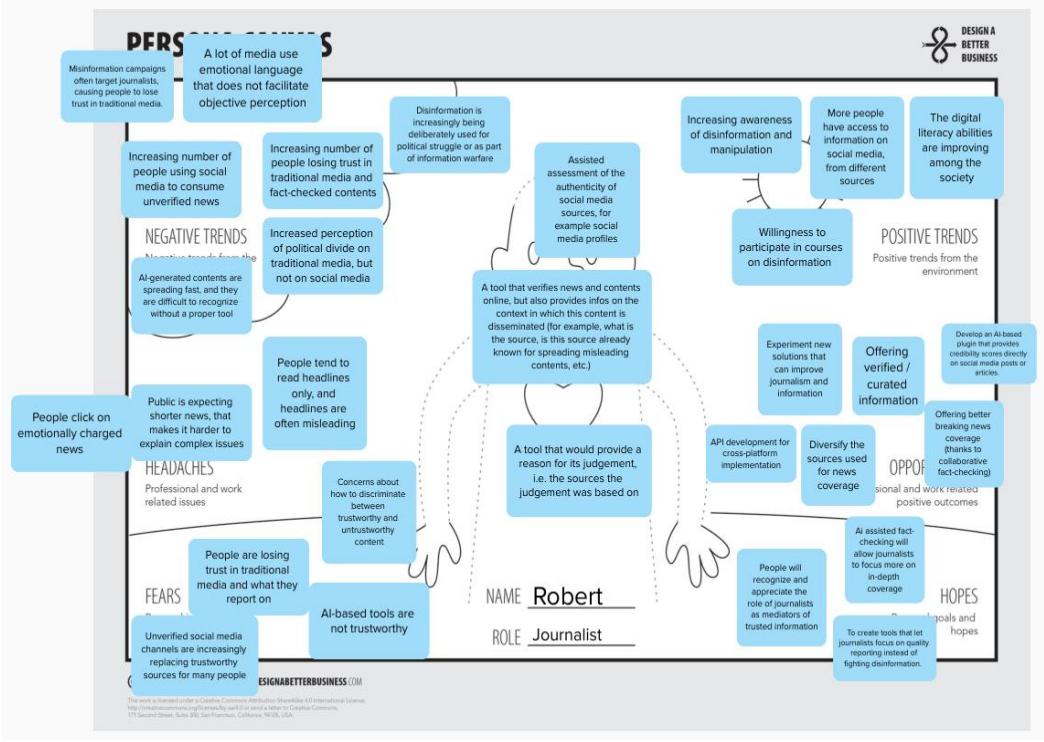

Figure 19: *Persona Canvas for the "Journalist" Target Customer*

Policymakers

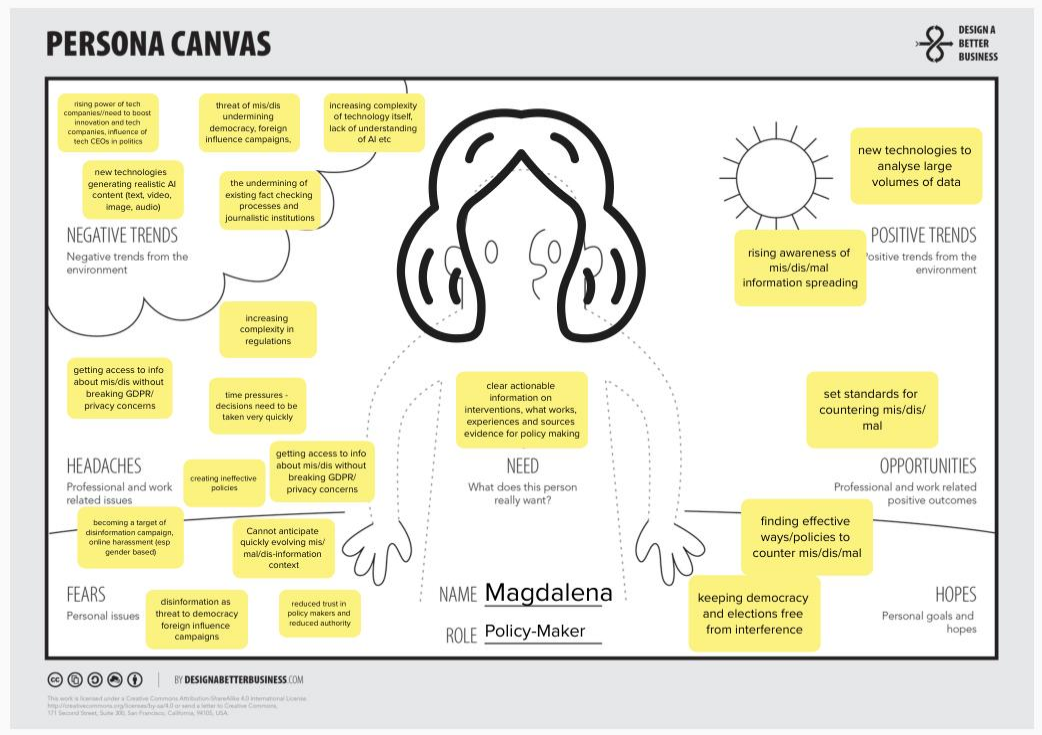

Figure 20: *Persona Canvas for the "Policymaker" Target Customer*

8.2. Value Proposition Canvas

In this section are reported **the feedback collected from the end-user partners related to the Value Proposition Canvas.** A separate canvas for each **Target Customer** is provided (Figures 21, 22, 23, and 24).

Researchers

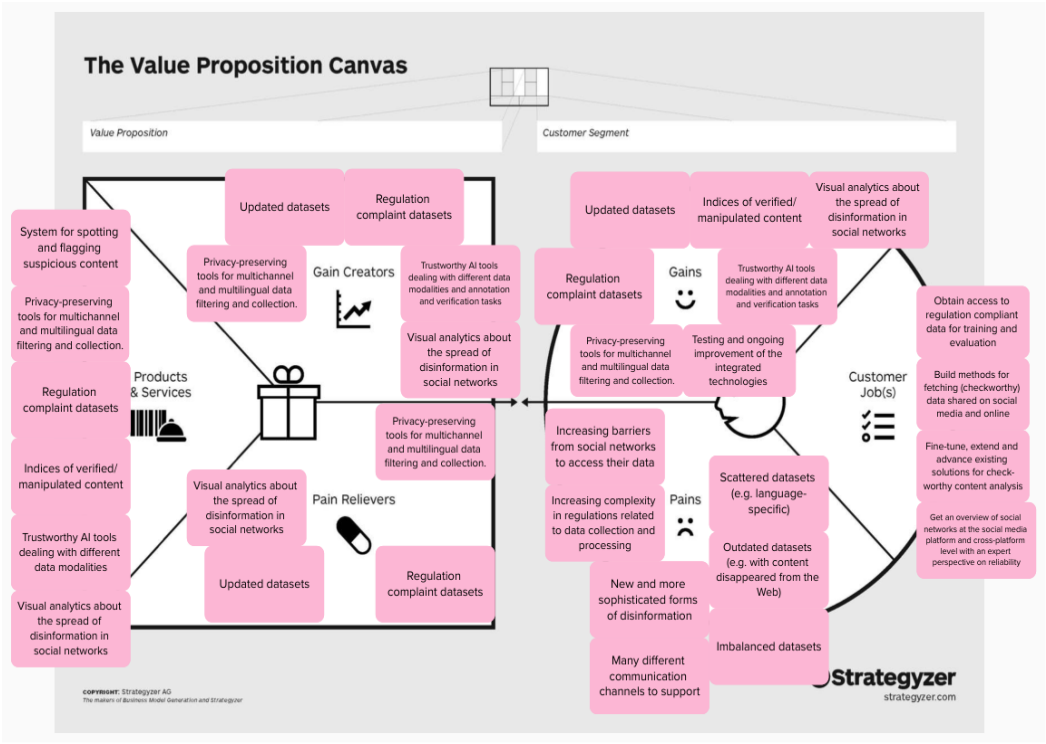

Figure 21: *Value Proposition Canvas for the "Researcher" Target Customer*

Fact-checkers

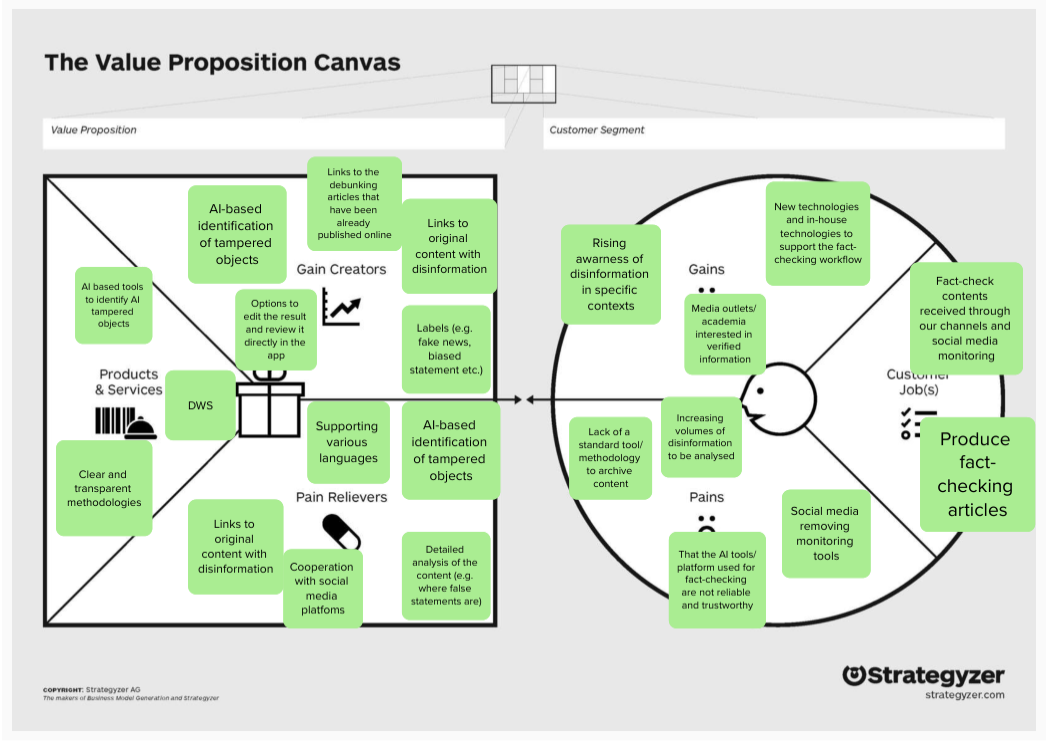

Figure 22: *Value Proposition Canvas for the "Fact-checker" Target Customer*

Journalists

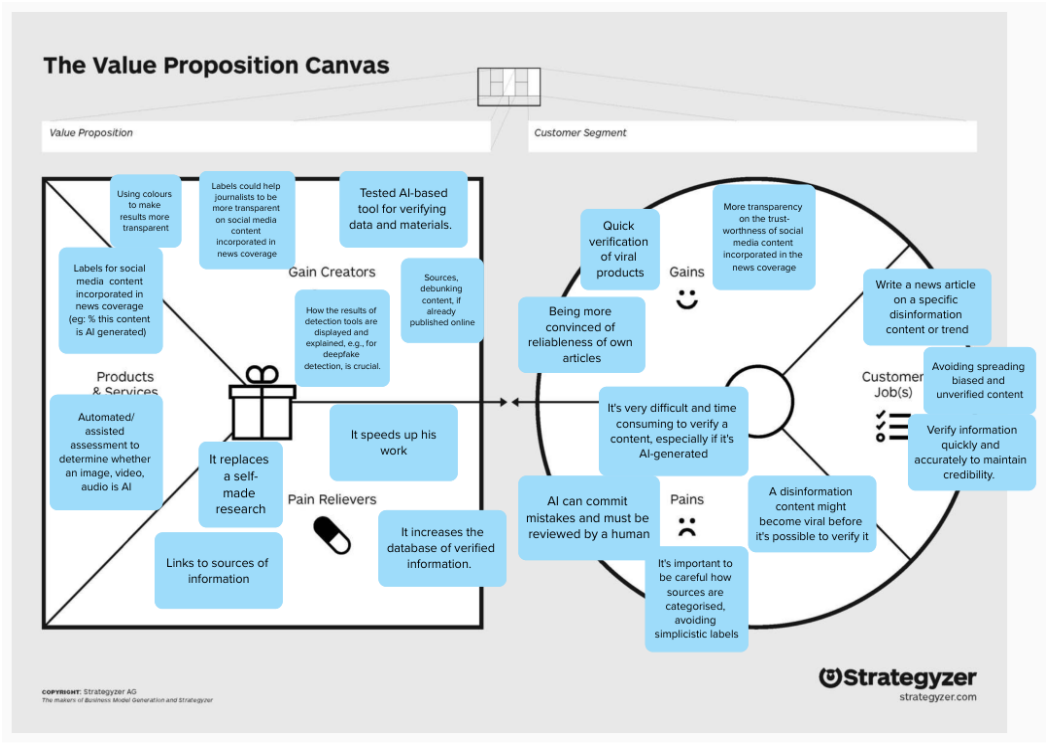

Figure 23: *Value Proposition Canvas for the "Journalist" Target Customer*

Policymakers

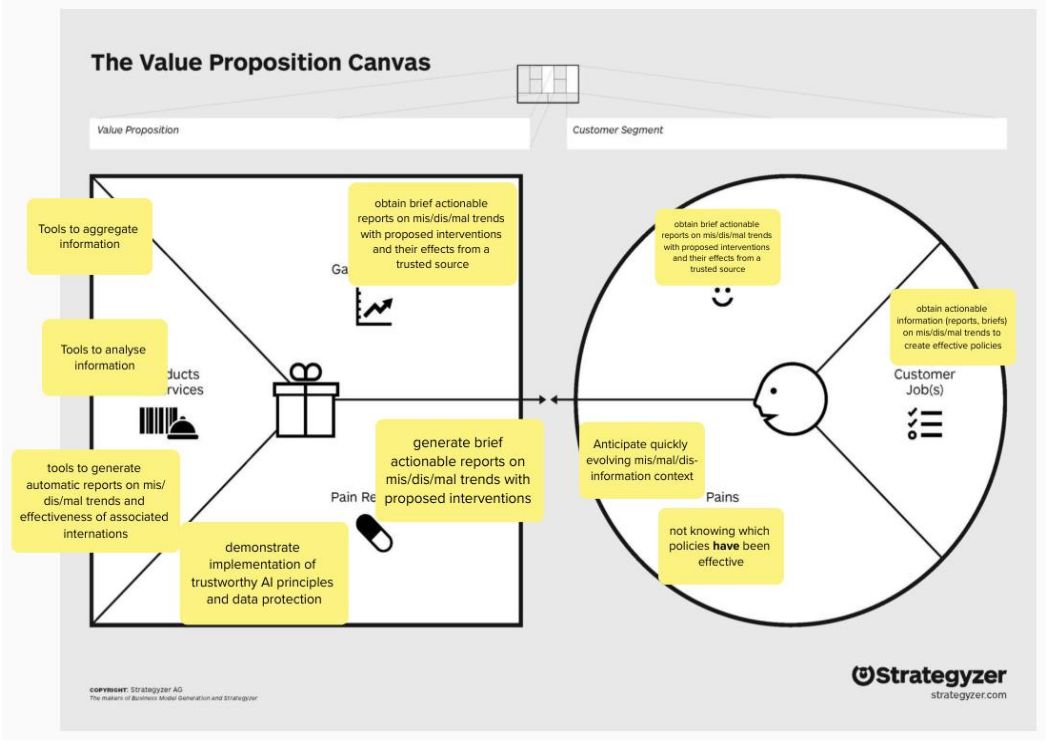

Figure 24: *Value Proposition Canvas for the "Policymaker" Target Customer*

8.3. Ad-lib Value Proposition Template

This section presents the feedback collected from the **end-user partners** related to the **Ad-lib Value Proposition Template**. One or more canvases for each **Target Customer** are provided (Figures 25, 26, 27, and 28).

Researchers

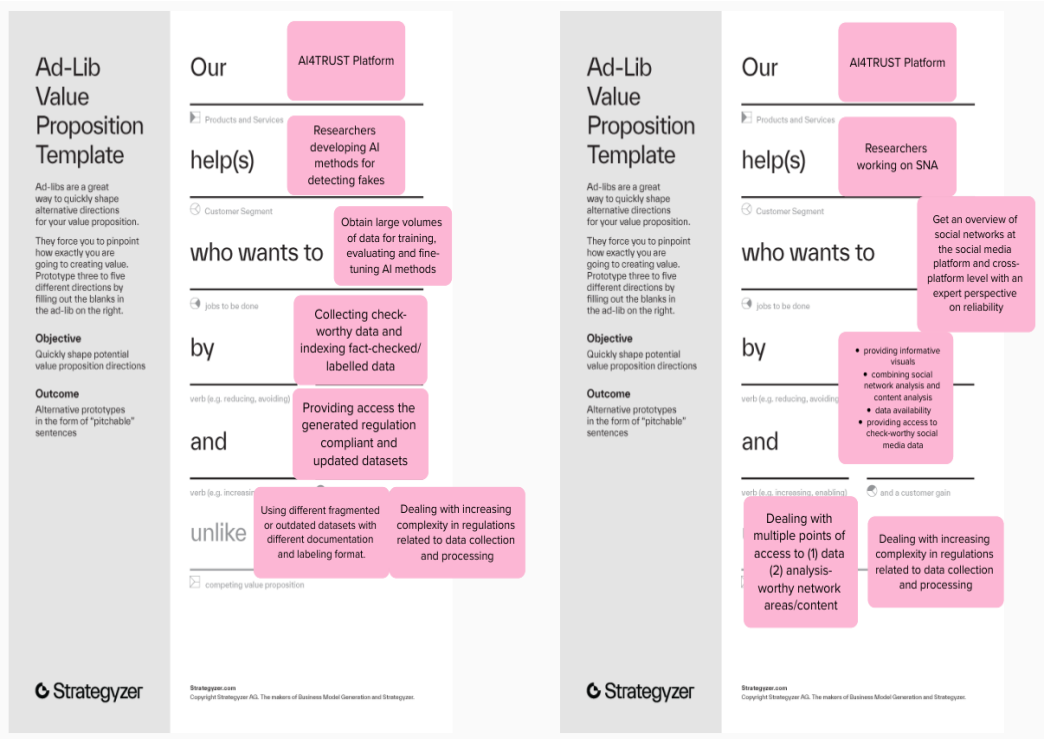

Figure 25: *Ad-lib Value Proposition Template for the "Researcher" Target Customer*

Fact-checkers

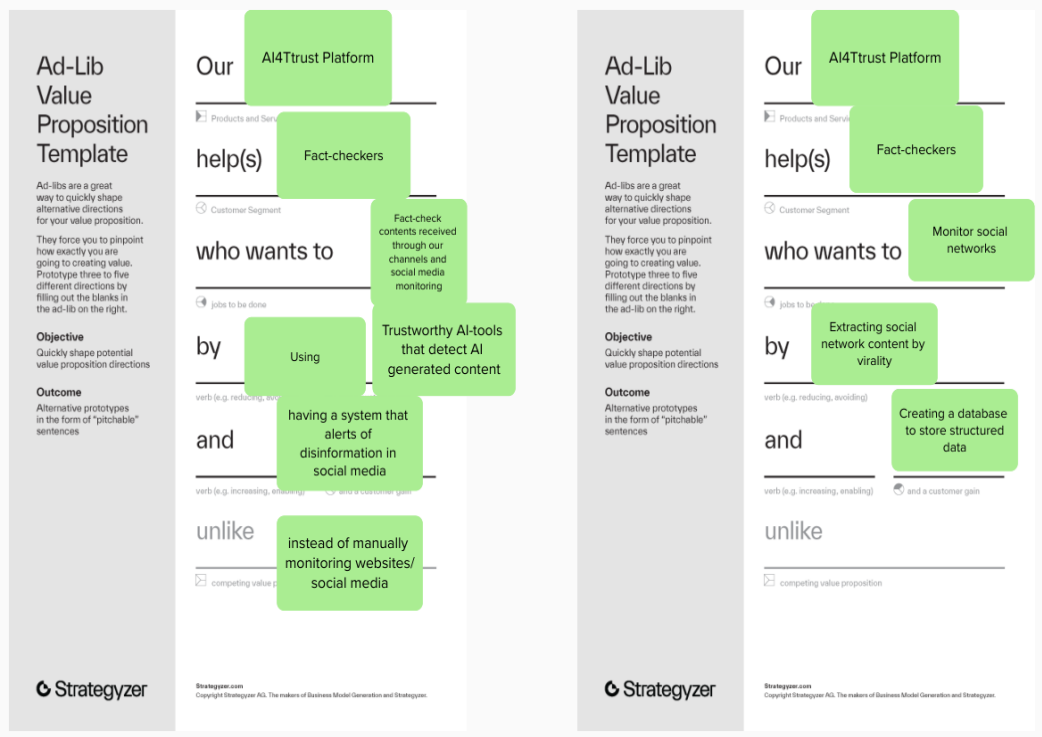

Figure 26: *Ad-lib Value Proposition Template for the "Fact-checker" Target Customer*

Journalists

Figure 27: *Ad-lib Value Proposition Template for the "Journalist" Target Customer*

Policymaker

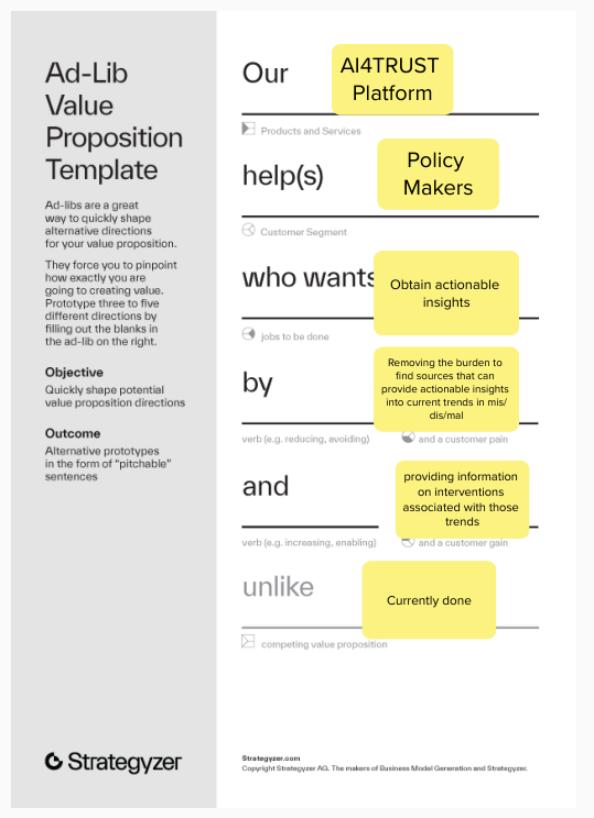

Figure 28: *Ad-lib Value Proposition Template for the "Policymaker" Target Customer*

8.4. Prototype Canvas

This section presents the feedback collected from the **end-user partners** related to the **Prototype Canvas**. One canvas for each **Target Customer** is provided (Figures 29, 30, 31, and 32).

Researchers

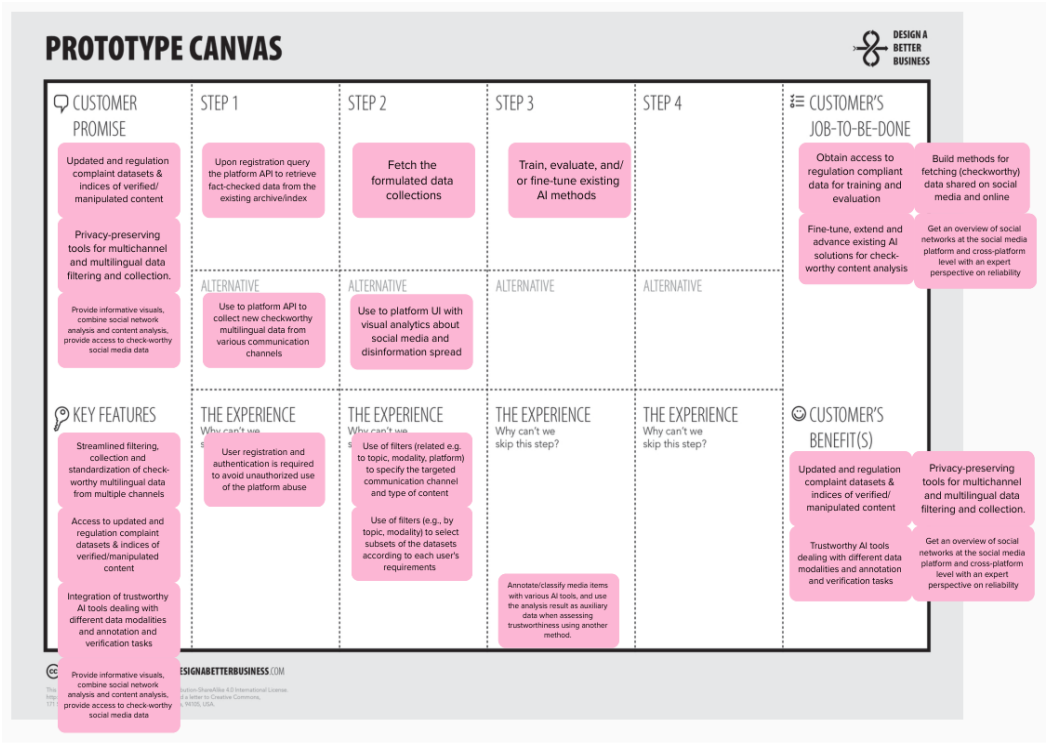

Figure 29: *Prototype Canvas for the "Researcher" Target Customer*

Fact-checkers

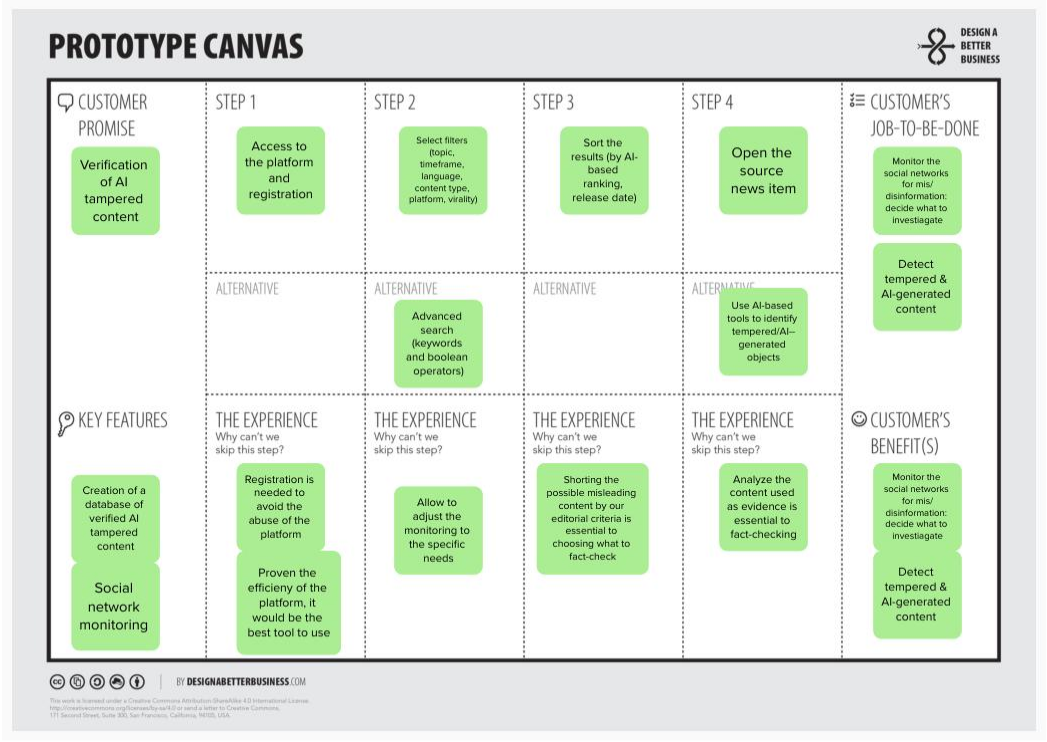

Figure 30: *Prototype Canvas for the "Fact-checker" Target Customer*

Journalists

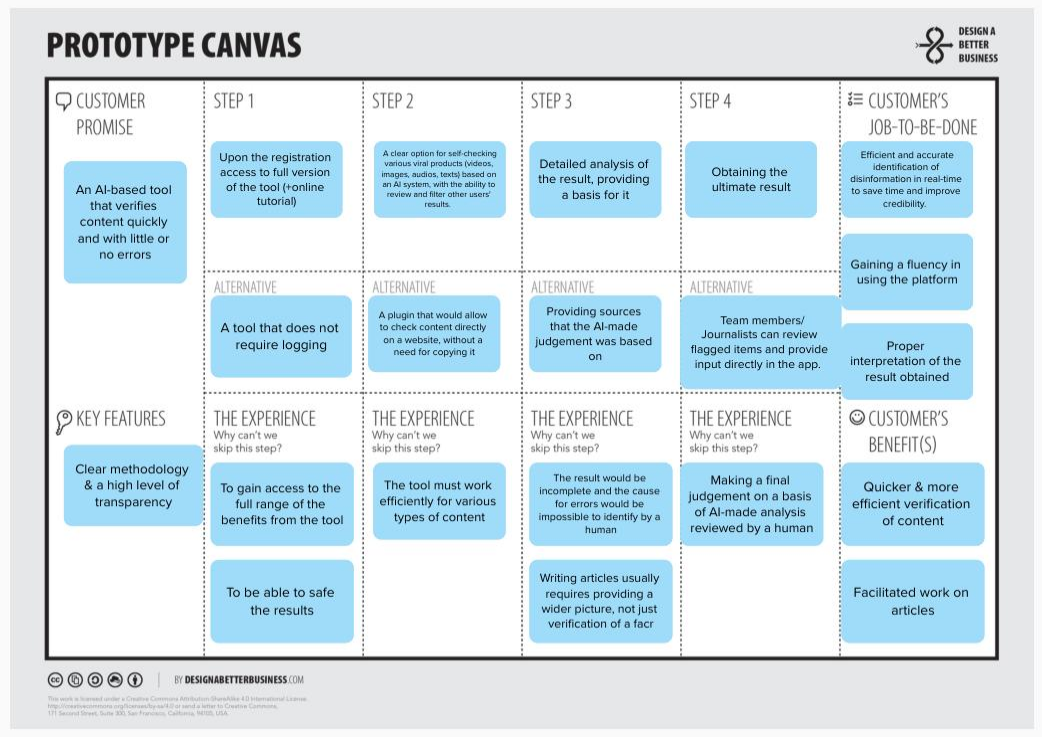

Figure 31: *Prototype Canvas for the "Journalist" Target Customer*

Policymakers

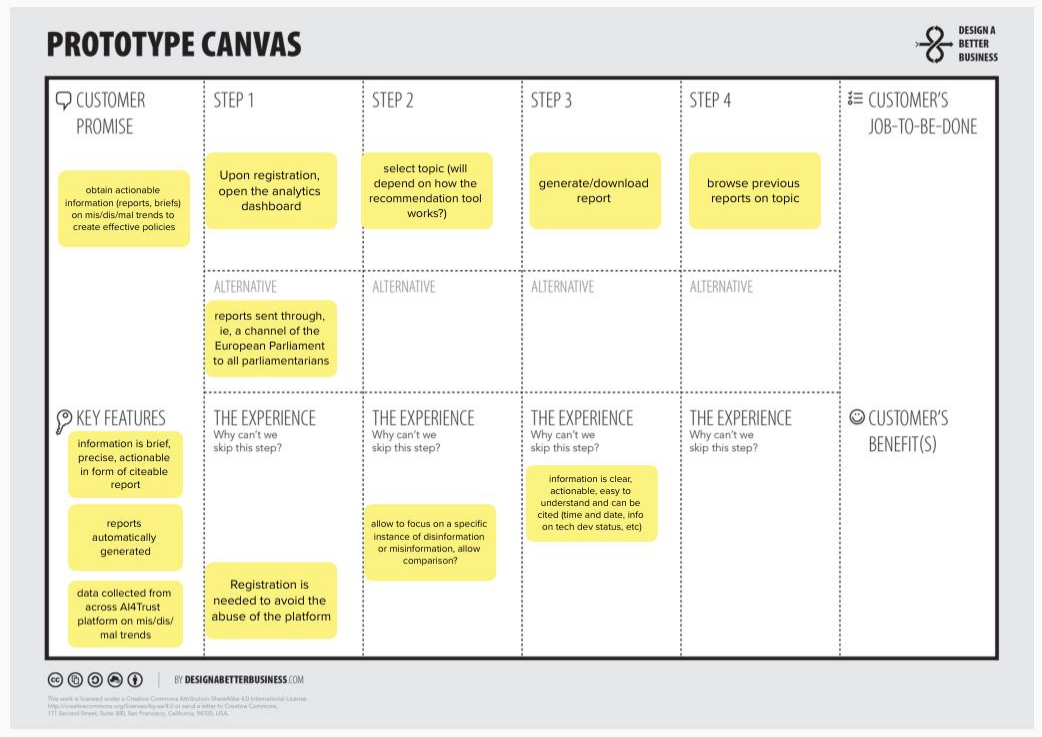

Figure 32: *Prototype Canvas for the "Policymaker" Target Customer*

9. Annex II

This annex presents in detail the **updated Individual Innovation, Exploitation, and Sustainability Plan for each asset of each partner within the AI4TRUST project**. It outlines the classification and assessment of individual assets based on their relevance and potential impact, derived through consultations and interviews with **consortium partners**. Both **tangible** and **intangible assets** are identified, emphasising key resources, expertise, and strategic advantages. Additionally, each partner has developed **customised plans** to foster innovation, maximise exploitation potential, and ensure long-term sustainability.

9.1. Fondazione Bruno Kessler (FBK)

Fondazione Bruno Kessler (FBK) is bringing to this project an extensive set of experience in the field of **Data Engineering, Data Science, Social Network Analysis (SNA), Natural Language Processing (NLP), and the development of Large Language Models (LLMs)**.

The AI4TRUST Platform builds upon the **experience of the Covid19 Infodemics Observatory**[67] and other important technological **assets developed by FBK's Digital Society and Augmented Intelligence Centres**. FBK aims to widely expand from this early experience creating a **new social listening data stream platform** for automated collection and processing of social and news media, compliant with EU-GDPR and other related European and national legislation and tailored to the needs of scientific research in the field of mis/disinformation.

FBK also develops **AI models for assessing the check-worthiness of textual claims** in multiple languages, and for retrieving previously fact-checked claims across languages. Additionally, FBK pioneers AI models for generating verdicts on claim veracity, addressing pragmatic aspects like style and emotion.

Finally, the research experience and expertise that will be gained thanks to the dynamic partnerships within AI4TRUST with media and fact-checking organisations not only enhances FBK capacity for **technology transfer** but also fuels its ability to continually generate **fresh research ideas**, driving innovation and keeping this institute at the forefront of the fields of Artificial Intelligence (AI) and the study of Online Communication, strengthening its position in future projects.

The initial focus on scientific applications has been and will be expanded to include **potential commercialisation avenues** for some technological assets, specifically the social listening data

[67] https://covid19obs.fbk.eu

stream platform and AI models for claim assessment and fact-checking. In this regard, FBK - adequately supported by suitable companies or SMEs - will address markets in: **1) Research and academia**, by providing GDPR-compliant data and tools to support scientific studies on disinformation; **2) Media and journalism**, offering AI models to enhance fact-checking workflows and combat mis/disinformation effectively; **3) Public and private sector entities** interested in advanced tools for monitoring and analysing social and news media. FBK plans to engage with the above **potential users/customers** by organising **workshops and conferences** to showcase the capabilities of its tools, collaborating on **further pilot projects** and establishing agreements with media and fact-checking organisations, and providing **continuous support and updates to professional users** of its platform and AI tools.

To this aim, FBK envisions a **dual business model**: **a) Open access for the scientific community,** ensuring broad reuse and collaboration; **b) Open-source platform with optional subscription-based APIs or licensing models for professional users** (e.g., media organisations and fact-checking entities), providing tailored solutions while ensuring long-term sustainability. In this framework, FBK will focus on releasing certain tools under open-source licenses, promoting transparency, collaboration, and community-driven development. For novel AI methodologies, FBK will explore patent applications where relevant, while ensuring the core platform remains freely accessible to encourage innovation. The anticipated **timeline for market entry post-project** will be as follows: **Year 1**: Finalise development and conduct pilot tests with early adopters; **Year 2**: Launch open-source platform, with optional premium API subscriptions and licensing models; **Year 3**: Expand market reach and scale operations based on community feedback and market trends. By addressing these aspects, FBK aims to maximise the impact of its contributions to AI4TRUST, ensuring both scientific and societal benefits while paving the way for long-term exploitation and sustainability.

Asset	**Social listening data stream**
Type	Technology/Data
Description	FBK will develop a social listening platform that automatically collects and processes social media and news from multiple sources. This platform will elaborate on the raw data, transforming them into GDPR-compliant data while capturing text, shared audiovisual content, and social interactions, making it suitable for scientific use.
Target Groups/Beneficiaries	Researchers

Asset	**Social listening data stream**
Innovation, Exploitation and Sustainability Plan	The Data Stream platform will be an asset also after the end of the AI4TRUST project, as it could continue to generate relevant data for scientific research and could be further adapted and expanded to cover other data sources. The data collected will inform the scientific work of the AI4TRUST consortium. Post project, part of the data collected will be, in aggregated form, shared with the whole scientific community to ensure further future research developments. FBK envisions further leveraging the social listening platform beyond the project's scope by building long-term partnerships with academic institutions and public organisations to ensure continuous data provision. The platform can also be expanded to incorporate additional real-time data sources and diverse media formats, enhancing its versatility for evolving research and market needs. Moreover, FBK plans to offer premium subscriptions for professional users, featuring advanced analytics and data visualisation capabilities tailored to their specific requirements.

Asset	**AI model for check worthiness of textual claims**
Type	Knowledge/Technology
Description	FBK will develop an AI model that is able to assess whether a textual claim is check-worthy or not in different languages. The model will be trained either on existing datasets for check-worthiness already available to the Natural Language Processing (NLP) community as well as on a dataset that has been specifically created during AI4TRUST for the Italian language.
Target Groups/Beneficiaries	Researchers and professionals
Innovation, Exploitation and Sustainability Plan	In AI4TRUST, FBK will build on (monolingual and multilingual) LLMs and fine-tune them both on existing datasets and on the Italian data that FBK has specifically created for the task. Multilingual solutions will also be explored and evaluated to cover as many languages as possible. The trained models will be exposed through APIs and will allow researchers and professionals to filter claims that may be worth checking for veracity, before further processing them. The trained model will

	be made available too for further reuse and improvement both from professional fact-checkers and researchers. To maximise the utility of this asset, FBK aims to collaborate with international organisations to adapt the model for diverse regional contexts and languages. Educational resources and workshops will be developed to help users integrate the tool effectively into their workflows. Additionally, FBK plans to monetise the model by incorporating it into larger fact-checking platforms or offering it as a standalone service, ensuring its accessibility to both researchers and professional users.

Asset	**AI model for retrieval of previously fact-checked claims**
Type	Technology
Description	FBK will develop an AI model that, given a claim, provides in output a ranked list of previously fact-checked claims that are like the one of interest. The tool can work cross-lingually and can be adapted to include previously fact-checked claims of professional fact-checkers.
Target Groups/Beneficiaries	Researchers and professionals
Innovation, Exploitation and Sustainability Plan	In AI4TRUST, FBK will build on works related to previously fact-checked claim retrieval and extend them to make them more robust and accurate, also cross-lingually. The source code of the application will be released, so that fact-checking companies will have the possibility to use the tool beyond the project ends and add their previously fact-checked claims to the pool of existing ones on their own servers. FBK intends to position this tool as a key resource for global fact-checking networks by integrating it into centralised repositories of fact-checked claims. Collaboration with media organisations will ensure its seamless integration into their workflows, while customisation options will be provided to address the specific needs of different stakeholders. These efforts will make the tool a cornerstone for combating mis/disinformation on a global scale.

Asset	**AI models for verdict generation**
Type	Knowledge/Technology

Description	A set of trained Language models for generating short texts that discuss the veracity of a claim provided a fact-checking article connected to the claim. The input of these models is textual (the claim and the debunking article concatenated together) while the output is (1) a text that can be used by stakeholders for performing social correction on social media platforms or similar tasks and (2) a list of the most relevant sentences from the input article. These language models are designed to focus primarily on the veracity of the claim, while also taking into account pragmatic aspects such as style and emotions, which are essential for effective communication, particularly on social media platforms.
Target Groups/Beneficiaries	Fact-checkers/journalists, citizens. In general, whoever needs to verify discusses the veracity of a claim from an argumentative point of view and not only from a classification perspective.
Innovation, Exploitation and Sustainability Plan	In AI4TRUST, FBK will build on existing LLMs that are already in place and fine-tune them on data that FBK has specifically created for the task. The trained models will be exposed through API and run internally to the platform, to provide suggestions to claims to which the veracity needs to be assessed by the users. FBK plans to explore partnerships with social media platforms to embed these models for automated or assisted moderation and misinformation correction. Additionally, the models could be integrated into public communication campaigns led by governments, media channels, or NGO/CSOs, enabling more effective engagement with audiences. Subscription-based access to the models will be offered for customised use cases, such as content verification or sentiment-aware communication strategies, ensuring their impact extends across diverse domains.

Asset	**Infodemic Observatory Tool**
Type	Technology
Description	The “Infodemic Observatory” will be capable of tracking aggregated statistical information on the quantity of misleading news circulating in the past week on different topics and across various social media platforms, both in absolute terms and relative exposure to the public. The observatory design is inspired by the COVID-19 Infodemics Observatory platform developed by FBK. To characterise the risk for social media users, aggregated information will be made available to media

	practitioners, fact-checkers, journalists and policy makers end-users: - The total number of news collected in the platform; - The total volume of messages sharing unreliable news items - The total exposure these messages are expected to have on social media; - An infodemic risk index illustrating how likely it is to encounter unreliable news at that moment in time. This information will be provided as a time series illustrating the evolution in time of these indicators in the different languages, topics and platforms. At the same time, once aggregated over all different languages, these indicators will be also grouped geographically and represented as a map evolving over time.
Target Groups/Beneficiaries	Project partners and future users of the platform (i.e., fact-checkers, media workers, policy makers)
Innovation, Exploitation and Sustainability Plan	This tool will rely on social media data collected in WP2, fact-checking conducted by social media partners, and leverage disinformation predicting tools developed in WP3. The innovation lies in the unique alliance of fact-checking and technology to produce a novel tool for risk estimation. This component is as sustainable as the data and technologies are, and will be optimally developed in a pipeline that continuously collects and analyses social media data.

9.2. ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS (CERTH)

The **CERTH** team has substantial experience in the field of **media analysis** and **verification**, with particular expertise in the following methods:

- **Video fragmentation and keyframe selection**;
- **Detection of deepfake and AI-synthesised content**;
- **Visual content classification**;
- **Multimodal representation learning**.

CERTH's research experience and expertise in this area, combined with the **technology requirements** of the fact-checking and media organisations involved in AI4TRUST, will enable us to drive new advancements in our existing solutions. Furthermore, this collaboration has the

potential to lead to the development of innovative methods that could significantly improve the performance of our techniques and services.

Asset	**Tool for reverse video search on the Web**
Type	Technology
Description	A web-based technology that allows a user to temporally segment a video into visually coherent fragments, extract a set of representative keyframes and use them to perform automated keyframe-based search for this video on the Web through various search engines (e.g., Google Lens, Yandex, Bing). The detection of one or more near-duplicates of the query video that have been posted on the Web in the past, indicates a video re-use; then, based on the claim that is associated with the video, the fact-checker can decide on whether the video has been used out of its original context for misleading the viewers about an event.
Target Groups/Beneficiaries	Fact-checkers/journalists (to seek assistance in content verification). Media integrity teams at various companies/organisations (to verify content e.g., in a Know Your Customer (KYC) application setting), Researchers (for comparing their approach).
Innovation, Exploitation and Sustainability Plan	Contrary to the established approaches for detecting video re-use on the Web - that rely on the manual extraction of video frames and their submission to image-based search engines (e.g., TinEye[68], RevEye[69]), or the use of technologies that enable reverse search in closed collections of images and videos (e.g., Berify[70], Videntifier[71]) and specific video sharing platforms (DataViewer[72]) - the developed tool enables a more extended search on the Web. Moreover, it advances previous solutions by integrating an AI-based method for video thumbnail selection and automating interaction with multiple search engines, thus facilitating significantly the detection of near-duplicates of the video on the Web and the debunking of fakes that rely on video re-use. The developed tool can be used both as a stand-alone

[68] https://tineye.com/

[69] https://chromewebstore.google.com/detail/reveye-reverse-image-sear/keaaclcjhehbbapnphnmpiklalfhelgf

[70] https://berify.com/

[71] https://www.videntifier.com/

[72] https://citizenevidence.amnestyusa.org/

	technology via its web-based UI, as well as an integrated component of the AI4TRUST Platform. The stand-alone version of the tool is hosted at CERTH and is publicly usable for free. The development of tools for detecting dis/misinformation are among the core research topics of our team, and thus we will be seeking additional funding from different sources (e.g., research grants, commercial contracts/licensing for customized versions) to ensure maintenance and improvement of the developed tool beyond the end of AI4TRUST. We plan to sustain this asset for a period up to 3 years after the project ends in collaboration and in agreement with the rest of the consortium. Adjustments and modifications regarding the communication with the different search engines and the retrieval of results are within the scope of the sustainability plan.

Asset	**Deepfake image/video detection**
Type	Technology
Description	An image or video file is classified as being real or generated using one of the popular deep fake generation models. This includes fully AI-generated images and deepfake videos with visual or/and audio manipulations. A score between 0-100 is also produced and expresses the confidence of the decision. In addition, in the case of videos, distinct scores are assigned per video segment to help users localise the deep fake in the video.
Target Groups/Beneficiaries	Fact-checkers/journalists (to seek assistance in content verification). Media integrity teams at various companies/organisations (to verify content e.g., in a Know Your Customer (KYC) application setting), Researchers (for comparing their approach).
Innovation, Exploitation and Sustainability Plan	In AI4TRUST, CERTH built on an existing proprietary deepfake image/video detection service, and through pertinent research, improved it by developing and integrating new multimodal and video deepfake detection models that can detect content coming from recent generative AI models with increased accuracy compared to state-of-the-art methods. Also, explainability is achieved through the service by indicating the modified video parts letting the human evaluator take the final decision. Finally, further improvements in terms of robustness to real-world conditions have taken place (e.g., video parts without faces or

	talking are excluded from the analysis). Given that deepfake image and video detection are strategic research topics for our team, we will be seeking additional funding from different sources (e.g., research grants, commercial contracts, and licensing) to ensure maintenance and improvement of the service beyond the end of AI4TRUST. We plan to sustain this asset for a period up to 3 years after the project ends in collaboration and in agreement with the rest of the consortium.

Asset	**REST service for sensational content detection**
Type	Technology
Description	A REST service for annotating images and videos according to the existence of various sensational actions/events that are typically found in disinformation materials concerning e.g., war/conflict zones, refugees and migrants, or environmental disasters. The output of this service for a given image/video is the most relevant action/event from a predefined list of sensational actions/events, and a score in the range [0, 1], representing the similarity of the image/video to the identified action/event, with higher scores indicating greater similarity. This output can be used as evidence (in addition to the output of other data analysis technologies of AI4TRUST) for the check-worthiness of the news item or claim that is associated with the analysed image/video, and the prioritisation of its analysis with the different tools for trustworthiness/reliability evaluation.
Target Groups/Beneficiaries	Fact-checkers/journalists (to seek assistance in content verification). Media content owners/providers (for organising and retrieving their content; this extends beyond just sensational content detection, to any desired content that can be described in natural language), Researchers (for comparing their approach).
Innovation, Exploitation and Sustainability Plan	Differently from existing approaches for sensational content detection in images/videos (including the initial solution developed in AI4TRUST) that focus on specific categories of content, (e.g., visually disturbing, nudity, pornography), the developed solution is compatible with several categories of content that is typically found in disinformation materials. Moreover, grounded on a rich set of learned embeddings using image-text pairs from a wide variety of visual content, our

	technology can easily support the detection of newly defined actions/events on the fly; it is sufficient to describe the desired action/event in natural language. The developed service has been exposed through an API and is used as an integrated component of the AI4TRUST Platform, to annotate the collected visual data and provide an indication about their check worthiness to the Disinformation Warning System. The development of tools for annotating visual content are among the core research topics of our team, and thus we will be seeking additional funding from different sources (e.g., research grants, commercial contracts/licensing with Media content owners/providers) to ensure maintenance and improvement of the developed tool beyond the end of AI4TRUST. We plan to sustain this asset for a period up to 3 years after the project ends in collaboration and in agreement with the rest of the consortium.

Asset	**REST service for visual-text misalignment detection**
Type	Technology
Description	A REST service that gets as input an image/video and a text describing the visual content and makes an estimate about their contextual alignment. The output of this model for a pair of image/video and text is a score representing the misalignment between the image/video and the text (higher scores indicate greater misalignment), and a binary label “0” or “1” that specifies contextual alignment or misalignment, respectively. This output can be used as evidence (in addition to the output of other data analysis technologies of AI4TRUST) for the check-worthiness of the news item or claim that is associated with the analysed pair of image/video and text, and the prioritisation of its analysis with the different tools for trustworthiness/reliability evaluation.
Target Groups/Beneficiaries	Fact-checkers/journalists (to seek assistance in content verification). Media content owners/providers (for organising and retrieving their content; this extends to any desired content that can be described in natural language) Researchers (for comparing their approach).
Innovation, Exploitation and Sustainability Plan	The developed REST service integrates an AI model that has been trained using augmented training data. These data were

	generated by extending the VisualNews dataset that contains real pairs of images and captions, in order to also contain out-of-context pairs of images and captions, and for this we used an LLM. The developed service has been exposed through an API and is used as an integrated component of the AI4TRUST Platform, to annotate the collected pairs of visual and textual data and provide an indication about their check worthiness to the Disinformation Warning System. We envisage exploiting this asset as part of the AI4TRUST Platform and also seeking additional funding in relation to initiatives towards combating disinformation. We plan to sustain this asset for a period up to 3 years after the project ends in collaboration and in agreement with the rest of the consortium.

9.3. UNIVERSITÀ DEGLI STUDI DI TRENTO (UNITN)

The researchers of UNITN have a long-standing experience in the field of **computer vision and multimedia analysis** as well as competences in the social sciences. Specifically, UNITN has expertise on methods for: a) generating synthetic visual contents with deep generative models, b) visual content classification, d) video analysis and d) large multimodal models, e) digital sociology and sociological analysis of online communication dynamics. The collaboration between UNITN and the partners of AI4TRUST has the potential to not only enhance the performance of existing methods developed by UNITN but also pave the way for the **development of new and more effective approaches**.

UNITN aims to enhance its research expertise, focusing on image and video generation methods to support the development of more powerful tools for deep fake detection useful for fact-checkers and journalists in combating disinformation. Furthermore, UNITN is engaged in a multidisciplinary research effort that pivots around the tracing and innovative analysis of digital contents flow by mixing network analysis techniques with NLP and sociological theory. The outcomes of UNITN scientific knowledge and technological expertise will be shared through **high-impact scientific papers** that will be presented/submitted in national and international conferences/journals in multiple disciplinary areas spanning from computer and data sciences to sociology. Additionally, UNITN is **open to commercial opportunities**, considering licensing of developed tools, services, or providing consulting services to interested third parties.

UNITN envisions substantial advantages in participating in AI4TRUST by working on real-world cases and data provided by fact-checking and journalist partners. Additionally, the partnership with CERTH will allow the development of **more robust generative models**. Similarly, the collaboration with partners working on audio processing and Natural language processing (NLP) will allow to **improve the module of video anomaly detection with a multimodal pipeline**.

Asset	**Tool for image generation**
Type	Technology
Description	A technology that allows the generation of images. Given as input an image of a face and the indication of a certain attribute (e.g., colour of the hair) the software produces as output another image with the face and changed attributes.
Target Groups/Beneficiaries	Journalists and media (to create visual contents.) Researchers (for comparing their approach).
Innovation, Exploitation and Sustainability Plan	The AI4TRUST team will build on an existing model developed by UNITN for image generation and will target the further development of the software to handle arbitrary images - e.g., with different head poses, image clutter, focusing on face manipulation. The software will be made publicly available online (GitHub). The tool will allow partners to generate deepfake images to further develop deepfake detectors.

Asset	**Tool for video generation**
Type	Technology
Description	A software that permits generating short video clips. It takes as input a video and a segmentation mask which indicates the object in the video that needs to be modified and outputs another video where the indicated object is modified in appearance while maintaining motion.
Target Groups/Beneficiaries	Journalists and media (to create visual contents). Researchers (for comparing their approach).
Innovation, Exploitation and Sustainability Plan	Innovation, Exploitation & Sustainability Plan: In AI4TRUST we will build on a state-of-the-art text-to-image generation network available from the research community and will modify the network architecture with temporal layers and with conditioning input thanks to a ControlNet like architecture. Such modifications will improve the software, allowing the users to specify the object that needs to be edited with fine granularity. We plan to integrate other control signals, e.g., textual inputs, depth maps to further improve the software. The software will be made publicly available online (GitHub). The tool will allow partners to generate deepfake videos to further develop deepfake detectors.

Asset	**REST service for video anomaly detection**
Type	Technology
Description	A REST service that gets as input an image/video and estimates if the input contains a normal or an abnormal situation. The output of this model (a score in the range [0, 1], with 0 / 1 indicating normal/abnormal. The output of this model will be used within the Disinformation Warning System to estimate the check worthiness of a content.
Target Groups/Beneficiaries	Fact-checkers/journalists (to seek assistance in content verification). Researchers (for comparing their approach).
Innovation, Exploitation and Sustainability Plan	Innovation, Exploitation & Sustainability Plan: The developed solution has been exposed through an API and is used as an integrated component of the disinformation warning system within the AI4TRUST Platform. As the underlying technology is based on vision-language models and large language models, it is modular, and it can seamlessly include new advances in the fields. This ensures the sustainability of the tool in the long term. We expect to investigate other prompting strategies as well as frame selection tools to keep refining the efficiency and accuracy of the approach. We plan to sustain this asset for a period up to 3 years after the project ends in collaboration and in agreement with the rest of the consortium.

Asset	**Coordinated Inauthentic Behaviour Tool**
Type	Technology
Description	This functionality will rely on a re-elaboration of quantitative analysis of the reconstructed networks of social interaction that shall be technically possible on each social media platform mapped throughout the project, with a particular focus for development purposes on Telegram. The intended use is to allow end-users to explore how dis/misinformation circulates within and across platforms thanks to common languages/discourses. It should allow end-users to visualise some qualitative and quantitative indicators about the prominence of some contents within and across platforms.

Target Groups/Beneficiaries	Project partners and future users of the Platform (i.e., fact-checkers, media workers, policy makers)
Innovation, Exploitation and Sustainability Plan	This functionality exploits the results of the pre-processing phase via Social Network Analysis while rendering them visible and useful for end-users. The innovation in the simple indicators provided by the tool will rely on their unique computation from the combination of rich social media data, automatic labelling tools for disinformation detection, and social media analysis which is rarely combined with the two former. This component is as sustainable as the data and technologies are, but needs to be recomputed each time the data is updated.

9.4. NATIONAL CENTER FOR SCIENTIFIC RESEARCH "DEMOKRITOS" (NCSR-D)

NCSR-D and especially the Institute of Informatics and Telecommunications (IIT) consists of multiple researchers, Post-Docs, PhD, and MSc students, as well as practitioners which are active in the fields of N**atural Language Processing and machine/deep learning**. Lately, IIT has been focused on multidisciplinary research by involving partners from various fields beyond computer science majors, such as the disciplines encompassing **linguistics, law, and ethics**. Furthermore, NCSR-D has been a part of multiple relevant European projects, which provide experience in turning project and research outcomes into a competitive edge and opportunities for innovation. Furthermore, the results of **document intelligence** hold significant importance for the forward-looking digital innovation hub. This hub possesses expertise in advancing innovation outcomes and leveraging them independently. Moreover, it aims to amplify project innovations and contribute to the innovation hub sector, all while extending its reach to the broader research community.
In turn the experience, the expertise and multidisciplinary nature of the Institute allow for actions, which will drive innovation such as:

- Leveraging the latest advances and techniques in Natural Language Processing (NLP) and deep learning to develop document intelligence;
- Process and analyse textual datasets and extract valuable knowledge;
- Design and implement scalable and robust systems;
- Communicate and collaborate with other researchers, stakeholders, and users;
- Identify and address ethical and social issues and implications of research driven by development of document intelligence.

Asset	**Document intelligence - Technology**
Type	Technology
Description	Detecting disinformation is a complex task that often requires analysis beyond the surface level. An orchestrated approach to understand, clearly define, and detect these signals is essential. Document intelligence is an amalgamation of data-driven methods for detecting disinformation signals in textual content. The asset aims to deploy Natural Language Processing (NLP) methods to identify textual content produced for dis/mis/malinformation. Furthermore, it includes extensive analysis of disinformation signals. In turn, this analysis allows for a more verbose identification of disinformation. The verbosity of the asset enriches beneficiaries and stakeholders with informative results and permits deeper understanding of the disinformation agenda, while at the same time builds expertise and practical skills on dealing with such signals. From a technological point of view, document intelligence includes: (1) AI techniques that minimise human supervision by relying on a semi-supervised approach, complement existing tools with robust NLP and utilise state-of-the-art multilingual Large Language Models (LLMs) with prompt engineering and few-shot learning based on misinformative content for detecting disinformation signals of prevailing manipulation tactics as well as multilingual Pre-trained Language Models with cross-lingual fine-tuning for detecting hate speech, offensive language and clickbait cases. For cross-lingual fine-tuning, a mix of open-source datasets was leveraged. These advanced models benefit from their ability to understand and generate text across multiple languages, enabling more accurate and nuanced detection of disinformation in order to mitigate evolving challenges in document intelligence and moderate content. (2) Analysis, identification, and definition of the six most prevalent manipulation tactics, namely conspiracy theory, trolling, discredit, polarization, pseudoscience and science denialism with their respective dis/mis/malinformation signals, whose derivation requires deeper understanding of the content semantics that goes beyond superficial analysis, such as disinformation signal detection for verifying truthfulness, credibility, veracity, and authenticity — detection of bot/AI-generated text, etc.

Target Groups/Beneficiaries	Any platform or organisation that involves user-generated text, especially in domains with disinformation prone themes such as media outlets, researchers, fact-checkers, public health or climate change experts, policymakers, educators, etc.
Innovation, Exploitation and Sustainability Plan	The innovation outcomes of the asset mainly involve neural Natural Language Processing (NLP) models trained with up to date, fact-checked and unique data provided by the fact-checking organisations involved in the project's consortium. Besides, it includes any analytics or data products that may come up during training of these models, as well as expertise on dealing with disinformation signals. Furthermore, the fine-grained identification of disinformation signals, involvement of multilingual textual content and incorporating knowledge from domain experts of fact-checking organisations through carefully curated data instances and disinformation examples, as well as, through interactive consulting within the AI4TRUST project, allow for competitive advantage. As NCSR-D is a scientific research centre, the main direction for utilising the asset is mainly for academic exploitation. The focus of the exploitation would be to publish research outcomes in relevant journals and conferences with the intent to provide a competitive advantage and a specialised expertise to NCSR-D that will advance its collaborations with other institutes and further establish its position in relevant projects. Thus, enhancing its research reputation and strengthening its research in the relevant fields. Furthermore, the asset will be exploited within the AHEDD digital innovation hub either by providing the outcomes of the asset (e.g., trained neural models) within the services of the hub or providing expertise and know-how for the involved parties. In addition, it will present research results within its educational activities, thus providing fast access to research results to a new generation of researchers. Besides, the document intelligence tool can be exploited by other relevant digital European platforms such as the AI-on-Demand platform (AI4EUROPE, DeployAI) or the Greek AI Factory (Pharos). The plan to sustain the asset is to maintain it for a period up to 3 years after the project ends in collaboration and in agreement with the rest of the consortium. Adjustments and modifications are to be expected regarding the outcomes of the asset in terms of performance or including additional data resources to increase the generalisation and robustness of the trained models. Sustainability of document intelligence also depends on the maintenance of the rest of the tools, such as the AI4TRUST

	Platform. Adjustments in the integration of the asset within the Platform such as updating of connection protocols, maintaining the containerisation of the asset, or updating and maintaining technological dependencies are within the scope of the sustainability plan.

Asset	**Document intelligence - Knowledge**
Type	Knowledge
Description	Document intelligence includes expertise, practical skills, and knowledge of disinformation signals, such as understanding their features, translating this knowledge into building and training models to classify such signals. Such expertise ensures sustainability and effectiveness of combined methods over time, while they are encapsulated in the AI4TRUST Platform.
Target Groups/Beneficiaries	Any platform or organisation that involves user-generated text, especially in domains with disinformation prone themes such as media outlets, researchers, fact-checkers, public health or climate change experts, policymakers, educators, etc.
Innovation, Exploitation and Sustainability Plan	As NCSR-D is a scientific research centre, the main direction for utilising the asset is mainly for academic exploitation. The focus of the exploitation would be to publish research outcomes in relevant journals and conferences with the intent to provide a competitive advantage and a specialised expertise to NCSR-D that will advance its collaborations with other institutes and further establish its position in relevant projects. Thus, enhancing its research reputation and strengthening its research in the relevant fields. The exploitation of the technological asset within the AHEDD digital innovation hub will provide expertise and know-how for the involved parties. In addition, it will present research results within its educational activities, thus providing fast access to research results to a new generation of researchers.

9.5. CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS (CNRS)

The CNRS team involved in AI4TRUST has extensive experience researching social networks and their effects on organisations, markets, and political issues. Team members previously developed a **model of how actors use 'appropriateness judgments'** to give meaning to information and elaborate it interactively with their networks. Judgments depend on people's identification to reference groups, recognition of authorities, and alignment with priority norms. Adoption of mis/dis/malinformation should not be taken for granted and can rather be hypothesised to increase when judgments are similar and signalled as such in communication networks. AI4TRUST could build on these ideas and flag such signals to help users in their contextualization and interpretation of the phenomena described. The CNRS team also developed new research **mixing network and content analysis ('socio-semantic networks')** in response to findings on how network structures heavily affect the propagation of mis/dis/malinformation, regardless of the characteristics of individuals in these networks. Additionally, CNRS members have in-depth knowledge of cross-country, shadow market of paid engagements that contributes to injecting fabricated behavioural traces into social media and can influence the **speed and direction of propagation of false or misleading contents**.

Asset	**Expertise in social science, especially social network analysis**
Type	Knowledge
Description	CNRS brings to the project expertise in social science, with a distinctive focus on social network analysis
Target Groups/Beneficiaries	Researchers and policymakers
Innovation, Exploitation and Sustainability Plan	The CNRS team can provide insight into the motivations, constraints, and social pressures that lead actors, under uncertainty and embedded in their social networks, to interactively elaborate information and participate in the production and dissemination of harmful contents. Understanding the social dynamics that drive the production and diffusion of online information and contents is essential to identify forms of mis/dis/malinformation and to explain how some of them spread, while others do not. On this basis, CNRS can help to identify behavioural and social signals that, in addition to semantic signals, support the future AI4TRUST Platform in its effort to detect mis/dis/malinformation. CNRS can also provide suggestions on how to improve Platform design to

	account for both social and semantic signals and therefore, to improve its performance. The team's deep knowledge of the broad social media environment can offer insights into the political, economic and organisational factors that can impact observed outcomes as well as ethical issues specific to this type of setting. The large network datasets that can be accessed through the project can help the CNRS team to test and refine social network theories at a remarkable scale and level of detail. The expected outputs have potential to significantly enrich CNRS knowledge base and to stimulate further research in this area. Further, the collaboration with non-academic partners and the close contact with policymakers can provide valuable insights to pursue CNRS's social impact goals.

Asset	**Social Network Visualization Tool (formerly Social Network Analysis Tool)**
Type	Technology
Description	This Social Network Visualization tool will aim to visualize the distribution and dissemination of artifacts (such as URLs) or thematic elements (like embeddings) within a particular system, such as Telegram channels focused on specific topics. The tool will collect data by gathering artifacts from messages exchanged within specified Telegram channels and themes derived from content using methods like word embeddings or topic modelling. Visualizations will be created using an interactive mapping tool akin to Gephi, with nodes representing channels and edges indicating the relationships between them based on the frequency of information exchange. Spatialization of nodes will likely be determined by algorithms considering connection strength, often employing force-directed layouts to highlight clusters of closely related nodes. Different colours will be used to represent various themes or types of artifacts, with intensity possibly indicating frequency or importance within the network. Principles such as degree (measuring node connections), strength (measuring connection weight), and clustering (grouping nodes into communities) might ideally be employed to elucidate the important element of the network, possibly of its dynamics. Users will be able to interact with the map, zooming, panning, and accessing detailed information via tooltips.

Target Groups/Beneficiaries	Project partners and future users of the Platform (i.e., fact-checkers, media workers, policymakers)
Innovation, Exploitation and Sustainability Plan	The joint visualization of the network of interconnection across online channels and various semantic features related to disinformation would offer a novel tool that CNRS would rely on to analyse and showcase the connection between topological structure and misinformation dissemination. Such a progress would be further useful for CNRS broader understanding of joint social and semantic dynamics, for instance in terms of assessing whether there exist specific, assortative clusters of social groups which are feeding on the same type of information or not. This would also make it possible to hope for a comprehensive understanding of content dissemination dynamics on group-based platforms, such as Telegram.

Asset	**Reliability State of Social Media Tool**
Type	Technology
Description	This functionality will rely on a mostly quantitative analysis of the reconstructed network of social interactions on each social media. The intended use of this component is to provide the end-user with synthetical quantitative indices of the disinformation risk presented in given areas of social media that the Platform is able to map, to bring attention to areas of interest where the risk of disinformation flowing is higher. This simple tool will be beneficial to any category of end-user, as it will paint a straightforward picture of the reliability state of interactions under study.
Target Groups/Beneficiaries	Project partners and future users of the Platform (i.e., fact-checkers, media workers, policy makers)
Innovation, Exploitation and Sustainability Plan	The CNRS team, ideally using technologies provided in WP3 when applicable to the available data, combined with Social Network Analysis (SNA), can help provide synthetical measures of disinformation risk in areas of interaction between social media users, to assist the consortium and the end-user of the Platform in identifying less reliable areas. The innovation in these simple measures will rely on their unique computation from the combination of rich social media data, automatic labelling tools for disinformation detection, and social media analysis which is rarely combined with the two former. This

	component is as sustainable as the data and technologies are, but needs to be recomputed each time the data is updated.

Asset	**Relevance evaluator**
Type	Technology
Description	This functionality will establish content relevance based on three criteria: (1) whether its language is supported by the project (2) whether its author's virality is considered statistically important with respect to the virality of other (neighboring) authors in the social network (3) whether the post's virality is considered statistically important with respect to the virality of other (neighboring) posts in the social network/by the same author
Target Groups/Beneficiaries	Project partners and future users of the Platform (i.e., fact-checkers, media workers, policy makers)
Innovation, Exploitation and Sustainability Plan	The CNRS team, ideally using data and technologies provided in WP2 when applicable to the available data, combined with Social Network Analysis (SNA), can help provide synthetical measures of relevance and virality, to assist the consortium and the end-user of the Platform in identifying priority areas for various tasks including fact-checking and research. The innovation in this synthetic score will rely on its inclusion within a broader pipeline of content prioritization as developed in WP4. This component is as sustainable as the data and technologies are, but needs to be recomputed each time the data is updated.

9.6. POLITEHNICA BUCHAREST (POLITEHNICA)

The team at **POLITEHNICA** (formerly **UNIVERSITATEA POLITEHNICA DIN BUCUREȘTI - UPB**) possesses substantial experience in the field of **audio** and **speech processing**, with specific expertise in the following methods:

- **Speech-to-text transcription**;
- **Text-to-speech synthesis**;
- **Detection of audio deep fakes and AI-synthesised content**;
- **Emotion detection from speech**;
- **Speaker identification**.

POLITEHNICA's research experience and expertise, combined with the **technology requirements** from fact-checking and media organisations within the AI4TRUST consortium, will drive new advances in POLITEHNICA's existing solutions. This collaboration is expected to foster the development of new methods that could substantially improve the performance of POLITEHNICA's techniques and services.

Asset	**Speech-to-text technology and web service**
Type	Technology
Description	A technology that allows a user to transcribe automatically the spoken content of a multimedia file (audio or video that contains speech). The technology is deployed as a REST service that allows content upload, transcription progress assessment and transcription download. The technology is and will be adapted for transcribing news, press conferences and media content in general and is available for various European languages, such as RO, EN, ES, PL, DE, FR, and IT.
Target Groups/Beneficiaries	Media companies (to transcribe media content for writing news articles or reports); Fact checkers (to access faster the spoken content of long multimedia files).
Innovation, Exploitation and Sustainability Plan	POLITEHNICA entered into the project with a tool for speech to text in RO and EN, already in place in POLITEHNICA's workflows, and upgraded it as follows: a) updated the AI transcription models for RO and EN, b) added new AI models for other European languages (ES, PL, DE, IT, FR), and c) extended the API to support transcription of audio/video files based on their URLs. The tool will be further updated to include AI transcription models for EL. The tool is now an integrated component of the AI4TRUST Platform. Given that speech to text is a strategic research topic for our team, we will be seeking additional funding from different sources (e.g., research grants, commercial contracts, and licensing) to ensure maintenance and improvement of the service beyond the end of AI4TRUST. We plan to sustain this asset for a period up to 3 years after the project ends in collaboration and in agreement with the rest of the consortium.

Asset	**Audio deep fake detection technology and web service**
Type	Technology
Description	A technology that allows a user to assess whether the audio stream in a multimedia file (audio or video that contains speech) is computer generated or manipulated by cutting and pasting various real or computer-generated content.
Target Groups/Beneficiaries	Fact-checkers/journalists (to assess content veracity).
Innovation, Exploitation and Sustainability Plan	POLITEHNICA developed this technology and the related web service from scratch, as follows. A first version of the technology was developed using pretrained, self-supervised speech representations and basic classifiers. The first version of the technology was trained only on scientific datasets. This technology was presented at the top conference on speech processing and was evaluated to be above state-of-the-art on 6 scientific datasets. The robustness of the technology was addressed by further fine-tuning and evaluating the first version on a large dataset of audio deep fakes collected by the consortium. Coupled with architectural updates and an innovative data-centric training procedure, this led to an audio deep fake detection technology evaluated to be above the state-of-the-art on 5 scientific datasets and 2 realistic datasets. Explainability was obtained by accompanying the detection decision with information regarding the AI model used to generate the deep fake The technology and web service are already integrated in the AI4TRUST Platform. Given that deepfake detection became a strategic research topic for our team, we will be seeking additional funding from different sources (e.g., research grants, commercial contracts, and licensing) to ensure maintenance and improvement of the service beyond the end of AI4TRUST. We plan to sustain this asset for a period up to 3 years after the project ends in collaboration and in agreement with the rest of the consortium.

Asset	**Audio anomaly detection technology and web service**
Type	Technology

Description	A technology that allows a user to assess whether the audio stream in a multimedia file (audio or video that contains speech) contains anomalies, such as splicing points.
Target Groups/Beneficiaries	Fact-checkers/journalists (to assess content veracity).
Innovation, Exploitation and Sustainability Plan	POLITEHNICA developed this technology and the service from scratch, as follows. The first version of the technology used pretrained self-supervised speech representations and an audio segment classifier to detect whether an audio segment contains partly AI generated content or splicing points. POLITEHNICA now develops a splicing point detection technology that will be further integrated into a web service and exposed to the end users through AI4TRUST Platform. The technology and web service will be used stand-alone and integrated in the AI4TRUST Platform. They will be further used by POLITEHNICA in other research projects.

Asset	**Audio deep fake generation technology**
Type	Technology
Description	A technology that allows a user to input a text message, select the voice and generate the spoken message corresponding to the input text (text-to-speech technology).
Target Groups/Beneficiaries	AI4TRUST research team (to generate audio deepfakes to further improve the detection technology); Media companies (to create spoken news starting from text news).
Innovation, Exploitation and Sustainability Plan	POLITEHNICA entered the project with a text-to-speech tool for RO and EN. This tool was further improved within the project by updating various architectural components, e.g., the vocoder, and by addressing the adaptation capabilities to new voices using as little training data as possible. This tool will be used as a stand-alone technology in POLITEHNICA to create deepfakes for further improving the detection technology. Given that text-to-speech is a strategic research topic for our team, we will be seeking additional funding from different sources (e.g., research grants, commercial contracts, and licensing) to ensure maintenance and improvement of the service beyond the end of AI4TRUST. We plan to sustain this asset for a period up to 3

	years after the project ends in collaboration and in agreement with the rest of the consortium

9.7. SAHER (EUROPE)

As a company, SAHER (EUROPE) is highly experienced and skilled in **creating and exploiting its professional networks to maximise project impacts**. The knowledge that SAHER gains throughout the project is both used as a foundation for its commercial activity and for using it as a Platform for future collaborative work. Furthermore, SAHER with its renowned expertise in **ethical and legal aspects** related to the development and implementation of AI-based technologies is a key partner for the **release of trustworthy and EU-GDPR compliant solutions**.

Asset	**Networking through stakeholders**
Type	Knowledge
Description	Thanks to its expertise, SAHER will undertake key action to disseminate AI4TRUST outcomes to maximise the impact and reach out future academic, research and industrial partners in the project. This will involve attending domain-specific events and gatherings and seeking opportunities to promote the AI4TRUST network and the results generated by it. This networking will be undertaken by: ● individual discussion and negotiations with pre-selected partners both face to face and virtually; ● cooperation on research projects, by inputting to other networks and seeking to create new networks that will add value; ● pro-active networking to attract new partners and commercial opportunities to exploit based on the knowledge gained by participating in the AI4TRUST network and by contributing to platforms, EU initiatives and alongside activities to exploit the network through writing publications and contributions using online and new media opportunities such as LinkedIn posts.
Target Groups/Beneficiaries	Researchers and Professionals
Innovation, Exploitation and Sustainability Plan	The dissemination activity will be based on to three specific short-term goals: ● to build a network of potential academic partners (European universities and respected research institutions

	and think tanks). The potential academic partners will, in the first phase, be selected according to the strongest performing fields of research within each institute. This will be part of our knowledge transfer and knowledge exchange activities with the company; • to expand its established network of potential industrial partners; • to create a network of potential research partners (including end-users). The empowerment of the communication in different networks and related activities will sustain and support the exploitation of the AI4TRUST results in a broad range of domains and stakeholders.

Asset	**Extensive expertise in ethical and legal aspects**
Type	Knowledge
Description	The SAHER(EUROPE) strategy for implementing security innovations is reinforced by incorporating internationally recognised research legal and ethics expertise, as well as extensive security policy review and evaluation at both national and international levels.
Target Groups/Beneficiaries	Researchers and Professionals
Innovation, Exploitation and Sustainability Plan	The Legal & Ethics Team of SAHER comprises internationally recognised professionals who contribute a wealth of expertise to guide and advise on complex research and innovation programs. With a deep understanding of developing legal and ethical frameworks for international research, the team possesses unique proficiency in areas such as data protection, privacy, human rights, and the ethical considerations surrounding biometrics, digital governance, facial recognition, big data analytics, and the utilisation of AI. Emphasising and empowering the ethical and legal aspects will facilitate the exploitation of AI4TRUST outcomes across various domains and stakeholders in a manner that fully complies with the EU-GDPR and aligns with the requirements of the new AI Act and DSA and observe the aspirations of the AI Pact.

9.8. GDI GLOBAL DISINFORMATION INDEX GUGHAFTUNGSBESCHRANKT (GDI)

GDI has in-depth knowledge in the development of **technology which can automatically detect signs of online disinformation**. Also, GDI's team has subject matter expertise on disinformation **in various geographical contexts**. GDI leverages recent advances in **Natural Language Processing (NLP)** to detect domains which are at risk of sharing disinformation in various languages. The list of domains included in **GDI's data platform** relies on a review process performed by a team of trained intelligence analysts. The Manual Review is run across sites identified by its machine learning classifiers as carrying the highest potential disinformation risk. Websites are analysed to identify the presence of adversarial narratives[73] tracked by GDI. Each website is reviewed by a minimum of two intelligence analysts who perform a "blind" review meaning that they do not see each other's rating.

Asset	**Disinformation warning system (DWS)**
Type	Technology
Description	As part of Work Package 3- "AI-driven data analysis methods", GDI is going to develop a new model alongside other partners of Task 3.4. This new model will aim to detect whether a piece of content is likely to contain disinformation or not, also providing a confidence score.
Target Groups/Beneficiaries	Professional and Researchers
Innovation, Exploitation and Sustainability Plan	This model will be specifically designed by GDI and other partners involved in T3.4 to integrate outputs from technologies developed in WP3, as well as GDI's data platform. This model will be developed to respond to the specific needs of the AI4TRUST consortium, in terms of language and type of content reviewed by the tool (text, image, audio, multimodal). The timeline for this asset is tied to the delivery of technologies developed in T3.1-T3.3 given the disinformation warning system will integrate the outputs of these technologies. This asset will serve the consortium's work and help in the detection of disinformation. This asset is specially designed to meet the needs of the consortium and will be tailored to integrate technologies from T3.1-T3.3. In that regard, GDI does

[73] https://www.disinformationindex.org/blog/2022-06-22-disinformation-as-adversarial-narrative-conflict/

	not plan to exploit this asset outside of AI4TRUST consortium work.

Asset	**Data platform for indexing and documenting verified and manipulated content**
Type	Technology
Description	GDI will use its existing data platform for indexing and documenting verified and manipulated content, which will constitute a feature called Domain Disinformation Detection in the disinformation warning system. This data platform is the result of pre-existing work carried out by GDI, including the identification and labelling of domains which regularly share disinformation across various languages of the AI4TRUST project.
Target Groups/Beneficiaries	Professional and Researchers
Innovation, Exploitation and Sustainability Plan	The database used by GDI will be made up of domains that have been previously labelled by GDI as at a high risk of spreading disinformation in the languages used in the AI4TRUST project. The identification of these domains is the combined result of GDI's technology capabilities, and expertise of GDI's intelligence team. Each domain within this dataset has been manually reviewed by GDI's analysts. The exploitation of this asset in the AI4TRUST project will be completed by GDI during the development of the model of the disinformation warning system. The creation and exploitation of this asset by GDI pre-exists and is independent of the AI4TRUST project. In that regard, this asset will continue to be exploited by GDI for research/ commercial/ educational purposes during and after the AI4TRUST project.

9.9. STOWARZYSZENIE DEMAGOG (DMGG)

Demagog Association (DMGG) has a nuanced knowledge, experience, and understanding of the challenges that surround the process of verifying information on social media. Demagog is the first **fact-checking organisation** in Poland. For ten years, the organisation has been verifying politicians' statements. In 2019, DMGG joined the Third-Party Meta program, and in 2023, it began collaborating with TikTok. The association also belongs to the International Fact-Checking Network (since 2019) and the European Fact-Checking Standards Network (DMGG was heavily involved in

the creation of the entire organisation, moreover, Pawel Terpilowski, editor-in-chief of the portal sited on the EFCSN board).

The first resource DMGG brings to the project is a **strong expertise in fact-checking and recognising misinformation, disinformation, malinformation, and conspiracy theories**. The organisation specialises in recognising disinformation on health, war, and climate, targeting minority groups, technology & AI, conspiracy theories and hoaxes. Also, DMGG tracks actors who financially gain from spreading disinformation. The organisation has experience in obtaining information from open sources and expertise in the necessary technological implementations to reduce the spread of false information and information operations. These resources are fundamental for the AI4TRUST project. Moreover, DMGG will share **ground-truth data** with the Consortium. This data is a key asset for the project as it will be used **to train AI models** that will be integrated into the AI4TRUST Platform. The data will be shared with the Consortium only for the project's lifecycle and objectives, while any eventual further commercial use will be **agreed upon in terms of legitimate interests and protection of intellectual value**.

Asset	**Fact-checking analyses data of misleading information published on social media**
Type	Data
Description	Demagog Association (DMGG) provides a database containing fact-checking analyses of misleading information published on social media, including fake content generated by artificial intelligence. The database is essential for training the models used in the AI4TRUST project to establish a Platform to limit false information in the digital space. DMGG has verified more than 2,200 fake news stories in the social media and digital space during its activities. This data is going to be collected and shared as part of the AI4TRUST project to train the AI model that will power the Platform. Precise and transparent methodology, reliance on primary sources, correction policies, and financial transparency that are validated by the IFCN and EFCSN ensure that the shared database is representative and reliable.
Target Groups/Beneficiaries	Technical partners and Researchers
Innovation, Exploitation and Sustainability Plan	The data provided by DMGG is a valuable resource to develop more accurate and efficient fact-checking/ disinformation identification tools using artificial intelligence. By M14, this data will be available to the consortium, while DMGG plans to

	continually update and expand these datasets to ensure their continued relevance and usefulness, even after the end of the project, to guarantee sustainability of the overall results.

Asset	**Fact-checking and detecting misinformation, disinformation, and malinformation on social media**
Type	Knowledge
Description	DMGG shares critical insights and expertise in terms of fact-checking and identifying misinformation, disinformation, malinformation, and conspiracy theories. DMGG's assistance translates into improving the effectiveness of the tools implemented by technology partners, ensuring that the work results in a Platform that will be a useful, practical, and an effective tool in the fight against disinformation.
Target Groups/Beneficiaries	Technical partners of the project and developers of the Platform.
Innovation, Exploitation and Sustainability Plan	Similarly to other fact-checking organisations involved in AI4TRUST, DMGG's expertise plays a key role in driving innovation within the project. DMGG is committed to ongoing collaboration with technology partners to refine and improve AI tools and approaches, ensuring their relevance and applicability. Once the project is completed, there is an opportunity to continue DMGG's involvement, contributing to the adaptation and enhancement of the Platform in future evolutions.

Asset	**Fact-checkers everyday workflow**
Type	Process
Description	DMGG is involved in the process of testing the AI4TRUST Platform pilot. For this purpose, the process of using the Platform by fact-checkers and media professionals in their daily work will be designed, along with reporting on the level of satisfaction and usefulness of the pilot version of the AI4TRUST Platform.
Target Groups/Beneficiaries	Project partners and future users of the Platform (i.e., fact-checkers, media workers, policymakers)

Innovation, Exploitation and Sustainability Plan	The pilot testing is a key phase to the success of the project, providing immediate and practical feedback to technical partners. This information will be used to refine and optimise the AI4TRUST Platform. After the end of the project, further use and feedback will contribute to the Platform's long-term improvement and adaptation to the dynamic and changing needs of fact-checking.

9.10. FUNDACIÓN MALDITA.ES CONTRA LA DESINFORMACIÓN: PERIODISMO EDUCACIÓN INVESTIGACIÓN Y DATOS EN NUEVOS FORMATOS (MALDITA)

MALDITA brings to the table a wealth of experience and expertise in the field of **mis/disinformation detection and fact-checking**. Its team comprises skilled journalists, fact-checkers, and technology product and development experts who have a proven track record in combating dis/mis/malinformation and promoting media literacy. MALDITA's expertise aligns seamlessly with the assets outlined, particularly the integration of AI4TRUST Platform AI tools into our fact-checking workflow. MALDITA has a comprehensive understanding of the complexities and nuances of the disinformation landscape.

MALDITA's experience in debunking and analysing mis/disinformation patterns positions us as a valuable partner in fine-tuning AI solutions to effectively detect and combat mis/disinformation. We **bring firsthand knowledge of the challenges faced by fact-checkers**, allowing us to provide insights and feedback that are essential for the development and optimisation of AI tools. We also have the capacity to understand and solve technical challenges around disinformation. Leveraging its editorial team's expertise, MALDITA will actively **contribute to integrating AI4TRUST Platform AI tools into its news production processes**. MALDITA's journalists possess a deep understanding of reporting standards and journalistic ethics, ensuring that the advanced tools are seamlessly **integrated into MALDITA's workflows while upholding accuracy and integrity**.

MALDITA is committed to driving innovation in the fight against mis/disinformation while upholding the highest **standards of quality journalism**. MALDITA's dedication to accuracy, transparency, and accountability ensures that the integration of AI solutions into MALDITA's fact-checking processes is conducted with the utmost rigour and integrity. MALDITA actively seeks to **innovate and improve its practices**, continually refining its approaches to better serve its audience and combat dis/mis/malinformation effectively.

As a fact-checking partner, **MALDITA is well-positioned to deliver significant benefits to the consortium:**

- **Real-world scenarios and practical use cases:** MALDITA brings firsthand experience in combating mis/disinformation in real-world contexts. By providing access to **MALDITA's extensive database of mis/disinformation cases and fact-checking methodologies**, MALDITA offers valuable insights into the practical challenges and opportunities associated with disinformation detection and debunking. These real-world scenarios serve as essential test cases for evaluating the effectiveness and reliability of the AI4TRUST outcomes in real-world applications.
- **Validation and testing:** as a trusted fact-checking organisation, MALDITA provides **credibility and validation to the AI4TRUST outcomes**. MALDITA offers its platform and resources for testing and validation, ensuring that the AI tools effectively address the needs and requirements of fact-checkers. By rigorously **testing the AI4TRUST solutions in a real-world environment**, MALDITA contributes to enhancing their **accuracy, reliability, and usability**, ultimately improving their **value proposition** for the Consortium.
- **Shared interest in combating mis/disinformation:** MALDITA shares a common interest with other project partners in combating disinformation and promoting **media integrity**. By collaborating closely with academic and technology partners within the Consortium, MALDITA contributes to a **holistic approach to mis/disinformation mitigation**, leveraging its respective expertise and resources to achieve shared objectives. Together, MALDITA aims to develop **innovative solutions and methodologies** that effectively address the challenges posed by disinformation in today's digital landscape.

Asset	**Data used for ground truth**
Type	Data
Description	Maldita will share a dataset to contribute to the effort of Work Package 2-" Methodological design, data gathering and pre-processing" to collect ground truth data. This dataset will be shared with consortium partners.
Target Groups/Beneficiaries	Professionals and Researchers
Innovation, Exploitation and Sustainability Plan	The dataset provided is the result of the expertise of Maldita's team. To constitute this dataset, Maldita's team has included reports of contents with possible disinformation from users on WhatsApp and other social media such as Twitter, Facebook, Instagram, and Telegram.

	This dataset will be used by consortium partners, especially technical partners developing technology components in Work Package 3-"AI-driven data analysis methods". The timeline for the use will be regulated by a consortium's data sharing agreement. The creation and exploitation of this asset by Maldita pre-exists and is independent of the AI4TRUST project. In that regard, this asset will continue to be exploited by Maldita for research/ commercial/ educational purposes during and after the AI4TRUST project.

Asset	**Integration of AI4TRUST tools to Maldita's fact-checking workflow**
Type	Technology
Description	The AI4TRUST AI tool will be integrated to Maldita's fact-checking workflow to test and validate the AI4TRUST models and technologies on an existing, in production and every day used disinformation management framework and related data.
Target Groups/Beneficiaries	Maldita's newsroom team
Innovation, Exploitation and Sustainability Plan	The combination of AI4TRUST Platform AI tools with Maldita's fact-checking workflow has the possibility to represent a significant enhancement to our fact-checking capabilities. By leveraging the AI capabilities offered by AI4TRUST, we aim to improve several aspects of the fact-checking process, thereby improving the efficiency and effectiveness of our operations. The integration of AI4TRUST tools in daily activities will lead to a more efficient fact-checking workflow for our newsroom team. For instance, we anticipate significant time savings for our fact-checkers, enabling them to focus more of their time and energy on higher-value tasks, such as verifying the accuracy of information and producing high-quality debunk articles. One of the primary benefits of streamlining our fact-checking workflow is the ability to respond more quickly to user inquiries and reports of disinformation. By reducing the time, it takes to identify and debunk false information, we can improve our overall response time and provide more timely and accurate information to our audience. This will enhance trust in our

	organisation and increase engagement with our fact-checking services. Moreover, we aim at being better positioned to scale our fact-checking operations to handle larger volumes of dis/mis/malinformation. This scalability is essential for meeting the growing demand for fact-checking services and ensuring that we can effectively combat disinformation across various platforms and channels. Additionally, we are laying the foundation for long-term sustainability and continued innovation in our fact-checking efforts.

9.11. ASTIKI MI KERDOSKOPIKI ETAIRIA KENTRO KATAPOLEMISIS TIS PARAPLIROFORISIS / CIVIL NON-PROFIT COMPANY KENTRO KATAPOLEMISIS TIS PARAPLIROFORISIS (ELLINIKA)

ELLINIKA's expertise in fact-checking is deeply intertwined with each of the assets of the project. With years of experience in **identifying, analysing, and debunking misleading information**, ELLINIKA brings a nuanced understanding of the challenges and requirements in this field. This expertise is vital in shaping the development and application of the AI4TRUST Platform.

With an almost 10-year active presence in Greece, ELLINIKA has fact-checked numerous misleading claims and has produced more than 2,000 debunking articles. These articles constitute our **database**, which has now been shared with the Consortium to be used for the development of the AI4TRUST Platform. Our high standards, aligned with the international fact-checking standards set by the International Fact-Checking Network and the European Code of Standards for Independent Fact-Checking Organisations—both of which we are signatories to—ensure that our data are not only extensive but also accurately representative. Below are the main benefits ELLINIKA brings to the project:

- **Expertise in fact-checking:** Our experience and skills in the fact-checking field are directly transferred to the development of the AI4TRUST Platform. We have already and will continue to provide critical insights and feedback to the tech partners, ensuring that the tools developed are not only technologically advanced but also practically applicable and effective.
- **Ground Truth data and testing :** ELLINIKA's expertise extends to the practical application of the Platform, where we critically assess its performance in real-world scenarios. Our experience allows us to provide valuable feedback on the Platform's functionality and usability, directly influencing its refinement and optimisation. By providing our database

and engaging in the pilot testing of the Platform, ELLINIKA brings significant benefits to the consortium and aligns closely with the project's goals. Our data, which include all our fact-checking articles and cases containing AI-generated disinformation, serve as a valuable resource for all partners involved. At the same time ELLINIKA's commitment alongside other fact-checking organisations to pilot test the Platform is of great importance for the project as providing direct, actionable feedback on the Platform's efficiency is crucial for the Platform's iterative improvement. This real-life testing of the Platform ensures that the tools developed are not only technologically advanced but also practically applicable and effective in real-world fact-checking scenarios.

Asset	**Data related to fact-checking articles and dataset with generated AI content**
Type	Data
Description	In collaboration with other fact-checking partners in the project, ELLINIKA provides essential data, including all our fact-checking articles (around 2.000) and datasets containing AI-generated content used to spread mis/disinformation. This data is crucial for training the AI tools of the AI4TRUST Platform.
Target Groups/Beneficiaries	Technical partners and academic researchers in the field of dis/mis/malinformation
Innovation, Exploitation and Sustainability Plan	These datasets represent a valuable resource for developing more accurate and efficient AI tools. These data have already been shared with the consortium, and we plan to continually update and expand these datasets to ensure their ongoing relevance and usefulness.

Asset	**Fact-checking background and expertise**
Type	Knowledge
Description	ELLINIKA contributes expert knowledge and experience in the field of fact-checking. This expertise is essential to guide technical partners in the development of AI tools and test the Platform, ensuring they are accurately aligned with the typical, everyday requirements of fact-checking operations.
Target Groups/Beneficiaries	Project consortium and future users of the Platform.

Innovation, Exploitation and Sustainability Plan	As with all fact-checking organisations involved in the project, ELLINIKA's expertise plays a critical role in driving innovation within the project. We are committed to ongoing collaboration with tech partners to refine and enhance AI tools and approaches, ensuring their relevance and applicability. Post-project, we plan to continue our engagement, contributing to the Platform's adaptation and improvement if needed.

9.12. EURACTIV MEDIA B.V. (EURACTIV)

Euractiv's expertise derives from the 20 years of experience in **EU policy news** and the wide range of journalists with varying years of experience (junior and senior journalists who have different workflows and priorities). The skills offered by Euractiv will support the development of the AI tool every step of the way - providing **knowledge on journalists' workflows, testing pilot versions and ultimately the final product**. Euractiv offers **journalists in English, French and German** – aligning with the **multilingual angle of the project**.

Euractiv's involvement is a real need both internally and externally from the project. Internally, its expertise is required to **properly test the Platform** before it is made public. By sharing the tool with its journalists, Euractiv was also able to share its initial analysis of comparing how efficient the tool is versus what journalists use currently. Externally, **Euractiv's journalists will continue to combat disinformation** and increase the public's trust in news organisations by generating trustworthy content both during and after the project has finished. **Euractiv will disseminate the results of the project through its channels** (i.e., website, social media, events) through editorial and multimedia content. Euractiv added Bluesky as one of its dissemination and exploitation channels due to the rise in concern of X (Twitter). This will increase the reach of AI4TRUST, complimenting Euractiv's communication work which focuses on sharing general information on Artificial Intelligence. Despite Euractiv Media Network merging with another organisation, resulting in changing its domain name to Euractiv Media, the internal workings of its dissemination of content and the day-to-day structure of its journalists have not changed.

Asset	**Expertise in journalism**
Type	Knowledge
Description	The Euractiv Media's newsroom hosts a wide range of expertise in Brussels, Paris, and Berlin. The journalists will offer their knowledge in what is required to create news content; testing the efficacy and efficiency of the AI4TRUST fact checking tools.

Target Groups/Beneficiaries	Professionals, Journalists, EU Policymakers (as readers of Euractiv)
Innovation, Exploitation and Sustainability Plan	In a business perspective, Euractiv will benefit from the AI4TRUST Platform in the long term as its journalists will be able to use it to further enhance the public's trust in public institutions. By testing the Platform, Euractiv's skills will continue to develop in creating trustworthy news. The revenue structure of Euractiv has changed by the addition of a paywall and pro subscriptions to content. However, this does not impact the work of AI4TRUST, as all content created and disseminated for the project is made freely available.

Asset	**Access to content generated by Euractiv**
Type	Data
Description	Euractiv Media is an independent pan-European media organisation, with expertise across Europe. The newsroom covers 8 different hubs, ranging from Artificial Intelligence to Climate Change and European Politics. AI4TRUST and Fact-checkers will have access to content generated by Euractiv on these wide-ranging topics, as well as multi-faceted formats including editorial, video, infographics and podcasts. The paywall mentioned above does not impact the access that the AI4TRUST consortium has with regards to AI4TRUST content.
Target Groups/Beneficiaries	Professionals and Researchers, Institutions, EU policymakers, Consultancies.
Innovation, Exploitation and Sustainability Plan	Euractiv does not plan to exploit this asset outside of AI4TRUST consortium work.

9.13. SKYTG24

Sky TG24 offers comprehensive daily news coverage, addressing both domestic and international events, across television, its website, and social media platforms. With over 7,000 hours of live content each year, accompanied by in-depth reporting, **Sky TG24** provides a wealth of expertise in the field of disinformation countermeasures. Over the years, it has developed significant experience

with TV programs such as *"Numeri"*, *"Pillole di vaccino"*, and *"Impact"*, which focus on spreading high-quality information and promoting data literacy on critical issues such as economic and political topics, vaccination campaigns, and climate change, all aimed at tackling dis/mis/malinformation (as detailed in D6.1, section 2.3.1).

Sky TG24 brings valuable expertise in analysing and testing the **AI4TRUST** tools from the perspective of a media company and journalists. With daily journalistic activity and experience navigating the complexities of the modern news landscape, **Sky TG24** offers a broad viewpoint on the critical issues faced by media companies in their efforts to combat dis/mis/malinformation.

The **individual exploitation plans** of **Sky TG24** could place particular emphasis on leveraging its extensive journalism expertise and commitment to disseminating high-quality information, thereby enhancing the effectiveness of the **AI4TRUST** Platform in combating mis/disinformation. Given **Sky TG24**'s daily coverage of news and its experience with programmes designed to promote data literacy and counter disinformation, the company is well-positioned to play a pivotal role in testing and analysing the **AI4TRUST** tools from a media perspective.

Sky TG24 could utilise its deep understanding of the intricacies of the news landscape to provide valuable insights into the specific needs of journalists and media companies in relation to fact-checking and mis/disinformation detection. By collaborating with the **AI4TRUST** project, **Sky TG24** could assess the Platform's capacity to meet these needs and recommend modifications that would enhance its usability and functionality for media professionals.

Moreover, **Sky TG24's** involvement in the **AI4TRUST** initiative can facilitate knowledge exchange among media companies, academic institutions, and technology developers. This collaboration will enable a more thorough understanding of the challenges in the fight against disinformation and foster the development of innovative, tailored solutions for the journalism sector. By working closely with other stakeholders, **Sky TG24** can further enhance its own capacity to address mis/disinformation, ultimately benefiting both its operations and the wider media ecosystem.

To ensure continuous engagement and improvement, **Sky TG24** could interact with journalists and media companies through workshops and webinars, allowing for direct feedback and input. This approach could help to tailor the company's offerings to better meet the needs of users while fostering a sense of community within the industry.

In summary, **Sky TG24's** individual exploitation would involve utilising its journalism expertise to inform the development and enhancement of the **AI4TRUST** Platform, fostering collaboration with other media and academic partners, and contributing to the collective effort to combat disinformation effectively. This strategy not only strengthens **Sky TG24's** position as a leader in quality journalism but also amplifies the overall impact of the **AI4TRUST** solution in the fight against mis/disinformation.

Asset	**Expertise in journalism field**
Type	Knowledge
Description	Sky TG24 can provide its knowledge of day-to-day news coverage and expertise in covering all relevant news: this can help to analyse if/to what extent the Platform under development can meet the needs of journalists and media companies. Sky TG24 can also provide suggestions - if helpful - on how to change the Platform to make it more useful for journalists and media companies. Sky TG24 knowledge of the media ecosystem can also offer useful insights on the journalism field to inform all the phases of the project.
Target Groups/Beneficiaries	Journalists and Media Companies
Innovation, Exploitation and Sustainability Plan	Working alongside other media companies can help Sky TG24 better understand the needs of the sector, exchange valuable knowledge and expertise, and discuss different points of views on the best ways to tackle disinformation. Moreover, working alongside both industrial and academic partners allows us to build connections between studies on social media, journalism and technological research and development. Working with academic and industrial partners can broaden our view on different approaches and needs, further augmenting Sky TG24 abilities to understand the reality around us. Sky TG24 can provide to consortium partners its expertise in the field of journalism, allowing them to better understand what the needs of a media company in tackling dis/mis/malinformation are online.

9.14. ASOCIATIA DIGITAL BRIDGE (ADB)

Facilitating fact-checking and debunking in Romanian newsrooms is essential, given the rarity and inconsistency of such activities within the country. An AI-driven solution has the potential to significantly accelerate the development of fact-checking efforts. However, larger newsrooms often face limitations in terms of flexibility, making it difficult to experiment with various platforms and tools. From this perspective, not only **the tools provided by the AI4TRUST Platform**, but also the **Disinformation Warning System (DWS)**, would play a **crucial role in the work of both journalists and researchers.**

Individual exploitation plans have evolved since the shutdown of **Crowdtangle** in August 2024, as our association is no longer solely a media outlet but has increasingly focused on research activities. Consequently, we are now particularly interested in utilising **DWS** instead of **Crowdtangle**. **DWS** has the potential to **distinguish the AI4TRUST Platform from other fact-checking tools**, as identifying the most widely distributed disinformation pieces is the first critical step in any fact-checking and debunking effort.

ADB Euractiv Romania, structured as an NGO with a media outlet's specialisation, has expanded its research capabilities. The organisation possesses the expertise and resources to **integrate the AI4TRUST Platform into the piloting of AI-assisted fact-checking and DWS**, particularly in the area of EU-related topics and European policies. This approach will enhance the effectiveness of fact-checking in Romanian media, ultimately contributing to the broader fight against disinformation, particularly in the context of EU policy and governance.

Asset	**Expertise in Journalism**
Type	Knowledge
Description	ADB - Euractiv Romania boasts a small yet experienced team of journalists capable of thoroughly testing the Platform and offering feedback on both technical responsiveness and the quality of AI-generated results. The primary market focus remains Romanian newsrooms, fact-checking organisations, and researchers that require AI-powered verification tools. However, the secondary market now includes EU-focused policy journalists, academic institutions, and NGOs working on media and/or digital literacy.
Target Groups/Beneficiaries	Journalists, Media organisations and Academic Institutions
Innovation, Exploitation and Sustainability Plan	The team of journalists involved possess the expertise and flexibility necessary for engaging in projects and testing different stages of the Platform aimed at enhancing the quality of journalism and streamlining the journalistic workflow in the Romanian language newsroom. The testing timeline will be aligned with the project's requirements. Potential customers, including journalists, students in journalism, media companies and academic institutions, will be engaged through workshops and training sessions on AI-driven journalism and partnerships with newsrooms to integrate the AI4TRUST Platform into daily workflows. Once a piece of information is checked, ADB -

	Euractiv Romania will publish the respective item in Romanian language in the section Facts, not Fake, on euractiv.ro.

Asset	**Partnership with key media organisations and fact-checking NGOs / companies**
Type	Knowledge
Description	The core team of journalists can collaborate with ADB - Euractiv Romania's media partners, including newsrooms, fact-checking NGOs, or academics specialises in debunking, to expand the Platform testing validate the AI's ability to detect and humanly fact-check disinformation spread in the Romanian language, and the results turned by the DWS ADB - Euractiv Romania can offer feedback and suggestions based on their own testing as well as any testing extended to partners.
Target Groups/Beneficiaries	Fact-checkers, Journalists
Innovation, Exploitation and Sustainability Plan	ADB - Euractiv Romania aims to use the AI4TRUST outcomes and gains expertise to enlarge and differentiate its network.

Asset	**Native knowledge of Romanian language**
Type	Knowledge
Description	ADB - Euractiv Romania foresees the possibility to use the AI4TRUST results to improve the Romanian language's transcriptions tools.
Target Groups/Beneficiaries	Researchers and Professionals
Innovation, Exploitation and Sustainability Plan	The ADB - Euractiv Romania team of journalists has already begun using the speech-to-text app, which is a component integrated with the AI4TRUST Platform for the Romanian language. While utilising the app for transcriptions, they also make corrections within the app for the returned results in Romanian.

9.15. EUROPEJSKIE MEDIA SP ZOO (EMS)

As a media organisation, **EMS** delivers benefits to the Consortium by providing **real-world scenarios**, **practical use cases**, and access to its **extensive network** for validation and testing. The EMS organisation possesses a strong foundation in **EU news reporting, policy analysis, and debunking false claims**. In addition to its editorial expertise, EMS will provide its dataset of debunked fake news and actively engage in the review and assessment of claims to support collaboration with technical teams and researchers.

The editorial team's expertise aligns seamlessly with the AI4TRUST outcomes, ensuring a deep understanding of the disinformation challenges in Poland. Leveraging its experience, EMS actively **contributes to the development and fine-tuning of the AI4TRUST AI solutions, integrating them effectively into its news production processes**. EMS commitment to accuracy and journalistic integrity will drive the innovation of project outcomes, ensuring that the advanced tools **not only detect disinformation but also enhance the overall quality and reliability of its news content**. This synergy between EMS organisational skills and the project assets positions EMS as a key driver in achieving the Consortium's objectives.

Asset	**Knowledge of journalistic, editorial team as regards AI tools, disinformation, news technologies**
Type	Knowledge
Description	As a media organisation, EMS provides to the project consortium competencies in terms of EU news reporting, policy analysis, and debunking false information. In this case, the asset brought to the project is the skills and expertise of the editorial teams towards challenges belonging to every aspect of disinformation.
Target Groups/Beneficiaries	Professionals and Researchers
Innovation, Exploitation and Sustainability Plan	EMS aims to empower its team's expertise with enhanced capabilities in content verification, source analysis, and disinformation detection. This sort of training on the job will also help EMS to improve competency and familiarity with everyday AI technologies (like AI translation robots, AI content production and generation – i.e., ChatGPT). This innovation will enable more efficient debunking of fake news, ensuring the dissemination of accurate and trustworthy information to our audience.

Asset	**News process and workflows**
Type	Processes
Description	EMS provides the project the possibility to test new AI tools into its existing news process. The skills of its teams will be used to understand the impact of the AI4TRUST models on existing processes while training of the teams on new technologies and tools will empower global expertise of the organisation.
Target Groups/Beneficiaries	Professionals and Researchers
Innovation, Exploitation and Sustainability Plan	The advanced AI solutions offer a unique added value by providing robust tools to combat disinformation effectively within the EU landscape. After the implementation and integration of the tool, EMS would be ready to take on the role of AI4TRUST Platform ambassadors, both by raising awareness of the tool within the Polish media community and by sharing its own experience in using the Platform. EMS plan is to exploit these assets both industrially and academically. Industrially, we aim to integrate the AI solutions into their news production pipeline, offering a more reliable and fact-checked news service to our audience. Simultaneously, we see an opportunity to position ourselves as a provider of commercial solutions, offering expertise and tools to other media organisations facing similar challenges. Academically, we envision contributing to ongoing research and development efforts, fostering collaboration within the consortium and beyond.

9.16. University of Cambridge (UCAM)

The Minderoo Centre for Technology and Democracy consists of an independent team of academic researchers affiliated with the **University of Cambridge (UCAM)**. They are engaged in fundamentally reimagining the dynamics of power between digital technologies, society, and the planet.

In today's context, **understanding the collective challenges arising from power, technology, and democracy** is more crucial than ever. UCAM is of the belief that discussions regarding **technology and regulation** often revolve around comprehending, interpreting, and implementing the ideals of democracy, particularly within complex modern states heavily reliant on technologies that are often unfamiliar to the majority of citizens. This understanding is **vital for citizens** to effectively govern themselves and shape their collective state. UCAM's research is centred around four primary objectives:

1. **Improving public comprehension of digital technologies** and their societal impacts;
2. **Highlighting the global environmental repercussions** stemming from digital technology;
3. **Proposing solutions to address the adverse effects of digital technologies** on citizens' rights;
4. **Cultivating informed trust in digital technology** while emphasising the significance of democratic values over corporate interests.

Asset	**Knowledge of policy networks and human-centred AI techniques**
Type	Knowledge
Description	Policy and Human-Centred AI techniques knowledge includes expertise, practical skills, and knowledge of how people use systems and how policymakers and civil society use and navigate systems and entails translating this knowledge into building tools to serve such users. Such expertise ensures the 'human-centredness' of the AI4TRUST Platform and the relevance of the Platform for its intended users.
Target Groups/Beneficiaries	Any platform or organisation that involves user-generated text, especially in domains with disinformation prone themes such as media outlets, researchers, fact-checkers, public health or climate change experts, policymakers, educators, etc.
Innovation, Exploitation and Sustainability Plan	The knowledge developed for the AI4TRUST model has the potential for reuse in providing services for potential users and UCAM reserves the right to provide further services and commercial exploitation of the knowledge developed for the project.

Asset	**Hybrid Recommendation Tool**
Type	Technology
Description	This tool can help end users to overcome their information gaps about the current state of the information ecosystem with platforms walling off access to data and can help identify systemic challenges and risks (e.g., public health, fundamental rights). To do so, the recommendation tool draws on different functionalities of the AI4TRUST Platform, namely the detection and analysis of disinformation signals (e.g., claim validity and

	Social Network Analysis). The purpose of the tool is to link aggregated inputs from the AI4TRUST Platform with a classification of their severity level and a guide towards mitigation measures. The hybrid character is conferred by both the human-centred design of the tool and the human supervision of the recommendation inputs, particularly when dealing with disinformation classified as a systemic risk. Based on aggregate outputs from the AI4TRUST Platform, such as Social Network Analysis (SNA), Coordinated Inauthentic Behaviour (CIB), reliability state of social media, and the classification of mis- or disinformation according to levels of severity, the Platform can produce semi-automated reports tailored to end users (e.g., policy makers, journalists, researchers). These reports can compile the human-led analyses of spread of disinformation campaigns, their virality and their reaction to it on social media platforms and visualise the spread of disinformation. The reports will not list automatic suggestions for mitigation measures but can present a scientific overview of mitigation measures and their shortcomings. Journalists and fact-checkers can use the reports to identify relevant, viral disinformation and thus better allocate their scarce resources. Policy makers will be interested to inquire about specific topics, time frames, and locations/language (national level or EU level).
Target Groups/Beneficiaries	Project partners and future users of the Platform (i.e., fact-checkers, media workers, policy makers)
Innovation, Exploitation and Sustainability Plan	Innovation resides in the hybrid component combining automated disinformation signals with human supervision and analysis, which draw on some of the principles and techniques developed in WP4. Although widely desirable, this hybrid design is rarely deployed in AI-based platforms. The sustainability of the tool is assured by the constant access to platform data streams as well as the existence of a critical mass of researchers, who can access the data and either analyse the output (e.g., with SNA tools) or verify the validity of prior claims. In this sense, the exploitation of the hybrid recommendation tool rests on two pillars: contextual relevance and scale. On the one hand, we expect to promote the relevance and use value of the hybrid recommendation tool to end users within the context of their own work. On the other hand, by scaling up the number of end users, the hybrid recommendation tool has the potential to become an independent resource for both information mediators

	(journalists, fact-checkers) and policy makers. The publication of scientific research outputs and policy reports will further contribute to expose and exploit the potential of the hybrid recommendation tool.

9.17. FINCONS

FINCONS is a leading player in **international business consulting and system integration** in support of technological and digital transformation and **in AI4TRUST is responsible for the integration of the project Platform** and **build a strategy to bring it to the market**, bringing to the projects its competencies and expertise on system integration, IoT, Big Data and Blockchain. FINCONS is leader of WP5, dedicated to the **design and development of the AI4TRUST Platform** and provision of **suitable technical solutions for assuring EU-GDPR and ethics compliance**.

Asset	**Fight dis/mis/malinformation technological issues**
Type	Knowledge
Description	Fincons provides to the project its expertise on media domain in terms of platform, regulations, and data. This knowledge will be enriched by specific topics, barriers, and facilitators (not only technological) dedicated to dis/mis/malinformation.
Target Groups/Beneficiaries	Researchers and professionals
Innovation, Exploitation and Sustainability Plan	As an industrial partner, FINCONS will use the acquired knowledge from the AI4TRUST outcomes to support researchers and customers in facing dis/mis/malinformation challenges, first integration into existing workflows of new AI technologies and tools.

Asset	**AI4TRUST Platform integration**
Type	Technology
Description	Integrating the AI4TRUST Platform means to integrate novel AI models, technologies and data in a unique framework dedicated to different stakeholders. The technical skills of FINCONS as a system integrator have provided the correct methodologies and instruments to create a consistent and reliable Platform.

Target Groups/Beneficiaries	Professionals
Innovation, Exploitation and Sustainability Plan	As other industrial companies in the Consortium (EURACTIV, SKYTG24, ADB, EMS), FINCONS will exploit the pilots results and deployed solutions to increase the level of quality of product and services of their portfolio, to fight dis/mis/malinformation, and to empower scientific researchers, media practitioners and policymakers with advanced AI-based technologies.

Asset	**Human Validation Tool**
Type	Technology
Description	The Human Validation Tool integrated in the web application of the AI4TRUST Platform will complete the fact-checkers workflow, allowing them to provide manual feedback about the news items passed through their monitoring and analysis activities. This enables a hybrid approach where the AI modules provide filters and insights for fact-checking, while the fact-checkers can manually evaluate the news items, ensuring the high level of reliability guaranteed by the human-in-the-loop.
Target Groups/Beneficiaries	Fact-checkers/journalists
Innovation, Exploitation and Sustainability Plan	As an industrial company, FINCONS will exploit this asset in conjunction with the other results of the pilots, in order to enhance the quality of its products and services, empowering media practitioners, fact-checkers and journalists to fight dis/mis/malinformation.