

TITAN

Grant Agreement #767227



TITAN

DELIVERABLE REPORT D2.1

DELIVERABLE

SUBMISSION DATE

31.8.2023

NAME OF THE DELIVERABLE

TITAN Socio-technical Framework
and User Needs Analysis

WORK PACKAGE

WP2

NATURE

Public

AUTHOR(S)

Part 1: A. F. Inguanzo, T. Meyer, G. Corbelli,
A. Caforio, I. Picone, G. Petasis, D. Politikos, G.
Stamatellos, M. Wozniak

Part 2: S. Papafragkos, A. Ramfos, G. Petasis,
M. Hoff, S. Fibecker Ladegaard, I. Picone, G.
Corbelli, M. Cipolla, A. Lamsa, V. Ollikainen,
M. Wozniak, T. Van Lier, C. Van Den Bosch, B.
Velcheva

LEAD BENEFICIARY

VUB

PROJECT DETAILS

PROJECT ACRONYM

TITAN

GRANT AGREEMENT

101070658

CALL IDENTIFIER

Human-01-27

PROJECT DURATION

01.09.2022 – 30.09.2025

PROJECT OFFICER

Peter Friess

PROJECT COORDINATOR

Engineering S.p.A.

QUALITY CONTROL ASSESSMENT

VERSION	DATE	DESCRIPTION	NAME	ORG
v0.1	Date	Table of contents	Author name	Author Organisation
v0.2	03/2023	First draft T2.1	Ana Fernández/Trisha Meyer	VUB
v0.3	29/03/2023	External review T.2.1	External Ethics Advisory Board	
v0.4	04/2023	Internal review/update T.2.1	T2.1 partners	VUB, NCSR-D, UNINET
v0.5	06/2023	First draft T2.2	T2.2 partners	ATC, DBT, VRT
v0.6	17/08/2023	Merge T2.1 and T2.2	Ana Fernandez	VUB
V0.7	25/08/2023	Update T2.2	T2.2 partners	ATC, VUB NCSR-D, CSD UNINET
v0.8	09/2023	Internal review	Massimo Magaldi	ENG
v0.9	02/10/2023	Incorporating last changes	Ana Fernández/Trisha Meyer	VUB
V1.0	6/10/2023	Final quality check	Massimo Magaldi	ENG

DISCLAIMER

The opinions stated in this report reflect the opinions of the authors and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein. All intellectual property rights are owned by the TITAN consortium members and are protected by the applicable laws. Except where otherwise specified, all document contents are: “©TITAN Project - All rights reserved”. Reproduction is not authorised without prior written agreement. The commercial use of any information contained in this document may require a license from the owner of that information.

STATEMENT OF ORIGINALITY

This Deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

ACKNOWLEDGEMENT

This document is a deliverable of the TITAN project, which has received funding from the European Union's Horizon 2020 Programme under Grant Agreement (GA) #101070658 and by UK Research and innovation under the UK governments Horizon funding guarantee grant numbers 10040483 and 10055990.

TABLE OF CONTENTS	4
EXECUTIVE SUMMARY	7
PART 1: SOCIO-TECHNICAL FRAMEWORK	9
1 INTRODUCTION	9
1.1 INTRODUCTION	9
1.2 DISINFORMATION	11
1.3 FAKE NEWS IN SOCIAL MEDIA	12
1.4 COMPLEMENTING FAKE NEWS DETECTION WITH FACT-CHECKING	13
1.5 CURRENT SELF-FACT CHECKING STEPS	13
1.6 ARTIFICIAL INTELLIGENCE	13
1.7 DISINFORMATION TYPES	14
1.8 REFERENCES	15
2 SOCIO-DEMOGRAPHIC AND SOCIO-POLITICAL FACTORS AFFECTING CITIZEN (DIS)TRUST AND (DIS)ENGAGEMENT	18
2.1 WHY TRUST MATTERS WHEN LOOKING TO STRENGTHEN CRITICAL THINKING	18
2.2 THE STRUGGLE WITH TRUST IN NEWS	18
2.2.1 <i>The ambivalence of trust in news</i>	19
2.2.2 <i>To trust or not trust in news</i>	19
2.2.3 <i>To trust your gut</i>	20
2.2.4 <i>To trust untruthful information</i>	20
2.2.5 <i>The good the bad and the trusted</i>	21
2.3 SOCIO-DEMOGRAPHIC AND SOCIO-POLITICAL FACTORS AFFECTING TRUST IN / DISCONNECTION FROM NEWS	22
2.3.1 <i>The socio-demographics of news and information consumption</i>	22
2.3.2 <i>The socio-demographics of media trust</i>	22
2.3.3 <i>It's the politics, stupid!</i>	24
2.3.4 <i>Zooming in on news avoidance and disengagement</i>	27
2.3.5 <i>Zooming in on the likeliness to believe fake news and disinformation</i>	28
2.3.6 <i>Ceteris paribus in AI solutions, or the lack thereof</i>	31
2.4 THE AMBIVALENT RELATIONSHIP BETWEEN TRUST AND CRITICAL THINKING	32
2.5 REFERENCES	34
3 PSYCHOLOGICAL/MOTIVATIONAL/COGNITIVE MECHANISMS	37
3.1 SOCIAL COGNITIVE THEORY (SCT): THE BROAD PSYCHOLOGICAL FRAMEWORK	37
3.1.1 <i>Personal factors</i>	37
3.1.2 <i>Situational and contextual factors</i>	37
3.1.3 <i>Behavioral factors</i>	38
3.2 CRITICAL THINKING WITHIN AN AGENTIC PERSPECTIVE	38
3.2.1 <i>Main cognitive skills of critical thinking</i>	39
3.2.2 <i>Self-regulatory feedback loop for learning critical thinking</i>	40
3.3 SELF-REGULATED LEARNING AND CRITICAL THINKING AGAINST DISINFORMATION	40
3.3.1 <i>Critical thinking and individual vulnerability</i>	42
3.4 THE SOCRATIC METHOD AND CRITICAL THINKING ASSESSMENT	42
3.4.1 <i>The Socratic method and Socratic AI: philosophical and educational approaches</i>	42
3.4.2 <i>Improving critical thinking with collective argument mapping in a collaborative module</i>	43
3.5 REFERENCES	44
4 AI TECHNOLOGY TRENDS FOR MODELS, DATA AND RESOURCES TO FIGHT DISINFORMATION	47

4.1	STATE-OF-THE-ART PROCEDURES AND TOOLS FOR FACT-CHECKING	47
4.1.1	<i>Overview of the research, writing and review process</i>	48
4.1.2	<i>Source of an online claim</i>	48
4.1.3	<i>Review process</i>	49
4.1.4	<i>Editorial Materials</i>	49
4.1.5	<i>Skills</i>	50
4.1.6	<i>Training fact checkers to check online claims</i>	50
4.1.7	<i>Evidence: Tools, websites, and access to data and institutions</i>	51
4.1.8	<i>Publication and distribution of online fact checks</i>	51
4.1.9	<i>Technology</i>	52
4.2	RECOGNIZING DISINFORMATION SIGNALS	54
4.2.1	<i>Key disinformation signals</i>	54
4.2.2	<i>Categories of disinformation signals</i>	55
4.2.3	<i>Identifying logical fallacies</i>	55
4.2.4	<i>State-of-the-art AI models, data and approaches for detecting fact-checking, fallacies and hate speech</i>	57
4.3	REFERENCES	57
5	LEGAL AND ETHICAL DIMENSIONS OF TRUSTWORTHY AI TO FIGHT DISINFORMATION	58
5.1	THE CONCEPT OF TRUSTWORTHY AI.....	58
5.2	ETHICAL DIMENSIONS	58
5.2.1	<i>Section 1– How to apply ethics by design in AI Development: practical steps</i>	61
5.3	LEGAL DIMENSIONS	64
5.3.1	<i>General Data Protection Regulation (GDPR)</i>	65
5.3.2	<i>Artificial Intelligence Act</i>	66
5.3.3	<i>Digital Services Act (DSA)</i>	67
5.3.4	<i>Fundamental rights</i>	68
5.3.5	<i>International and national developments and best practices in AI</i>	69
5.4	NEXT STEPS: THE PROCESS FOR OBTAINING THE ETHICAL AND LEGAL REQUIREMENTS OF TITAN	70
5.5	REFERENCES.....	71
6	MAPPING EXISTING USER TACTICS TO COPE WITH DISINFORMATION.....	75
6.1	AUTHENTICATION.....	75
6.1.1	<i>Internal authentication</i>	75
6.1.2	<i>External authentication</i>	76
6.2	CORRECTION.....	76
6.2.1	<i>Intrapersonal tactics</i>	76
6.2.2	<i>Interpersonal and interventional tactics</i>	77
6.3	CONCLUSION.....	78
6.4	REFERENCES.....	78
7	CONCLUSION PART 1: A SET OF REQUIREMENTS FOR TITAN’S SOCIO-TECHNICAL FRAMEWORK.....	80
	PART 2: USER NEEDS ANALYSIS	82
8	TITAN COACHING SYSTEM: A TWO STEP APPROACH	83
8.1	STEP 1: RULE-BASED CHATBOT DEVELOPMENT.....	83
8.1.1	<i>The Socratic method</i>	84
8.1.2	<i>Compilation of micro-lessons</i>	86
8.2	STEP 2: INTEGRATION OF A LARGE LANGUAGE MODEL	89
9	USER NEEDS COLLECTION METHODOLOGY	90
9.1	APPROACH	90
9.2	EXPECTATIONS.....	91
9.3	TARGET USER GROUPS	92

10	USER REQUIREMENTS RESULTING FROM USER TACTICS TO COPE WITH DISINFORMATION	94
11	USE CASES DEFINITION AND USER NEEDS ANALYSIS	96
11.1	USE CASE 1: HIGH EDUCATION FIGHT AGAINST DISINFORMATION	96
11.1.1	<i>Use case 1 user needs analysis.....</i>	<i>97</i>
11.2	USE CASE 2: NGO'S CITIZENS AS ACTIVISTS	100
11.2.1	<i>Use case 2 user needs analysis.....</i>	<i>101</i>
11.3	USE CASE 3: IMMIGRATION.....	102
11.3.1	<i>Use case 3 user needs analysis.....</i>	<i>103</i>
12	CO-CREATION OF THE REQUIREMENTS FOR THE TITAN AI SERVICE/PLATFORM.....	105
12.1	CITIZEN INSIGHTS QUALIFIED BY EXPERTS - RECOMMENDATIONS FOR THE TITAN TOOL.....	105
12.2	USER REQUIREMENTS ON INFORMED CONSENT IN AI TOOLS.....	106
12.3	USER REQUIREMENTS ON DATA AND TRUST	107
12.4	USER REQUIREMENTS ON AI LEARNING TOOLS AND CITIZENS NEEDS	108
13	FOCUS GROUPS WITH TITAN PARTNERS	111
14	USER REQUIREMENTS CONSIDERING LEGAL AND ETHICAL ASPECTS OF TRUSTWORTHY AI	116
15	MOCK-UPS CREATION AND ITERATIONS.....	117
16	CONCLUSION AND NEXT STEPS PART 2: USER NEEDS ANALYSIS	119
17	ANNEX 1: FALLACIES AND FACT-CHECKING DATASETS.....	121
17.1	FALLACIES DATASET SUMMARY	121
17.2	FACT-CHECKING DATASET SUMMARY	127
18	ANNEX 2: CONCLUSION LIST OF USER REQUIREMENTS.....	150
19	ANNEX 3: ARTICLES AND CORRESPONDING SAMPLE DIALOGUES.....	166

EXECUTIVE SUMMARY

TITAN aims to develop a tool for enhancing critical thinking against disinformation, through guidance, and coaching. This report presents the results of tasks T2.1 and T2.2.

More specifically, **Part 1** states the basis and groundwork for the TITAN project, to explore and study relevant literature on citizen engagement in the fight against disinformation. It aims at three main objectives:

- (1) to identify the socio-demographic and socio-political factors affecting citizen (dis)trust and (dis)engagement and the psychological, motivational and cognitive mechanisms.
- (2) to identify AI technologies (models, data and resources) relevant to set the socio-political and psychological/motivational/cognitive context of citizen's engagement in the fight against disinformation.
- (3) to study the basis for an ethical, legal and trustworthy AI conversational tool to tackling online disinformation.

Part 1 sets the basis for the design process of the TITAN system, and it will accompany TITAN's ecosystem. It will be used to create an ethical and legal impact assessment (D1.4) that will also be accompanied by a questionnaire (to be filled in by all members) to consistently review their datasets and/or other practices that have risks to the rights of individuals and other ethical issues with the implementation of this technology. That is, this deliverable aims to build the necessary knowledge to set up the basis for an ethical, psychological, legal and trustworthy AI conversational tool to fight disinformation.

Part 2 marks the progress of our work and research conducted throughout the first year of the project. It includes our two-step approach, the trajectory for the critical thinking assessment and the user needs analysis. Our project aims to combat disinformation by enhancing critical thinking skills through a chatbot. To achieve this, we have devised a two-step approach:

- **Rule-Based Coaching System:** In the initial phase, we will develop a rule-based coaching system within the chatbot. This system will provide structured guidance and support to users in combating misinformation and improving their critical thinking abilities.
- **Integration with Large Language Model:** Subsequently, using the data collected from our living labs, we will integrate this rule-based system with a sophisticated large language model. This integration will enhance the chatbot's capabilities, enabling it to provide more dynamic and context-aware assistance to users.

In addition to this approach, we have made significant progress in the assessment of critical thinking skills. This report provides a glimpse of our ongoing efforts in this area and outlines the directions we plan to follow in the next phases of the project.

User Needs Analysis:

We've examined three primary user groups: higher education students, NGOs and activists, and migrants and refugees. In addition, through co-creation workshops citizens of diverse groups complement the user requirements collection.

During the user needs analysis, we identified key requirements from our user groups:

- Higher education students seek to have control over the tool, need encouragement, and personalized guidance.
- NGOs and activists require analytical information about their data usage, the ability to share verified information and interaction with the tool for further engagement and motivation.
- Migrants and refugees need culturally sensitive language support, access to local news, and tools to navigate disinformation.

Moreover, from the co-creation workshops, we identified user requirements that span from transparency and security concerns to ensuring a seamless user experience.

The insights from this analysis will serve as the foundation for tailoring our chatbot to meet the unique needs of these user groups. We will also continue to gather user feedback for ongoing improvements.

This report underscores our commitment to developing an effective solution to combat disinformation and improve critical thinking skills. As we progress, we will work diligently to fulfil the objectives of the project and make a meaningful impact in the fight against disinformation.

1 INTRODUCTION

1.1 INTRODUCTION

The increased capacity of algorithms and Artificial Intelligence (hereinafter AI) to gather and analyze data influence the opinions of citizens and their access to news, as they allow for personalized recommendations, categorization and ranking of content. The news landscape, especially in political communication, is changing rapidly with these technologies facilitating the rapid spread of news; and while the digital has brought many positive advantages, it has also brought issues that affect democratic principles. There is a considerable amount of disinformation all over the world, blurring the fine line between facts and opinions and impacting the capacity of citizens to gather information and develop opinions based on facts.

In this regard, UNESCO published a study [1] in 2018 highlighting the increased difficulties for citizens in recent years to combat misinformation through social networks and the lack of regulation in this regard. In the same year, the European Commission published a report [2] about misinformation online as a growing concern and offered recommendations to combat disinformation and to promote values and standards for the next years. In 2019, the European Parliament published two studies on disinformation and AI. One conducted by EU DisinfoLab [3] studied current and future threats to disinformation as a growing phenomenon. The second by Brussels School of Governance at Vrije Universiteit Brussel [4] studied the consequences of using AI to fight disinformation. The second strongly suggests going against the increased use of AI in content moderation, without prioritizing human review processes. Both studies expose how disinformation online is a growing issue that affects democracies.

Disinformation is gaining importance from other authors that analyze digital infrastructure. Several scholars [5] [6] have studied how disinformation occurs, such as Vosoughi et al. (2018) who demonstrate that falsehoods spread six times faster on Twitter than regular or factual content [5]. The results of this study were further analyzed by Juul and Ugander (2021) who argued that to understand how news are spread, we need a bigger sample of the ‘tuer’ news, as they show, Vosoughi et al. were based on 24,409 verified true news and 82,605 false news). They further argue that to limit the spread of false news, it may be sufficient to focus on reducing the transmission of false information (what they call “the infectiousness”) [7].

In addition, Zimmermann and Kohring (2020) [8] conclude that in 2017 fake news online played a considerable role in diminishing the trust of German citizens in politics. Tandoc et al (2018) argue that developments in AI, such as bots, facilitate the creation of fake websites, thus helping the spread of fake news [9]. Van Raemdonck and Meyer (2022) [10] contend that digital technology is “a force multiplier” that offers new ways to rapidly spread and create disinformation, such as the capacity of algorithms to reinforce conspiracy theories in an unprecedented manner. The authors point out how disinformation can affect people’s political opinions because it “threatens civil discourse and unravels a shared sense of reality. Disinformation not only harms democracy but is most effective when democracy is in peril and public trust in institutions is low”. Similarly, Bennett & Livingston (2018) claim that disinformation is “obscure” [11] and can appear for different reasons: “(a) authoritative information (b) emanating from social and political institutions that (c) engage trusting and credulous publics”. They highlight that the loss of trust in institutions is a growing danger emanating from this ‘obscurity’ and complexity, which helps increase the consumption of misinformative claims. Citizens being misled or influenced (negatively) through social networks, and the difficulty to tackle these issues, form an important politico-social driver for the TITAN project.

Nevertheless, the term ‘disinformation’ is not new, and it has existed in mainstream media history before [12], with narratives that aim at deceiving or disinforming people. As an example, a US Fox News program reported that there were massive Muslim refugees and immigrant crimes committed in Sweden. Although soon proven to be inaccurate (based on movie images) [11], Donald Trump projected the story as true, which helped to

spread it further and strengthened the right-wing narrative against migrants. Disinformation also results in the need to continually prove something as true, such as in the case of former U.S president Barack Obama, who had to show his official birth certificate because of claims of him being non-American.

Social networks (e.g., Twitter, Instagram, YouTube, TikTok, Facebook...) allow for a quick exchange of information with a vast number of channels and different sources – also with non-reliable sources outside mainstream media that emphasize virality over veracity, as well as more independent discourses or narratives that challenge logical and democratic principles. The disinformation scheme can be found in different forms, namely: misinformation, fake news, propaganda, conspiracy theories, clickbait, post-truth, even rumors to name a few. All these types are part of the disinformation label, and they have the intent to partially present the truth or falsely claim connections or events, thus it can be both incorrect and partially accurate, where social networks increase exponentially the spread and consumption of false claims. Examples are numerous, such as disinformation campaigns conducted in social networks that use bots to increase and spread fake news [13].

Tools such as bots are facilitating the processing, creation of summaries and writing of content, which could potentially increase the issues that we face among social networks and other digital technologies. The past months have brought even more innovation, with the introduction of OpenAI, particularly the project called ChatGPT [14], a language model chatbot that is capable of creating *new* content by studying patterns from large datasets. This is not the only language model, there are others such as Google’s Bard, and Facebook LLaMA, however ChatGPT was released to the public to be used freely in November 2022 and is gaining fame, to the extent that prominent AI experts have been called for a moratorium on training more powerful large language models[15]. It was trained on large data sets from the Internet, such as Wikipedia, journalistic articles, and scientific articles. It can, for instance, autonomously write pieces of journalism (or any other type of written content such as student essays, poetry etc.) by entering quite simple instructions. The release of ChatGPT and its integration into search engines shook policy-making, tech and knowledge-based communities, as it is currently the most powerful trained publicly available AI language model. Owned by a few investors, this startup joined forces with Microsoft to train the data.

Considering the disinformation issues already occurring in our digital society, ChatGPT could potentially increase them by being capable of producing a vast amount of novel content, at an even greater speed. Like many other tools, it can be used with good or bad intentions, or as Chomsky claimed, ChatGPT brings optimism because “intelligence is the means by which we solve problems” [16], and to have a technology capable of finding patterns almost impossible for the human brain brings new possibilities. However, he follows by warning that “we fear [...] machine learning — will degrade our science and debase our ethics by incorporating into our technology a fundamentally flawed conception of language and knowledge”. Chomsky, who is an acclaimed expert in linguistics, warns that these systems have limitations, especially because they do not produce new knowledge but find patterns in the preexisting vast data points[17]. AI chatbots in this regard raise new legal and ethical questions, which TITAN needs to consider.

Issues with ChatGPT have started to emerge, such as how these tools named ‘intelligent’ still rely on humans, particularly Kenyan workers, to review content on topics including child sexual abuse, suicide, etc to try to make the tool (in the author’s words) ‘less toxic’ [18]. Copyright is also an issue, as the use of original sources is being reduced with people gathering information from the summaries created by language models, even regarding human creativity and art, where the AI Dall-E, also developed by OpenAI, was benefiting from thousands of existing copyright images [19]; as well as far less controversial examples, nonetheless important, of AI language models being capable of gathering information behind paywalls, which also bring issues of paid subscriptions, licenses, and access to content. An individual can access particular information through an AI language model that was behind paywalls in the original source [20].

This deliverable provides a baseline for TITAN, reflecting on social, political, cognitive, legal and ethical drivers relevant to the need for and response to disinformation in an increasingly AI-driven society (T2.1); and on the practical consequences and choices such drivers have for the TITAN project (T2.2). In the remainder of this introduction, we briefly define two key concepts in the project, namely, disinformation and artificial intelligence.

1.2 DISINFORMATION

Disinformation in the digital context is a complex concept, there is not one type of agreed definition on disinformation. The term can refer to online manipulation, misinformation or ‘fake news’. These different connotations lead to different issues, but often focus on disinformation as a manipulative practice. *Manipulation* in the digital context has been defined as “the use of information technology to covertly influence another person’s decision-making, by targeting and exploiting their decision-making vulnerabilities” [21]. Similarly, *fake news* has been defined as a “deliberate creation of pseudojournalistic disinformation” [22].

Two terms – misinformation and disinformation – are often used in similar but slightly different contexts. **Disinformation** is defined in the EU guidelines as “verifiably false or misleading information that is created, presented, and disseminated for economic gain or to intentionally deceive the public. It can cause public harm” [23] while in the same report, **misinformation** is defined as “verifiably false information that is spread without the intention to mislead, and often shared because the users believe it to be true”, the latter shows no purpose or intention. These definitions share a common thrust, which is the act of misguiding people through false information, but the intentionality or motive is what distinguishes them. In this line, misinformation refers to inadvertently sharing false information.

The growing problem of disinformation also comes with new characteristics. **Fake news** refers to the spread of fake content – that have inter alia often been created by autonomous systems and the “deliberate creation of pseudojournalistic disinformation” [22]. Yet, it has also been used to delegitimize journalism by political actors, particularly if the media contradicts their claims. Thus, fake news is often directed at mainstream media. Seeking to salvage the popularized term, Egelhofer and Lecheler (2019) studied the concept of fake news and argue that three characteristics must be met to call it ‘fake news’: it (1) has journalistic form (2) is low in facticity and (3) has the intent to deceive [24]. Tandoc et al (2018) argued that – even if the content is not false (e.g., if the content is not merely fabricated) – it can contain elements that can also be considered fake news, such as if it is only partly untrue (e.g., a false connection of events), incorrect, or written to provoke or entertain [9]. Thus, these authors argue that the content of fake news does not necessarily to be false or fake, it suffices to be partially incorrect with the goal to deceive people.

The TITAN project adopts the EU definitions of disinformation and misinformation. It will use fact-checked instances of disinformation in its interactions with citizens and addresses misinformation by increasing logical reasoning, critical thinking and digital media literacy through its micro-lessons. As for ‘fake news’, we focus on its similarities with disinformation, which is the intention to confuse or manipulate the public with false or misleading claims.

These terms bring to light a prominent issue in current digital societies: the fine line between how facts and beliefs are presented to the public. Certain citizens not only do not trust the media but increasingly consider scientifically proven concepts as lies, such as on climate change or that the earth is round. In this regard, Van Raemdonck and Meyer (2022) present disinformation as a ‘hybrid threat’ related to several factors, “the incentives, technical drivers and socio-psychological factors that are present in these hyperconnected times. Abuses of power, dysfunctional political systems, inequalities and exclusion are breeding grounds for disinformation” [10]. Psychological factors such as beliefs, individual’s emotional, cognitive aspects, being a vulnerable or disadvantaged group in society, also influence the acceptance or agreements with such

misinformative claims. In this sense, disinformation has become a complex issue that takes place in many areas of society, where on the one hand digital developments in information allow to easily spread news, reinforce specific content or create it and on the other hand, the psychological and social characteristics of individuals are also important.

1.3 FAKE NEWS IN SOCIAL MEDIA

Disinformation in social media is presented with unique challenges and unique research problems that make existing detection methods from traditional media ineffective or not applicable. First, fake news circulated in the social media is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content. The content of fake news is rather diverse in terms of topics, styles and media platforms, and fake news attempts to distort truth with diverse linguistic styles while simultaneously mocking true news. Existing hand-crafted and data-specific textual features are generally not sufficient for fake news detection. Therefore, we need to include auxiliary information, such as user social engagements on social media, to help make a determination.

As there is no agreement upon the definition of the term “fake news” [11], an indicative and non-exhaustive list of features of fake news, that could be (partially) detected by AI-tools, may include:

- *Linguistic features* that capture the different writing styles and sensational headlines to detect fake news
- *Visual-based features* are extracted from visual elements (e.g., images and videos) to capture the different characteristics for fake news
- *User-based features* that represent the characteristics of those users who have interactions with the news on social media.
- *Post-based features* that focus on identifying useful information to infer the veracity of news from various aspects of relevant social media posts
- *Network-based features* that are extracted via constructing specific networks among the users who published related social media posts, including
 - The *stance network* can be built with nodes indicating all the tweets relevant to the news and the edge indicating the weights of similarity of stances
 - The *co-occurrence network*, which is built based on the user engagements by counting whether those users write posts relevant to the same news articles
 - The *friendship network* indicates the following/followee structure of users who post related tweets
 - The *diffusion network*, which tracks the trajectory of the spread of news, where nodes represent the users and edges represent the information diffusion paths among them

Nowadays, the sophistication of spreading fake news in online communities of citizens has risen. Experience from vibrant online communities has revealed high levels of sophistication in spreading fake news, such as [12] the rise of claims designed to “game the system”, i.e., claims that do not technically break any rules of the online community, but that are designed with a partisan goal in mind, and written in a way to support that goal, including:

- a claim that is difficult or impossible to fact-check, and then supporting it with evidence that seems to imply an answer. At best, users could only respond with “I don't know” or “We don't have enough information”, and if such a claim is to be tagged, it would undoubtedly be “Unverifiable”. However, the lack of a refutation, along with links to evidence supporting the claim in question, could appear to many as support for that claim, even if that is not the case.
- a claim that is technically correct, but implies something that is false, because the claim is missing a greater context. An example of such a claim would be “everyone who drinks water in Brussels dies” - this is technically true, because everyone dies, but it implies something that is not true - that Brussels water is the cause. Again, we can respond to this by tagging the claim with something like “Partisan

Bias”, but those pushing these claims will undoubtedly respond by saying “if it's factually true, how can it be partisan bias?”

It is, after all, very possible to produce fake news by starting from true factual statements and drawing false conclusions by applying skewed, biased, or otherwise defective reasoning. That is why TITAN argues that fact-checking should be supplemented with reason-checking: evaluating whether the complete argumentative reasoning is acceptable, relevant, and sufficient.

1.4 COMPLEMENTING FAKE NEWS DETECTION WITH FACT-CHECKING

TITAN will harness AI-driven Socratic dialogues to optimize citizen self-searching movements for truth in relation to the suspicious statement or claim in-hand and according to his/her ‘critical-thinking’ maturity level. AI-enabled formal argumentation theory structures [13] will intelligently and interactively guide the generation of self-searching Socratic dialogues, while evaluating citizen responses to ensure quality of the self-searching process. In this way, the citizen will be able to work through self-checking hypotheses critically and systematically, hence, efficiently.

1.5 CURRENT SELF-FACT CHECKING STEPS

Indicative elements of current citizen self-searching steps, (which will guide the generation of Socratic questions), may concern any of the following:

- learning more about the source,
- checking if an image is used with the correct content,
- getting the whole story beyond provocative headlines,
- checking the author,
- verifying evidence,
- spotting twisted news stories or satire,
- seeing how (and if) different news agencies have reported on the same event,
- searching in real-world checks from trusted sources around the world,
- verifying location of event by using public tools such as Google Maps, Google Earth, or Google Street View

1.6 ARTIFICIAL INTELLIGENCE

In 2018, the European Union published a communication notice regarding Artificial Intelligence, that was developed by the EU high-level expert group on AI (AI HLEG). Artificial Intelligence systems are defined as [25]: “[...] systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).”

The term “Artificial Intelligence” (AI) is changing rapidly, and policymakers and stakeholders have adopted different definitions of the concept, even within the proposed AI Act [26]. It is a constantly changing technology that has been considered differently across countries, and the definition must account for all variables, especially if there will be new regulatory actions based on that definition. Renda and Engler [26] argue that the concept should involve a broad definition, and account for specific sectors. The authors explain how the EU considers that an AI has “some degree of autonomy” however, that leaves out “simpler” algorithmic techniques such as rule-based AI, which can also lead to limitations in fundamental rights.

Following AI HLEG, TITAN will adopt this definition as a conceptual basis, as it offers the foundations from which new AI regulatory actions are being developed in EU. Their definition of AI is broad and includes language models such as TITAN’s chatbot, thus making it applicable to new regulatory actions. In addition, the introduction has shown how disinformation is a complex topic that refers to several cognitive, social and technical characteristics. The AI HLEG has also developed ethical guidelines for a trustworthy AI that will also guide TITAN’s ecosystem, and with it, we aim to facilitate the creation of ethical tools to combat disinformation. These guidelines will be discussed in detail in section T.2.1.4, which refers to the legal and ethical dimensions of AI in TITAN’s system.

1.7 DISINFORMATION TYPES

Spreading of false or inaccurate information is a phenomenon being present in all human societies. The digitalization of media has accelerated the scale, volume, and distribution speed of disinformation over the past years, raising concerns in governments, businesses, and citizens [27,28]. According to the high-level group [29], “Disinformation” refers to “all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit”. United Nations High Commissioner for Refugees [4, page 230-231] released a fact sheet with the ten common types of disinformation, which are: Fabricated Content, Manipulated content, Imposter content, Misleading content, False context, Satire and Parody, False connections, Sponsored content, Propaganda, and Error. Additional disinformation types are Junk Science, Rumor [30,31], and Bots/Cyborgs [32]. A concise text for each type of disinformation is,

1. *Fabricated content* is 100% false content. For instance, a fake video created by filmmakers, to draw attention to the horrors taking place in a war.
2. *Manipulated content* occurs when genuine information or imagery is distorted. For instance, a populist ‘click bait’, the exaggeration about the health status of a public person or slowing down a footage so that a person appears drunk and slurring his words [33].
3. *Imposter content* refers to the impersonation of genuine sources by agents of disinformation using the logos of established brands to peddle false and misleading content. For instance, traffickers created pages with the UNHCR logo and posted images of beautiful yachts and “call this number” advertisements, promising safe passage of refugees across the Mediterranean Sea [33]. Accordingly, sophisticated imposter versions of well-known newspapers distribute fake news (e.g., the president of France was being funded by Russia).
4. *Misleading content* relies on context and nuance, with the aim of misrepresenting information to frame issues or individuals. This may be achieved by several ways, such as by reframing stories in headlines, cropping images or quotes to support a wider point, citing selective statistics in a way that aligns with a position or deciding not to present something on purpose. The goal in all cases is to change the meaning of the original message¹.
5. *False context* combines genuine content with false contextual information, e.g., when the headline of an article does not reflect the content, or when a photograph captured from the past is presented as a recent one. It is a powerful form of information disorder, as the content used is genuine, so it cannot be denied, but is reframed in a way to support a certain argument.
6. *Satire and Parody* attempts to pass false stories as true with humor and entertaining mood. Although, there is not always intention to harm, readers may be fooled. In some cases, repeated satire and parody may polarize and divide the audience.
7. *False connections* take place when headlines, visuals or captions do not support the content. They tend to have “Clickbait” headlines that don’t always give you the whole story; e.g., “*Sugar as addictive as cocaine and heroin, studies suggest*”, but if someone reads the article, he will realize that this theory has only been suggested to exist in rats, not humans.

¹ <https://commonslibrary.org/disinformation-and-7-common-forms-of-information-disorder/>.

8. *Sponsored content* puts out advertisements for a product, service, or brand that look like editorial content². It's a strategy from the media to charge companies for this advertising and a strategy from companies to reach customers. Studies show that young people have a difficulty to recognize the difference between a sponsored post and an actual news article on the same website³. For instance, while reading online news, you may click an article about the condition of women's prisons. But if you look closely at the illustration, the article turns out to be an advertisement for a new TV show⁴.
9. *Propaganda* spreads information, rumors or ideas to manage attitudes, values and knowledge. It's based on a systematic propagation of information or ideas in a tendentious way to encourage or instill a certain attitude or response⁵.
10. *Error* when a mistake is made by established new agencies in their reporting.
11. *Junk Science* refers to sources that promote pseudoscience, metaphysics, naturalistic fallacies, and other scientifically dubious claims.
12. *Rumor* traffics rumours, gossip, innuendo, and unverified claims.
13. *Bots* spread disinformation mainly through hashtag campaigns or by sharing news in the form of links, videos, memes, photos or other content types. They are algorithms that are programmed to engage autonomously on social media. For instance, hashtag campaigns involve bots constantly tweeting the same hashtag, or set of hashtags, in coordination. The goal is often to trick Twitter's trending algorithm into adding a specific hashtag to the trending topics list. An example is "#Hillarysick," which was propagated widely by bots after Hillary Clinton stumbled in September 2016, shortly before the presidential election⁶.
14. *Cyborgs* are hybrid accounts in which a human periodically takes over a bot account from time to time. An anonymous account which is inhumanly active, and which hyper-tweeting one point of view is likely to be a bot. The more impersonal an account's handle, screen name, bio and avatar, the more likely it is to be fake.

1.8 REFERENCES

- [1] <https://plus.google.com/+UNESCO>, "Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training," *UNESCO*, Sep. 03, 2018. <https://en.unesco.org/fightfakenews> (accessed Mar. 10, 2023).
- [2] "Final report of the High Level Expert Group on Fake News and Online Disinformation | Shaping Europe's digital future." <https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation> (accessed Mar. 24, 2023).
- [3] "Automated tackling of disinformation-Major challenges ahead | Panel for the Future of Science and Technology (STOA) | European Parliament." [https://www.europarl.europa.eu/stoa/en/document/EPRS_STU\(2019\)624278](https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624278) (accessed Mar. 24, 2023).
- [4] "Regulating disinformation with artificial intelligence | Panel for the Future of Science and Technology (STOA) | European Parliament." [https://www.europarl.europa.eu/stoa/en/document/EPRS_STU\(2019\)624279](https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624279) (accessed Mar. 24, 2023).
- [5] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018, doi: 10.1126/science.aap9559.
- [6] V. Bakir and A. McStay, *Optimising Emotions, Incubating Falsehoods: How to Protect the Global Civic Body from Disinformation and Misinformation*. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-031-13551-4.
- [7] J. L. Juul and J. Ugander, "Comparing information diffusion mechanisms by matching on cascade size," *Proc. Natl. Acad. Sci.*, vol. 118, no. 46, p. e2100786118, Nov. 2021, doi: 10.1073/pnas.2100786118.

² <https://bakersfieldcollege.libguides.com/c.php?g=729002&p=7907555>.

³ <https://www.theverge.com/2016/11/22/13712996/fake-news-facebook-google-sponsored-content-study>.

⁴ <https://edu.gcfglobal.org/en/digital-media-literacy/what-is-sponsored-content/1/#>.

⁵ <https://guides.lib.wayne.edu/c.php?g=401320&p=2729574>.

⁶ <https://datajournalism.com/read/handbook/verification-3/investigating-actors-content/3-spotting-bots-cyborgs-and-inauthentic-activity>.

- [8] “Zimmermann and Kohring - 2020 - Mistrust, Disinforming News, and Vote Choice A Pa.pdf.”
- [9] E. C. Tandoc, Z. W. Lim, and R. Ling, “Defining ‘Fake News’: A typology of scholarly definitions,” *Digit. Journal.*, vol. 6, no. 2, pp. 137–153, Feb. 2018, doi: 10.1080/21670811.2017.1360143.
- [10] N. Van Raemdonck and T. Meyer, “Why Disinformation is Here to Stay. A Socio-technical Analysis of Disinformation as a Hybrid Threat,” in *Addressing Hybrid Threats: European Law and Policies*, L. Lonardo, Ed., Edward Elgar, 2022.
- [11] W. L. Bennett and S. Livingston, “The disinformation order: Disruptive communication and the decline of democratic institutions,” *Eur. J. Commun.*, vol. 33, no. 2, pp. 122–139, Apr. 2018, doi: 10.1177/0267323118760317.
- [12] F. Zimmermann and M. Kohring, “Mistrust, Disinforming News, and Vote Choice: A Panel Survey on the Origins and Consequences of Believing Disinformation in the 2017 German Parliamentary Election,” *Polit. Commun.*, vol. 37, no. 2, pp. 215–237, Mar. 2020, doi: 10.1080/10584609.2019.1686095.
- [13] M. Himelein-Wachowiak *et al.*, “Bots and Misinformation Spread on Social Media: Implications for COVID-19,” *J. Med. Internet Res.*, vol. 23, no. 5, p. e26933, May 2021, doi: 10.2196/26933.
- [14] A. Hern, “AI bot ChatGPT stuns academics with essay-writing skills and usability,” *The Guardian*, Dec. 04, 2022. Accessed: Mar. 09, 2023. [Online]. Available: <https://www.theguardian.com/technology/2022/dec/04/ai-bot-chatgpt-stuns-academics-with-essay-writing-skills-and-usability>
- [15] “Pause Giant AI Experiments: An Open Letter - Future of Life Institute.” <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> (accessed Apr. 07, 2023).
- [16] N. Chomsky, I. Roberts, and J. Watumull, “Opinion | Noam Chomsky: The False Promise of ChatGPT,” *The New York Times*, Mar. 08, 2023. Accessed: Mar. 13, 2023. [Online]. Available: <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- [17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, in FAccT ’21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.
- [18] P. Marx, “Artificial intelligence’s dirty secret,” *Business Insider*. <https://www.businessinsider.com/chatgpt-ai-will-not-take-jobs-create-future-work-opportunities-2023-2> (accessed Mar. 14, 2023).
- [19] “Art and artificial intelligence collide in landmark legal dispute | Financial Times.” <https://www.ft.com/content/d691d599-3cdb-48d8-9824-9b2784a17d90> (accessed Mar. 09, 2023).
- [20] “Artificial intelligence is reaching behind newspaper paywalls,” *The Economist*. Accessed: Mar. 09, 2023. [Online]. Available: <https://www.economist.com/business/2023/03/02/artificial-intelligence-is-reaching-behind-newspaper-paywalls>
- [21] D. Susser, B. Roessler, and H. Nissenbaum, “Technology, autonomy, and manipulation,” *Internet Policy Rev.*, vol. 8, no. 2, Jun. 2019, doi: 10.14763/2019.2.1410.
- [22] J. L. Egelhofer and S. Lecheler, “Fake news as a two-dimensional phenomenon: a framework and research agenda,” *Ann. Int. Commun. Assoc.*, vol. 43, no. 2, pp. 97–116, Apr. 2019, doi: 10.1080/23808985.2019.1602782.
- [23] Y. Directorate-General for Education, *Guidelines for teachers and educators on tackling disinformation and promoting digital literacy through education and training*. LU: Publications Office of the European Union, 2022. Accessed: Jan. 30, 2023. [Online]. Available: <https://data.europa.eu/doi/10.2766/28248>
- [24] J. L. Egelhofer and S. Lecheler, “Fake news as a two-dimensional phenomenon: a framework and research agenda,” *Ann. Int. Commun. Assoc.*, vol. 43, no. 2, pp. 97–116, Apr. 2019, doi: 10.1080/23808985.2019.1602782.
- [25] “COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Artificial Intelligence for Europe (2018).” Accessed: Jan. 30, 2023. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN>
- [26] “What’s in a name?,” *CEPS*, Feb. 22, 2023. <https://www.ceps.eu/ceps-publications/whats-in-a-name/> (accessed Apr. 17, 2023).
- [27] E. Kapantai, A. Christopoulou, C. Berberidis, and V. Peristeras, “A systematic literature review on disinformation: Toward a unified taxonomical framework,” *New Media Soc.*, vol. 23, no. 5, pp. 1301–1326, 2021, doi: 10.1177/1461444820959296.
- [28] P. de P. Bak, J. G. Walter, and A. Bechmann, “Digital false information at scale in the European Union: Current state of research in various disciplines, and future directions,” *New Media Soc.*, 2022, doi: 10.1177/14614448221122146.

- [29] HLEG 2018. A multi-dimensional approach to disinformation: report of the independent high-level group (HLEG) on fake news and online disinformation. European Commission. Publications Office of the European Union. <https://blog.wan-ifra.org/sites/>.
- [30] UNHCR, 2021. Using Social Media in Community-Based Protection. A Guide, pp. 242.
- [31] C. Carrasco-Farré, "The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions," *Humanit. Soc. Sci. Commun.*, vol. 9, no. 1, pp. 1–18, 2022, doi: 10.1057/s41599-022-01174-9.
- [32] W. Shahid, Y. Li, D. Staples, G. Amin, S. Hakak and A. Ghorbani, "Are You a Cyborg, Bot or Human?—A Survey on Detecting Fake News Spreaders," in *IEEE Access*, vol. 10, pp. 27069- 27083, 2022, doi: 10.1109/ACCESS.2022.3157724.
- [33] C. Wardle, "FIRST DRAFT'S Essential Guide to Understanding information disorder," First Draft, no. October, p. 61, 2019, [Online]. Available: https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf?x47711.

2 SOCIO-DEMOGRAPHIC AND SOCIO-POLITICAL FACTORS AFFECTING CITIZEN (DIS)TRUST AND (DIS)ENGAGEMENT

2.1 WHY TRUST MATTERS WHEN LOOKING TO STRENGTHEN CRITICAL THINKING

When citizens encounter a piece of information, their levels of trust in the news in general and the news source specifically are likely to affect how they will treat the information. Here lies an interesting conundrum we will further explore: is trust in the news a good or a bad thing?

Next, when they decide to engage critically with the piece of information, trust in the 'alternative' information sources becomes key: what makes them trust these alternative sources above the original information? Another conundrum: what if citizens trust untruthful information?

Finally, in the process of critically assessing information, citizens will need to rely on their own critical skills to assess information. Ultimately, they will still need to trust a piece of information to be truthful, and for that, trust their own judgement.

Clearly, trust or distrust in the news is likely to affect the way citizens deal with information online. Hence, in the first part of this literature review, before looking at the cognitive processes that affect citizens' ability to critically assess information, we look at the socio-political factors shaping citizens' dispositions of trust or distrust towards information, and subsequently their capacity to critically engage with it. But first, we start by discussing how trust in the news has been conceived and measured in the literature and explore the ambivalence between trusting information and critically engaging with it.

2.2 THE STRUGGLE WITH TRUST IN NEWS

The ideal of an informed populace, one that comprises of logical, information-seeking, politically involved citizens, which is still the default attitude among media professionals, underlies the focus on fact-checking and critical thinking as a response to disinformation [1]. Even though this perspective has been challenged in the past, this viewpoint is prevalent among policymakers, scholars, and newsrooms [2]. Will this opinion still hold true in light of the "global democratic recession" [1] characterized by the rise in support for right-wing nationalist parties, a serious political crisis in the EU, and the unsettling political theatre of the former US president, among other things?

The lack of accountability in the media system is at the center of this crisis-like feeling. According to Van Aelst et al. [3], the emergence of partisan media poses a significant threat to the environment in which political information is produced because it creates "opportunity structures for selective exposure based on political attitudes and beliefs" as opposed to factual data. In its most radical formulation, this implies that facts and the truth are irrelevant. The current distributed media environment appears to further amplify individuals who practice "post-truth politics," which undermines the democratic public sphere and fosters social conflict [4].

At this point, trust comes into play. According to Strömbäck et al. [5], if the public does not use or trust the new information, it will be difficult to achieve the democratic ideal of an educated citizenry. Overall previous research suggests that media trust is associated with greater use of news media while media distrust is associated with greater use of non-mainstream news sources, but that the relationship between media trust and media use is quite modest [5].

We also know that news consumption is not always a rational and informative act driven by a desire to seek knowledge. It can also be motivated by factors like distraction, emotional attachment, and habit [6]. In turn, this can result in people consuming news they don't trust, either because their desire for information

outweighs their distrust [7] or because they prioritize convenience over trustworthiness [8]. This raises questions about the idea of an informed citizenry and the centrality of trust in news media for a democratic and factual society, or as Fisher [9] would state it:

“Relying on the assumption that the news consumer will interpret trust based on traditional conceptions of reliability and accuracy bound up in the ideal of the informed citizen, does not adequately accommodate how and why people are accessing news media.”

The kind of citizenship she is hinting at is one with “an orientation towards a public world, including politics and broader public issues, beyond matters of purely private concern” [10]. Such a view on citizenry adheres much more to Schudson’s [11] notion of a monitorial citizen.

Previous studies have primarily focused on how news contributes to people's sense of identity as democratic citizens in electoral politics and formal democratic settings (such as studies by Couldry et al. [10], Ekström et al. [12], and Hovden & Moe [13]). However, it is important to note that news can also serve as a means of connecting people on a cultural level, for instance within the workplace, neighborhood, or family.

2.2.1 The ambivalence of trust in news

As already touched upon in the introduction, trust as something to strive for in a healthy news ecosystem, is not as straightforward as we might think. As became clear in the previous section, to trust someone or something always requires a leap of faith, and this is not always reconcilable with the critical thinking expected from savvy news users. Let us unpack this tension to fully understand its ramifications for the development of a tool such as Titan.

2.2.2 To trust or not trust in news

As Strömbäck et al. [5] suggest, informative news media will hardly lead to the democratic ideal of an informed citizenry if citizens do not consume or do not trust the news. Indeed, trust in news media seems a prerequisite for people to be willing to accept the information media convey.

One condition for a trusted relationship between two parties to be achieved is the rational knowledge that the parties involved will perform a particular action [14]. In general, citizens make a rational assessment when they trust media: based on the public discourse on news media and the knowledge that the information if provided by experienced and trained professionals abiding by a deontological code, they have good reason to believe that the information provided is largely truthful.

A second condition, however, is an element of commitment, a leap of faith that momentarily brackets out the lack of rational knowledge we have on the fact that the said action will actually be performed [14]. It is this leap of faith that leads citizens to consume news – from media they trust at least – without questioning each story they encounter on an individual basis: they believe that the action of providing accurate information will follow. It is also this leap of faith that seems to conflict with the notion of critical thinking, or at least how it is often put forward in media literacy, namely as the reflex to not take for granted the information offered by the media.

An interesting conundrum indeed. Citizens need to believe the news and news media for them to fulfil their public role, but at the same time are thought to always be critical towards it. Trust is important, but it is also necessary to be cautious with trust and not blindly trust others. Being too trusting can lead to an unequal and unbalanced relationship, so it's important to maintain a level of skepticism towards others. The key here is the following: distrust isn't the opposite of trust, but an essential part of its practice [15].

This shows how ambivalent the notion of trust in news is. Admittedly, a certain level of distrust is what is fueling critical thinking. But we know from research that there is a strong link between low trust in the media and a preference for alternative news sources [16]. Titan will need to consider how to help citizens to think critically while at the same time not undermining the relations of trust between citizens and truthful news sources.

2.2.3 *To trust your gut*

The idea of remaining critical towards information is a widely supported in media literacy education. When teachers teach students to find reliable information, they teach them to question the source of the information and consider factors such as the credibility of the publisher and potential biases of the author. One important perverse effect of promoting critical thinking, however, is that it can lead citizens to trust their gut above truthful sources. Danah boyd [17] points out in *Did media literacy backfire? How the above-mentioned approach assumes that reputable sources like the New York Times, scholarly journals, and experts with advanced degrees are generally considered trustworthy. But this is not always the case, for example amongst conservatives who see the mainstream media as too liberal, amongst religious people that believe in God above science, or amongst working people that see degrees as a weapon of oppressing elites.*

The combination of building media literacy around the notion of it being one's personal responsibility to question the information they are receiving and finding out the truth for themselves and a distrust of media and scientific sources, according to boyd [17] [18], creates a dynamic that is potentially directing citizens towards untruthful sources. Left to their own devices, citizens are quick to accept the information that fits their belief system and quick to trade evidence, reason and facts for experience, faith and alternative evidence. To cite boyd [18]:

“No matter what worldview or way of knowing someone holds dear, they always believe that they are engaging in critical thinking when developing a sense of what is right and wrong, true and false, honest and deceptive. But much of what they conclude may be more rooted in their way of knowing than any specific source of information. If we're not careful, “media literacy” and “critical thinking” will simply be deployed as an assertion of authority over epistemology.”

For Titan, this perverse effect is important to consider, as the premise of the project is precisely to help citizens assess information through critical thinking. A reflection is needed on how to confront users with their own dispositions and biases in order for Titan not to become an instrument that is strengthening existing biases.

2.2.4 *To trust untruthful information*

A consequence of people following their gut feelings when assessing information is that they might start to trust untruthful information. A good example of the perverse effect of a too trusting relationship with information providers are the many alternative media, populist politicians and fearmongers on any side of the political spectrum that all mastered ways to gain people's trust, regardless of the truth [19].

Here too, we must realize that the way people engage with news is not always an instrumental practice driven by a rational, informative and selective orientation towards the media, but can also be more ritual, driven by distraction, affection and habits [6]. This can even lead to people using news they do not trust, e.g., when their urge to know the latest developments of a story is higher than their distrust in the news [7] or their orientation towards convenience leads them to choose less trusted over more trusted news [8].

More broadly speaking, it is important to acknowledge the importance of emotion in how citizens deal with information, which is certainly not a new element, but one that can easily be overlooked, especially when promoting critical thinking. As Beckett and Deuze [20] argue, research tells us that people respond to emotions, rather than ideas or facts. Networked technologies have further amplified this, as various kinds of

information, ranging from professional journalism to opinions of peers, now inhabit the same digital spaces, best exemplified by the social media news feed. In such a news feed, Papacharissi [21] suggests, emotion, drama, opinion, and news are blended in a manner that departs from the conventional deliberative logic and aligns with emotion.

Titan, in the way it seeks to help people in thinking critically, will need to be aware of the affective dimension of networked publics, especially if Titan aims to be accessible to people right from the digital spaces, they reside in.

2.2.5 *The good the bad and the trusted*

When it comes to news and information, trust is an ambiguous concept. Trust and distrust are certainly not opposites. Trust without distrust can lead to blind trust in untruthful news sources. Distrust without trust can lead to disengagement from news, and the public sphere altogether. The best approach lies somewhere in the middle, and we actually see this reflected in people's news use.

Schwarzenegger [22] talks about pragmatic trust to describe how "users pragmatically confide in specific news sources in their repertoire, even if skeptical of media to varying degrees". People realize that they need information, and that they cannot question everything all the time. The author further shows that this pragmatism applies to various degrees, ranging from naïve to informed. The latter's critical stance is based on knowledge of the news production process and the biases it can entail, the former's on the lack of such knowledge and a more gullible attitude towards institutions and authorities. Also interesting, is that Schwarzenegger sees pragmatic trust emerge amongst both mainstream and alternative media focused news users, which shows that citizen turning to alternative media should not be put away as uncritical.

On the contrary, following the argumentation of boyd [17] [18] laid out above, many of those turning to alternative media might come to that point because of critical thinking. They follow what they have been told through media literacy education: be critical and do your own research. When developing critical thinking programs or tools, it can therefore be important to picture your ideal user as a twin: the more conventional one committed to a way of knowing that is rooted in evidence, reason and facts, and the more alternative one committed to another way that is rooted in experience, faith and non-scientific evidence. They diverge in the way they come to accept information, but they are equal in both considering their reasoning to be a form of critical thinking.

Again, referring to danah boyd [18], it is important to realize that critical thinking can also be weaponized by anti-democratic forces, e.g. through what is known as "red pilling": "[i]n some online communities, taking the red pill refers to the idea of waking up to how education and media are designed to deceive you into progressive propaganda. In these environments, visitors are asked to question more. They're invited to rid themselves of their politically correct shackles. There's an entire online university designed to undo accepted ideas about diversity, climate, and history. Some communities are even more extreme in their agenda. These are all meant to fill in the gaps for those who are opening to questioning what they've been taught."

It is therefore important to keep in mind when developing critical thinking tools which would qualify as media literacy education, that such approach has long focused on personal responsibility. This can not only imbue individuals with a false sense of confidence in their skills, but also put the onus of monitoring media effects on the audience, rather than media creators, social media platforms, or regulators [23]. Titan will need to take into consideration this ambivalence when developing its AI platform.

2.3 SOCIO-DEMOGRAPHIC AND SOCIO-POLITICAL FACTORS AFFECTING TRUST IN / DISCONNECTION FROM NEWS

The literature review already showed that trust, distrust, and critical thinking are particularly ambiguous in themselves and relationally within the context of news and media. Awareness of that ambiguity is the first step in developing tools to enhance critical thinking.

At the same time, it does not stop there. Indeed, research shows significant heterogeneity between different socio-demographic groups regarding trust, distrust, and likeliness to believe different types of news and information (see, for example: Bedard & Schoenthaler [24]). Moreover, developing an AI tool to enhance critical thinking should also consider the underlying socio-demographic dynamics regarding using such a tool.

2.3.1 *The socio-demographics of news and information consumption*

Before delving deeper into the heterogeneity regarding trust in news and information according to socio-demographic positions, looking at socio-demographic dynamics within news readership is relevant. For instance, Edgerly [25] categorizes different media users based on age, gender, race, education, political party identification (in the US political context), and perceived news bias.

First, the online-only group is typically younger and more male. But race and education do not significantly impact the likelihood of belonging to the online-only group. The group that primarily consumes news in the form of print and television is typically slightly older, more female, white, Democrat/Independent, and has a lower perceived media bias. The conservative-only group (which only wants to hear politically conservative voices regardless of the medium) is typically older, more male, white and Republican. And perhaps more interestingly, this group has a high perceived media bias. Finally, there are also the news omnivores. This last group is more likely to be non-white and more likely to identify as Democrat/Independent. Lastly, so-called news avoiders are statistically likelier to be non-white and less educated. In addition, a news avoider is 2.47 times more likely to be Republican than Democrat or Independent. At the same time, other research points to a gender gap in news avoidance: "Even in wealthy post-industrial countries where gender equity has markedly improved, women are still significantly more likely to say they avoid news, particularly concerning politics" [26].

Hence, it quickly becomes apparent that news and information consumption are closely linked to socio-demographic factors. This conclusion is repeatedly confirmed in research, with an additional important role attributed to socioeconomic position (own income and/or that of parents) [27] [28]. Similarly, research by Reuters Institute [29] reveals: "Lower social grade individuals on average use fewer sources of online news, are less likely to go directly to news organizations for news online and are consequently more reliant on the distributed discovery of news via social media and search engines."

2.3.2 *The socio-demographics of media trust*

The subsequent question is whether the same socio-demographic dynamics play into different levels of trust in consumed news and information. For example, Strömbäck et al. [5] point out that previous research suggests that media trust is associated with more news media use, and media distrust is associated with more use of non-mainstream news. Given the previous analysis regarding overall news use, there is thus already an indication that socio-demographic factors also play a role regarding trust in news. Knowing non-mainstream news is more often consumed by people with lower socioeconomic positions [28], a hypothesis thus seems to be forming that people with lower socioeconomic positions have less trust in news. However, rejecting this hypothesis based on research, Kalogeropoulos et al. [30] find that both the use of mainstream news and the use of alternative news correlate with higher trust in news. But interestingly, they also find that using social media as a primary news source correlates with lower trust in news. The following figure provides an overview of these results [30]:

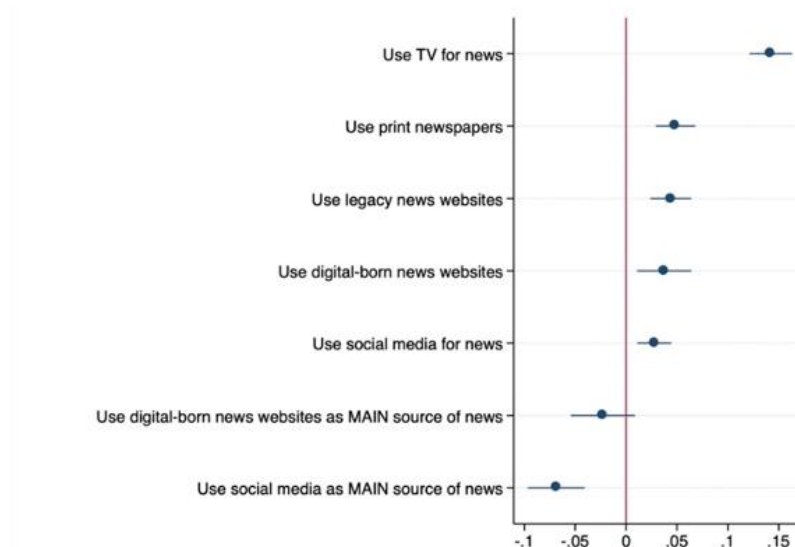


Figure 1. Effects of using different types of news on trust in news. Coefficients depicted with 95% confidence intervals. The model includes individual-level controls (Table 1).

Figure 2.1: Effects of using different types of news on trust in news. Source [30]: Kalogereopoulos, Suiter, Udris and Eisenegger (2019 p.3682)

Table 1. Individual-Level Factors Predicting Trust in the Media (Multilevel Models).

Trust in news media	Model 0	Model 1
Age		.003*** (.000)
Gender (female)		.071*** (.009)
Education		-.019*** (.004)
Income		.019** (.006)
Internal political efficacy		.029*** (.005)
Ideological strength		-.053*** (.011)
Interest in news		.068*** (.006)
Use of TV for news		.142*** (.010)
Use of newspapers/magazines for news		.048*** (.009)
Use of legacy websites for news		.043*** (.010)
Use of digital-born websites for news		.037** (.013)
Use of social media for news		.027*** (.008)
Count of offline sources used		.006*** (.001)
Count of online sources used		-.007*** (.001)
Use of digital-born websites as the MAIN news source		-.022 (.016)
Use of social media as the MAIN news source		-.068*** (.014)
Perceptions of outside influences		-.235*** (.008)
Constant	3.14***	3.44*** (.073)
Individual-level variance	.989	.729
Country-level variance	.034	.017

Note. Unstandardized coefficients. Robust standard errors in parentheses. $N = 60,579$. Countries = 35.
 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 2.1: Individual-level factors predicting trust in media. Source [30]: Kalogereopoulos, Suiter, Udris, and Eisenegger (2019 p.3681)

Regarding the socio-demographic factors (Age, Gender, Education, Income) in the multi-level model shown above, we see that age, female, and income positively correlate with trust in the media, which is at 99% of confidence or even more. Education, in turn, is negatively correlated (at 99% of confidence) with trust in the media. These results are also confirmed in research by Lee [31]:

Table 1. Factors Predicting Trust in the News Media

Dependent variables	Model 1	Model 2	Model 3	Model 4 ("Clean")
	β	β	β	β
Block 1				
Age	.00	.01	.00	
Gender (male)	-.01	-.01	.01	
Race (White)	-.04	-.04*	.00	
Education	.02	.00	.00	
Income	-.07**	-.08**	-.04	
Block 2				
Political efficacy		.11***	.03	
Political interest		-.04	-.02	
Media exposure		-.00	-.01	
Block 3				
Economy evaluation			.02	
Political trust			.23***	
Personal trust			.03	
Liberal-conservative ideology			-.06*	
Democrat-Republican partisanship			-.19***	
R^2	.00	.02	.11	

N = 2,203.

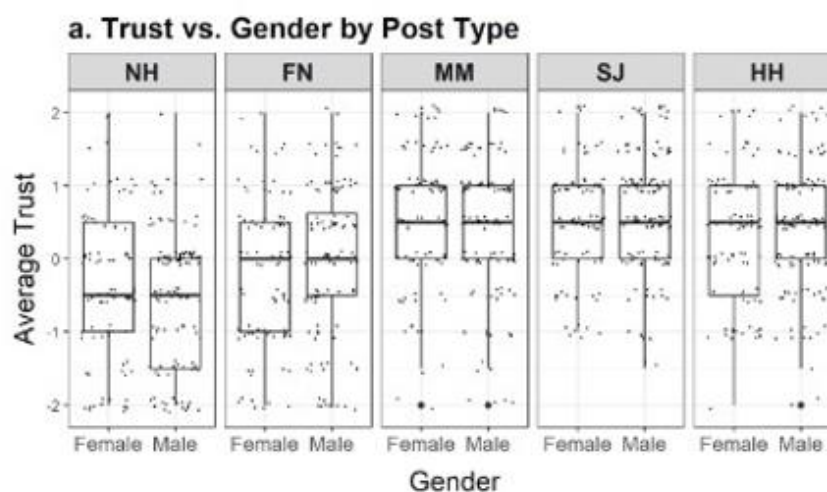
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 2.2: Factors predicting trust in News Media. Source [31] Lee (2010 p.13)

In Model 2 of the regression (shown above), we see that being white has a negative influence (at 95% of confidence) on media trust and, contrary to other research, we see that income has a negative impact (at 99% of confidence) on media trust. When one also controls for politics-related factors (Model 3), this significant effect of income disappears completely. This shows that the assumed impact of income on media trust might be partly due to not controlling for political trust.

2.3.3 It's the politics, stupid!

So, what can be concluded from this? Based on the previous results, it thus seems that most socio-demographic indicators have little or no impact on trust in media. On the other hand, political trust and orientation play an essential role in media trust. Indeed, this hypothesis is confirmed in research by Verma et al. 2018 [32]: MWU tests (shown below) in that study indicate no significant differences between men and women on trust in different news sources (denoted here as 'NH', 'FN', 'MM', 'SJ', 'HH'). Interestingly, there are no significant differences in age and educational attainment on trust in different news sources. And in line with the earlier hypothesis, the study shows that political leaning significantly impacts trust distribution.



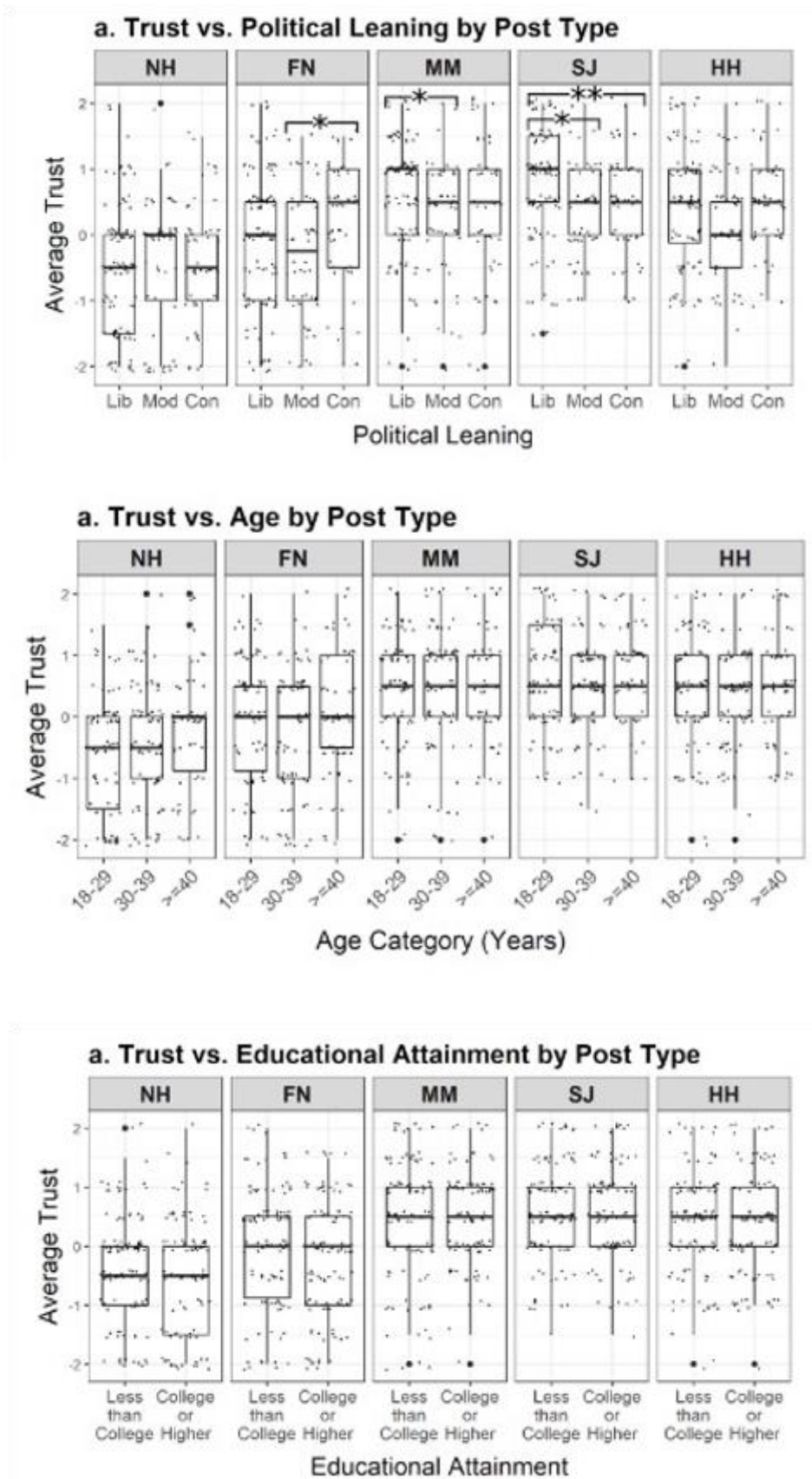


Figure 2.2: Source [32]: Verma, Fleischman, and Koltai (2018, pp-528-530)

Thus, based on research, we can confidently say that political trust (and, in some research also, political orientation) is related to media trust. Given this, it is also necessary to test whether there is an interaction effect: Is the relationship between political trust and political orientation on the one hand and media trust on the other different for different age groups? Brosius et al. [33] test these interaction effects and use the following classification of age groups (p. 655): “Generation Z (n = 298) includes those between 17 and 24 years at the time of the survey (2019), Millennials (n = 1542) those between 25 and 39 years, Generation X (n = 2060) those between 40 and 54 years, and Baby Boomers (n = 2537) those between 55 and 74 years.”

The results, shown below, show that the effect of political trust on media trust is almost the same across all age groups (i.e. there is no interaction effect). Regarding the relationship between political orientation and media trust, there are two notable findings: the study by Brosius et al. shows, on the one hand, that across all age groups, there is no apparent effect of political orientation on media trust, but on the other hand, politically-left-minded individuals from Gen Z on average have less trust in media (i.e. there is an interaction effect). The authors say the sample's composition could explain this effect, so more research is needed. Additionally, the study shows that there are also interaction effects for perceptions of bias and inaccuracy, as well as for political interest.

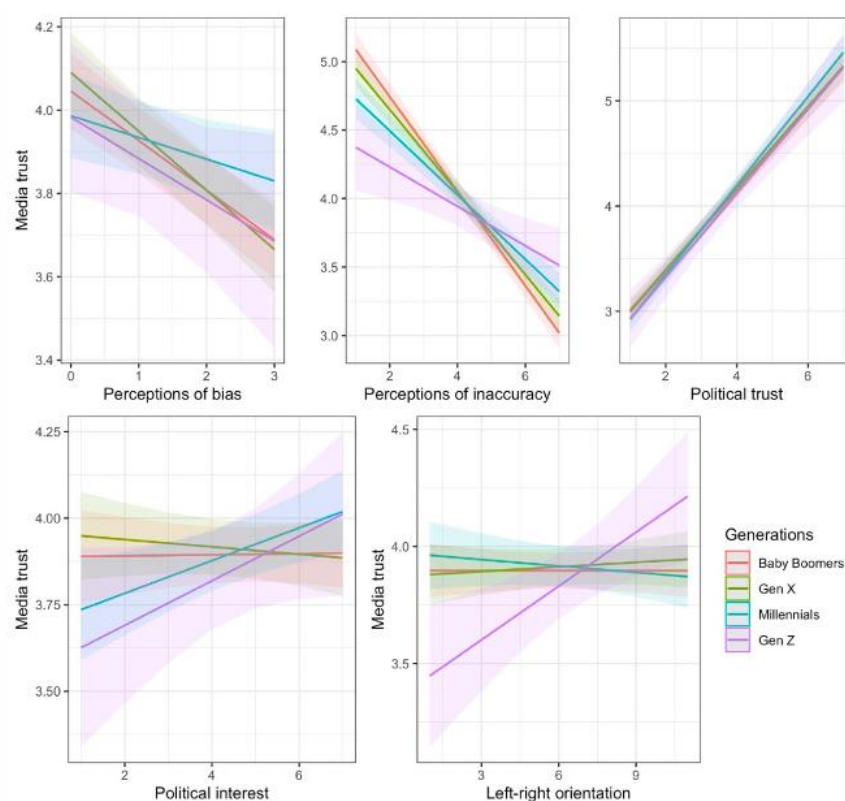


Figure 2.3: Source [33] : Brosius, Ohme, and De Vreese (2021, p.658)

A central conclusion thus follows from all the preceding analyses: rather than socio-demographic factors, political trust is mainly strongly associated with media trust. At the same time, it is impossible to give exclusions about the direction of causality, given that the relationship "could follow the logic of a reinforcing spiral" [33]. In contrast, we see that the form of news and information consumption does get influenced by socio-demographic factors, with individuals' socioeconomic position (sometimes referred to as 'social order') in particular. One finding that is undoubtedly relevant within this report's context is that individuals who avoid news are statistically more likely to place themselves on the right side of the political spectrum.

2.3.4 Zooming in on news avoidance and disengagement

The democratic ideal of a free, self-governing, and informed citizenry inherently assumes that citizens are politically informed and politically engaged [5]. An essential prerequisite for achieving this state of being informed and engaged is exposure through news coverage to events shaping society. Therefore, increasing news avoidance and civic disengagement [34] are worrisome trends within the context of the democratic ideal.

Research shows that so-called news avoiders are statistically more likely to be non-white, politically right-wing-oriented, and less educated. Still, this analysis does not give a complete picture, for which a more in-depth look at the different types of 'drivers' of news avoidance and disengagement is needed [25].

But first, how does academia define news avoidance? Although definitions differ slightly, there seems to be a consensus around the existence of two types of news avoidance: *intentional* and *unintentional* avoidance of news (see Skovsgaard & Andersen [35]; Van den Bulck [36]). Intentional news avoidance refers to an act in which individuals consciously avoid the news for a variety of reasons, unintentional news avoidance, on the other hand, is mainly a consequence of dynamics at play within the market logic of media (f.e. an increasing supply of entertainment responds to a greater demand for that genre, and these dynamics lead to specific individuals being less exposed to news). Focusing on intentional news avoidance, Villi et al. [37] seek to contextualize news avoidance by going beyond individual-level characteristics (i.e. socio-demographic factors), and hence they identify two contextual drivers of news avoidance.

First, cognitive drivers of news avoidance consist of a persistent and high frequency of so-called 'high-profile' news items (mainly politically related news), ultimately leading to a perception of being news 'overloaded'. Villi et al. [37] argue that this higher frequency can be caused by certain actual events (f.e. the presidential election of Trump) but also by the 'always-on'-state of social media. Chan et al. [38] also refer to this, using the concept 'news fatigue'.

Second, emotional drivers of news avoidance mainly result from an individual's attempt to protect himself from emotional reactions (fear, despair, anger, distrust) towards negative news. News themes that typically trigger such reactions include car accidents, terrorist attacks and natural disasters. Villi et al. [37] also show that the cognitive drivers of news avoidance are more country-specific (i.e. there is interaction: the effect of repetitive news coverage on news avoidance is different for different countries). Still, the emotional drivers regarding the impact on news avoidance are universal.

In addition, there is another contextual driver, one that at first glance seems purely socio-demographic but which in reality has an essential socio-cultural context attached to it: the role of gender in news avoidance [34] [26]. Toff and Palmer [26] identify three causes of this gender gap, all related to role socialization and structural inequalities: "news-is-for-men" perceptions", labor division within households along gender lines, and the physical and emotional toll of caretaking (falling predominantly on women).

Finally, factors related to the country's media system also influence the frequency of news avoidance. In particular, research shows that stronger public service broadcasting correlates with less news avoidance since "the quality of news from public service broadcasting is generally higher and more trusted" [34]. Also, Chan et al. mention that media markets that are less inclusive to mass audiences will generate more potential entry points for fake news. This higher potential production and dissemination of fake news may increase a tendency to avoid news already at the individual level. Hence, Chan et al. [34] find that individual-level characteristics such as news fatigue interact with media market characteristics: the effect of news fatigue on news avoidance is different in different media markets.

Table 1. Multilevel models predicting news avoidance and news authentication.

	News avoidance			News authentication		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed effects – Intercept	1.951***	1.967***	1.968***	.854***	.885***	.885***
<i>Level 2 – Media system</i>						
Press market		.008	.003		.043	.043
Political parallelism		.006	.035		.016	.010
Journalistic professionalism		-.057	-.058		-.090*	-.100*
Public service broadcasting		-.075***	-.070***		-.080**	-.092***
<i>Level 1 – Individual</i>						
Gender		.044***	.045***		-.074***	-.074***
Age		-.005***	-.005***		-.003***	-.003***
Education		-.000	-.000		.042***	.042***
News interest		-.098***	-.098***		.152***	.152***
News trust		-.071***	-.071***		-.036***	-.036***
News use frequency		-.007	-.007		.040***	.040***
Social media use for news		-.035***	-.035***		.289***	.289***
Social media as main news source		.126***	.126***		-.107***	-.107***
Concern with fake news (CFN)		.002	.001		.140***	.139***
News fatigue (NF)		.269***	.268***		.006	.007
CFN × NF		.042***	.042***		.003	.003
<i>Two-way cross-level interaction</i>						
CFN × Press market			-.011			-.003
CFN × Political parallelism			-.013			-.008
CFN × Journalistic professionalism			.009			-.009
CFN × Public broadcasting			.011			-.036**
NF × Press market			-.019			-.007
NF × Political parallelism			-.063**			.003
NF × Journalistic professionalism			-.001			.007
NF × Public broadcasting			.004			.005
<i>Three-way cross-level interaction</i>						
CFN × NF × Press market			-.015*			-.008
CFN × NF × Political parallelism			.003			.021*
CFN × NF × Journalistic prof.			.016			.024**
CFN × NF × Public broadcasting			-.010*			.003
<i>Random effects</i>						
Residual	.848***	.709***	.708***	.758***	.673***	.672***
Intercept	.026**	.004*	.003*	.037**	.005**	.001**
Concern with fake news		.001***	.001*		.002*	.000
News fatigue		.004***	.001*		.000	.000
N	31940	31491	31491	32819	32127	32127

*** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 2.3: Source [38]: Chan, Lee and Chen (2022, p.10)

So, what can be concluded from this? From the reasoning of a democratic ideal, increasing news avoidance is a worrying evolution as it may lead to a less informed and engaged citizenry (see for example Strömback [5]). Research shows that media trust (or lack thereof) is essential in understanding news avoidance. Still, at the same time, one should be careful not to reduce this to false essentialism that states that individuals avoid news because they don't trust it or because they are less educated or the like. Indeed, the analysis above shows that news avoidance can also be driven by cognitive, emotional, socio-cultural, and even media market factors unrelated to media trust. To address news avoidance and thus understand it in the first place, it is necessary to let go of the sometimes-assumed causal link between media trust and news avoidance and nuance this link using contextual factors (cognitive, emotional, cultural, economic).

2.3.5 Zooming in on the likeliness to believe fake news and disinformation

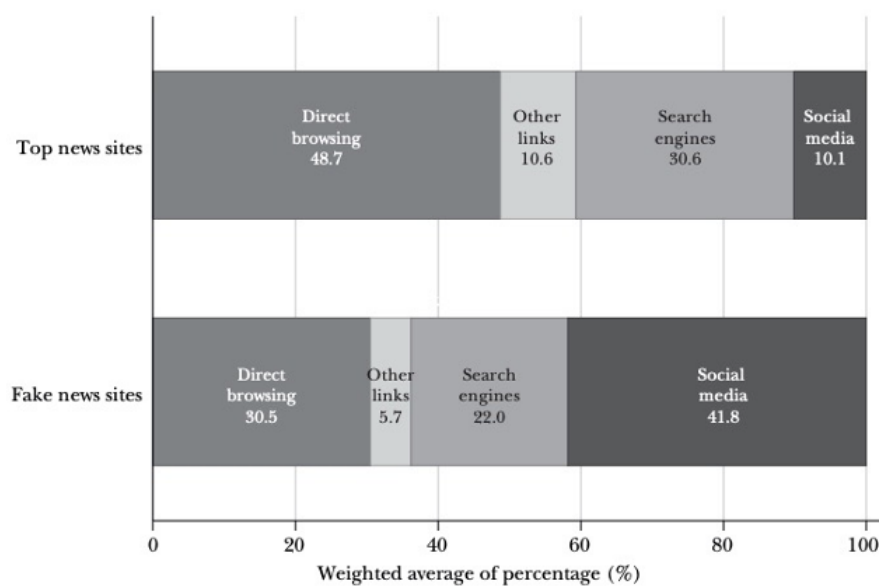
As reasoned above, research suggests that news avoidance can be detrimental to developing a well-informed, politically engaged citizenry. But, at the same time, the consumption of news can be detrimental when fake news or disinformation is believed to be truthful or when truthful news is believed to be false (see for example: Mason et al. [39]; Morgan [40]; Waisbord [41]).

However, the goal of this section is not to argue why incorrect belief in truthful or false news is a challenge for democracy, as this is the report's starting point. But the goal is here to analyze the underlying socio-demographics of being able to identify false news as being false and to identify truthful news as being truthful (i.e. 'correct belief').

Before doing so, it's relevant to examine *how* fake news is accessed: through which mediums run access to fake news? An influential study of the 2016 US presidential elections by Allcott and Gentzkow [42] finds a

significant difference in the centrality of social media as a medium through which news is accessed. Only 10,1% of visits to the leading US news websites can be attributed to social media, while social media contributes for 41,8% of visits to fake news sites. This finding has significant consequences, as Allcott and Gentzkow [42] write: “This suggests that people who get news from Facebook (or other social media) are less likely to receive evidence about the true state of the world that would counter an ideologically aligned but false story.” Yet, at the same time, social media enables increased exposure to news coverage, thus potentially being beneficial to democracy. Hence, referring to this ambiguity, Shu et al. [43] write: “Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news from social media. On the other hand, it enables the wide spread of “fake news”, i.e., low-quality news with intentionally false information. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society.”

Figure 3
Share of Visits to US News Websites by Source



Note: This figure presents the share of traffic from different sources for the top 690 US news websites and for 65 fake news websites. “Other links” means impressions that were referred from sources other than search engines and social media. “Direct browsing” means impressions that did not have a referral source. Sites are weighted by number of monthly visits. Data are from Alexa.

Figure 2.4: Share of Visits to US News Websites by Source. Source [42]: Allcott and Gentzkow (2017, p.222)

But then, who are the individuals less able to distinguish true from false news? Knowing that political orientation is essential for studying media trust and attitudes towards different types of news, controlling for *Democrat* and *Republican* [42] provides relevant insights. Results (shown below) in column 3 indicate that Republicans are less likely than Democrats to identify false news as false. Reversed, in column 4, Democrats are less likely than Republicans to label true news as truthful. Hence, Allcott and Gentzkow [42] state: “These results suggest that in our data, Republicans were not generally worse at inference: instead, they tended to be more credulous of both true and false articles.” Besides, education (column 4) is positively associated with correct beliefs about news headlines, but age does not affect the ability to distinguish real from false news. Interestingly, in contrast with the findings of Allcott and Gentzkow [42], Rampersad and Althiyabi [44] do find an effect of age on acceptance of fake news: age is positively associated with fake news acceptance, meaning that younger individuals will typically more easily recognize fake news as being such.

Table 1
What Predicts Correct Beliefs about News Headlines?

	(1)	(2)	(3)	(4)	(5)
Democrat	0.029 (0.020)	-0.004 (0.023)	0.028 (0.019)	-0.010 (0.021)	0.015 (0.013)
Republican	-0.024 (0.024)	0.040 (0.027)	-0.037* (0.020)	0.021 (0.023)	-0.018 (0.014)
ln(Daily media time)			-0.002 (0.007)	0.042*** (0.008)	0.013*** (0.004)
Social media most important			-0.066*** (0.025)	0.065*** (0.024)	-0.023 (0.016)
Use social media			0.014 (0.030)	-0.023 (0.038)	0.002 (0.019)
Social media ideological segregation			-0.027 (0.036)	0.028 (0.046)	-0.008 (0.024)
Education			0.014*** (0.004)	0.004 (0.004)	0.011*** (0.003)
Undecided			-0.011 (0.017)	0.006 (0.022)	-0.005 (0.013)
Age			0.002*** (0.000)	0.000 (0.001)	0.002*** (0.000)
N	12,080	6,040	12,080	6,040	18,120
p-value (Democrat = Republican)	0.029	0.124	0.004	0.207	0.035
Articles in sample	False	True	False	True	All

Note: This table presents estimates of a regression of a dependent variable measuring correct beliefs about headlines on individual characteristics. Columns 1 and 3 include only false headlines, columns 2 and 4 contain only true headlines, and column 5 contains all headlines. All columns include additional demographic controls: income, race, and gender. "Social media most important" means social media were the respondent's most important sources of election news. "Social media ideological segregation" is the self-reported share (from 0 to 1) of social media friends that preferred the same presidential candidate. "Undecided" is an indicator variable for whether the respondent decided which candidate to vote for less than three months before the election. Observations are weighted for national representativeness. Standard errors are robust and clustered by survey respondent.

*, **, *** indicate statistically significantly different from zero with 90, 95, and 99 percent confidence, respectively.

Table 2.4: What Predicts Correct Beliefs about News Headlines?. Source [42]: Allcott and Gentzkow (2017, p.229)

Research by Rampersad and Althiyabi [44] confirms that education is negatively associated with accepting fake news (i.e., higher-educated individuals have a higher probability of being able to distinguish real from false news correctly). Similarly, Koch et al. [45] find that education and analytical thinking, taken together under the variable *cognitive style*, are negatively associated with fake news amplification.

Similarly, Bedard and Schoenthaler [24] conduct a study in which they research if media users' ability to distinguish satire from fake news is related to socio-demographic factors such as age, education political affiliation, and gender. Their results show that older *Millennials* and individuals belonging to *Generation X* are, on average, better able to identify fake news correctly. This correct identification improves, on average, even further when the individual is a woman rather than a man. And, in line with previous research, higher-educated and politically Democrat-oriented individuals score better in terms of fake news identification.

Regarding fake news amplification, Arin et al. [46] distinguish between *deliberate* and *accidental* fake news sharing. First, results show that accidental sharing of fake news is much more prevalent than deliberate sharing of fake news. And second, this accidental type of sharing false information decreases with age and is less prevalent among politically left-leaning news users. Interestingly, deliberate fake news sharing is more common among younger news users. Lastly, as previous research shows, Arin et al. [46] also find that older, high-income, politically left-leaning news users can better detect fake news.

2.3.6 *Ceteris paribus* in AI solutions, or the lack thereof

Within the public sphere today, largely thanks to the success of ChatGPT, there is much debate about the impact of artificial intelligence (AI) on various areas of society. The views within these debates range from pure techno-pessimism to a techno-determinism that believes AI is the solution to many problems. Notably, often socioeconomic inequality is a theme around which many of these AI debates center.

Surprisingly, in contrast with AI’s potential impact on equality, much less attention is paid to inequalities *within* the perception of, access to, and use of AI. Hence, supposing use of AI has potential beneficial effects on both an individual and collective level, significant socio-demographic differences in access to AI can thus imply inequality.

Interestingly, Park and Jones-Jang [47] find no statistically significant effect of any socio-demographic factor (age, gender, income, education) on the intention to use AI, on the perceived accuracy of AI, on the perceived authenticity of AI, or on the perceived credibility of AI.

Table 3 Regression analyses for direct and indirect relationships

	PEOU	PU	Intention to use	AI Accuracy	AI authenticity	AI credibility
IV: security	- 0.08 (0.02) **	- 0.01 (0.02)	- 0.01 (0.00)	- 0.13 (0.04) **	- 0.17 (0.05) **	- 0.17 (0.05) **
IV: surveillance	- 0.01 (0.02)	- 0.03 (0.02)	- 0.02 (0.00) **	- 0.15 (0.04) **	- 0.19 (0.05) ***	- 0.16 (0.05) **
M1: PEOU	-	0.61 (0.03) ***	0.07 (0.01) ***	0.18 (0.07) *	0.16 (0.08) *	0.11 (0.08)
M2: PU	-	-	0.22 (0.01) ***	0.58 (0.07) **	0.61 (0.08) ***	0.65 (0.08) ***
M3: intention	-	-	-	1.58 (0.18) ***	1.67 (0.20) ***	1.57 (0.20) ***
Age	- 0.02 (0.00) **	0.00 (0.00)	- 0.00 (0.00)	- 0.02 (0.01) **	- 0.01 (0.01)	- 0.02 (0.01)
Female	- 0.21 (0.19)	0.08 (0.18)	- 0.12 (0.06)	- 0.58 (0.33)	- 0.55 (0.36)	- 0.71 (0.36)
Income	- 0.19 (0.05) ***	- 0.04 (0.05)	- 0.02 (0.01)	0.10 (0.09)	0.01 (0.10)	0.11 (0.10)
Education	0.13 (0.09)	- 0.00 (0.09)	0.00 (0.03)	0.16 (0.16)	0.18 (0.18)	0.18 (0.18)
AI device owned	1.48 (0.22) ***	1.28 (0.21) ***	0.44 (0.08) ***	0.69 (0.41)	0.84 (0.44)	0.61 (0.45)
Constant	11.90 (0.73)	3.73 (0.80)	0.52 (0.29)	6.89 (1.48)	6.11 (1.61)	6.65 (1.62)
Adjusted R	0.12	0.39	0.53	0.48	0.46	0.44
F statistic	17.03 ***	68.54 ***	108.66 ***	79.01 ***	70.53 ***	66.48 ***

Unstandardized coefficients (*B*) reported. Standard errors in parentheses

p* < 0.05, *p* < 0.01, ****p* < 0.001, *M* mediator. *F*-statistic, constant, and effects of covariates reported pertain to the model with IV of security concern. Separate analyses were run for surveillance concern

Table 2.5: Regression analyses for direct and indirect relationships. Source [47] Park and Jones-Jang (2022,p.8)

However, Antes et al. [48] conduct a similar study and conclude that men are typically associated with more openness towards AI, lower concern regarding AI, and lower attributed benefit to AI. In line with expectations, individuals with higher levels of household income are associated with more openness towards AI, less concern, and higher perceived benefit. In addition, full-time employed individuals are more open towards and less concerned about AI and attribute higher benefit to AI. Education, finally, is positively (albeit very mildly) correlated with all three dependent variables: openness, concern, and benefit.

Table 4 Correlations of openness, concern, and benefit scores with all study variables

	Openness		Concern		Benefit	
	r	95% CI	r	95% CI	r	95% CI
<i>Socio-demographics</i>						
Age	-.12	[-.18, -.06]	.06	[.00, .12]	-.03	[-.09, .03]
Sex (1 = Male, 0 = Female)	.10	[.04, .16]	-.20	[-.26, -.14]	-.03	[-.09, .04]
Race (1 = White, 0 = Non-White) ^a	-.05	[-.11, .01]	.01	[-.05, .07]	-.08	[-.14, -.02]
Ethnicity (1 = Latino, 0 = non-Latino)	.06	[.00, .12]	-.09	[-.15, -.03]	-.02	[-.08, .04]
Household income	.07	[.01, .13]	-.08	[-.14, -.02]	.07	[.01, .13]
Community type	.06	[.00, .12]	-.06	[-.12, .00]	.01	[-.05, .07]
Employment status ^b	.17	[.11, .23]	-.18	[-.24, -.12]	.05	[-.01, .11]
Education	.04	[-.02, .10]	.03	[-.03, .09]	.01	[-.05, .07]
<i>Health status and access</i>						
Health status	.08	[.02, .14]	-.12	[-.18, -.06]	-.02	[-.08, .04]
Healthcare location ^c	.03	[-.03, .09]	-.01	[-.07, .05]	.02	[-.04, .08]
Healthcare choice ^d	.08	[.02, .14]	-.06	[-.12, .00]	.11	[.05, .17]
Health insurance (1 = Yes, 0 = No)	.09	[.03, .15]	-.10	[-.16, -.04]	.05	[-.01, .11]
Healthcare satisfaction (n = 735)	.11	[.04, .18]	-.07	[-.14, .00]	.14	[.07, .21]
<i>Psychosocial variables</i>						
Health System Trust Index	.27	[.21, .33]	-.27	[-.33, -.21]	.21	[.15, .27]
Trust in technology	.41	[.36, .46]	-.21	[-.27, -.15]	.41	[.36, .46]
Faith in technology	.38	[.32, .43]	-.10	[-.16, -.04]	.46	[.41, .51]
Conscientiousness	.02	[-.04, .08]	.11	[.05, .17]	.15	[.09, .21]
Agreeableness	.08	[.02, .14]	.11	[.05, .17]	.20	[.14, .26]
Extraversion	.08	[.02, .14]	-.12	[-.18, -.06]	.04	[-.02, .10]
Emotional stability	.08	[.02, .14]	-.06	[-.12, .00]	.07	[.01, .13]
Openness (trait-based)	.07	[.01, .13]	.07	[.01, .13]	.05	[-.01, .11]
Social conservatism	-.01	[-.07, .05]	-.10	[-.16, -.04]	.05	[-.01, .11]
Economic conservatism	-.06	[-.12, .00]	-.06	[-.12, .00]	.02	[-.04, .08]

N = 936 (except as noted for specific variables in Tables 1 and 2)

^a Participants who selected any race other than White, or in addition to White, were classified as Non-White for purposes of this analysis

^b 1 = full-time employment, 0 = all other options

^c 1 = doctor office or private clinic, 0 = all other options

^d 1 = great or some choice; 0 = little to no choice

Table 2.6: Correlation of openness, concern, and benefit scores with all study variables. Source [48]: Antes, Burros, Sisk, Schuelke, Keune, and DuBois (2021, P.10)

Hence, it becomes clear that further research on socio-demographic factors tied to access to and use of AI is needed. As Følstad et al. [49] conclude: “Moreover, research still needs to assess whether there are systematic differences in the adoption and usage of chatbots driven by socio-demographic characteristics.” Yet, based on results and debates today, differences in attitudes and practices towards AI are mostly explained by individuals’ socioeconomic positions, gender, and education.

2.4 THE AMBIVALENT RELATIONSHIP BETWEEN TRUST AND CRITICAL THINKING

Trust in news media is crucial for achieving the democratic ideal of an informed citizenry. For citizens to be well-informed, they need to trust the news sources they consume. Trust acts as a prerequisite for people to accept and believe the information presented by the media, or any social actor offering it. This **trust is built upon the understanding that news is provided by experienced professionals who adhere to ethical standards**, giving citizens good reason to believe in the accuracy of the information. When developing an AI instrument like Titan, it will be equally important to be transparent about how the AI is driven and who is responsible for it.

However, trust is not ‘all good’. Too much trust or blind trust can lead to an imbalanced relationship and vulnerability. A certain level of distrust might in turn be necessary to fuel critical thinking and encourages people to question information. The notion that **distrust is not the opposite of trust**, but rather an essential part of its practice, suggests that a healthy level of skepticism is necessary. In developing AI-solutions to engage citizens in critical thinking, we need to recognize the challenge of striking a balance – citizens need to think critically, yet maintain a level of trust in reliable news sources, and in the AI system mediating the two. The dilemma is to encourage critical thinking without eroding the trust between citizens and credible news sources, especially considering the rise of alternative news sources driven by low trust in mainstream media.

Still, as our review shows, we also need to acknowledge that **trust involves an element of commitment and faith**, which can sometimes **conflict with the concept of critical thinking**, where individuals are encouraged to question information. Promoting critical thinking and media literacy, while essential for evaluating information, can inadvertently lead individuals to rely on their intuition and personal biases rather than trusting credible sources. Even reputable sources might be mistrusted by certain groups due to their own belief systems. This complex interplay between critical thinking, personal beliefs, and distrust of established sources can potentially direct people towards misinformation and untruthful sources. There we highlight the need for projects like Titan to address this challenge by helping users recognize their predispositions and biases, preventing the reinforcement of existing biases while still encouraging critical assessment of information.

Furthermore, people's relationship with information and news is not solely driven by rationality and critical assessment. **Emotional factors play a significant role in how individuals trust and consume information.** Studies show that citizens often respond more to emotions than to ideas or facts, a trend amplified by networked technologies and social media. As the information landscape blends news with emotions, opinions, and drama, projects like Titan that promote critical thinking must recognize and address the emotional aspect of information consumption, particularly if they aim to be effective within digital spaces where emotions are highly influential.

Of course, not everyone trusts or distrusts in the same way. Some will be more trusting, others more critical. Therefore, we delved into the literature on socio-demographic factors of news trust and distrust, for which we started from the observation that **media trust is a crucial in the consumption of news and information, but at the same time not the all-predicting factor.** The aim of the review was therefore, on the one hand, to gain insight into the multitude of factors that determine news and information consumption and, on the other, to investigate the underlying socio-demographic heterogeneity regarding the intensity of these different factors.

Looking at different forms of news and information consumption, we witness the existence of different types of media consumers, for example an *online-only* group, a *print and television* group, a *conservative-only* group, *news omnivores*, and also *news avoiders*. For each type of media user, we see that there are clear associations with certain socio-demographic factors such as age, gender, race, education, and political orientation. For example, news avoiders tend to be more often less educated, politically right-wing oriented, and male. Most striking is that **mainly an individual's socioeconomic position has an influence.** Individuals with lower incomes, for instance, rely more on news consumption via social media than those with a higher socio-economic status.

But while socio-demographic factors such as age, gender, race, education and income at first sight are important factors in predicting an individual's media trust, **when adding socio-political factors into the mix, the significant effect of socio-demographic factors disappears.** The literature shows that gender, age, and educational attainment do not significantly affect media trust distribution across different outlets. And further, we also see the hypothesis confirmed that political orientation has a significant effect on media trust. The conclusion here is that **political orientation and political trust are more important in predicting media trust than any socio-demographic factor.** Yet, we cannot be certain about the direction of causality, given that the relationship between media trust and political trust "could follow the logic of a reinforcing spiral" (Brosius et al., 2022, p. 658). What is clear is that political orientation, political trust, and trust interrelate strongly.

Disconnection from news is mainly studied in terms of news avoidance, which is problematic from a democratic point of view. By distinguishing between intentional and unintentional news avoidance, we gain insight into predictive factors that are not socio-demographic in nature as well: cognitive and emotional drivers of news avoidance, socio-cultural processes of role socialization between men and women, and finally also to the market system within which media operates. This therefore shows that in addition to socio-

demographic factors, **contextually related factors are also crucial to understanding and predicting news avoidance.**

At the same time, we know that not only avoiding news is problematic, but also incorrectly believing false information and news. Looking at how individuals access news, research suggests that social media account for only a very limited share of visits to top news sites, but account for a substantial share of visits to fake news sites. Social media is thus a central factor in the production, consumption, and dissemination of fake news. But who then are those people who are less able to recognize truthful and false news? We see that **politically right-leaning news users are more credulous of both true and false information.** Education is also found to be an important factor in several studies: **higher-educated individuals are more likely to recognize fake news,** and also less likely to accept and spread it.

Finally, we looked at underlying socio-demographic factors in access to artificial intelligence (AI) solutions. However, there is still little research on this due to the novelty of AI, and the **evidence seems to be inconclusive.** While some studies suggest no significant effect of age, gender, income, and education on intention to use AI, the perceived accuracy of AI, others show that men tend to be more open towards AI, less concerned about AI and see more benefit in AI. The same is true for individuals with higher household incomes. In addition, level of employment also has a significant influence, but education only weakly correlates with these factors.

The central conclusion is thus, in terms of socio-demographical analysis, that socio-demographics play an essential role in predicting media use. However, these are much less important than political trust and political orientation when predicting media trust. Regarding likeliness to believe false news and information, we expect that improving media literacy (which can be related to education) will lower acceptance of fake news and hence increase detection of fake news. At the same time, one should not overestimate the effect of education since political orientation remains a crucial factor in understanding and predicting who believes false information and news. For the development of an AI tool to assist in critical thinking such as TITAN, this means that it will not be able to remain neutral towards citizen's political predispositions but might want to factor in that **people from different political backgrounds might engage quite differently with the technology.**

2.5 REFERENCES

- [1] L. Graves, "The Monitorial citizen in the 'Democratic recession,'" *Journalism Studies*, Jun. 2017, doi: 10.1080/1461670x.2017.1338153.
- [2] Graeff, *Monitorial Citizenship*, *The international encyclopedia of media literacy*. 2019. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118978238.ieml0169>
- [3] P. Van Aelst et al., "Political communication in a high-choice media environment: a challenge for democracy?," *Annals of the International Communication Association*, vol. 41, no. 1, pp. 3–27, Jan. 2017, doi: 10.1080/23808985.2017.1288551.
- [4] J. Suiter, "Post-truth politics," *Political Insight*, vol. 7, no. 3, pp. 25–27, Nov. 2016, doi: 10.1177/2041905816680417.
- [5] J. Strömbäck et al., "News media trust and its impact on media use: toward a framework for future research," *Annals of the International Communication Association*, vol. 44, no. 2, pp. 139–156, Apr. 2020, doi: 10.1080/23808985.2020.1755338.
- [6] A. M. Rubin, "USES-AND-GRATIFICATIONS PERSPECTIVE ON MEDIA EFFECTS," in *Routledge eBooks*, 2009, pp. 181–200. doi: 10.4324/9780203877111-14.
- [7] Y. Tsifti and J. N. Cappella, "Why do people watch news they do not trust? The need for cognition as a moderator in the association between news media skepticism and exposure," *Media Psychology*, vol. 7, no. 3, pp. 251–271, Aug. 2005, doi: 10.1207/s1532785xmep0703_2.
- [8] S. E. Jarvis, N. J. Stroud, and A. A. Gilliland, "College students, news use, and trust," *Communication Research Reports*, vol. 26, no. 1, pp. 30–39, Feb. 2009, doi: 10.1080/08824090802636991.
- [9] C. Fisher, "The trouble with 'trust' in news media," *Communication Research and Practice*, vol. 2, no. 4, pp. 451–465, Oct. 2016, doi: 10.1080/22041451.2016.1261251.

- [10] N. Couldry, S. Livingstone, and T. Markham, *Media consumption and public engagement: beyond the presumption of attention*. 2007. [Online]. Available: <http://ci.nii.ac.jp/ncid/BB01285508>
- [11] Schudson, "The Power of News," Harvard University Press, 1995, [Online]. Available: https://books.google.be/books/about/The_Power_of_News.html?id=jr9V0ku5rzoC&redir_esc=y
- [12] M. Ekström, T. Olsson, and A. Shehata, "Spaces for public orientation? Longitudinal effects of Internet use in adolescence," *Information, Communication & Society*, vol. 17, no. 2, pp. 168–183, Nov. 2013, doi: 10.1080/1369118x.2013.862288.
- [13] Hovden and Moe, "sociocultural approach to study public connection across and beyond media: The example of Norway," *Convergence*, vol. 23, no. 4, 2017.
- [14] V. Bakir and D. M. Barlow, "Exploring Relationships between Trust Studies and Media Studies," in *Communication in the Age of Suspicion. Trust and the Media*, 2007. doi: 10.1057/9780230206243_2.
- [15] A. Cerra and C. James, *Identity Shift: Where identity meets technology in the Networked-Community Age*. 2011. [Online]. Available: https://openlibrary.org/books/OL28706346M/Identity_shift
- [16] R. Fletcher and S. Park, "The impact of trust in the news media on online news consumption and participation," *Digital Journalism*, vol. 5, no. 10, pp. 1281–1299, Feb. 2017, doi: 10.1080/21670811.2017.1279979.
- [17] Boyd, "Did media literacy backfire?," *Journal of Applied Youth Studies*, vol. 1, no. 4, 2017.
- [18] D. Boyd, "You think you want media literacy... do you? - Data & Society: Points," Medium, Jun. 19, 2018. [Online]. Available: <https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2>
- [19] I. Picone and K. Donders, "Reach or trust optimisation? A citizen trust analysis in the Flemish public broadcaster VRT," *Media and Communication*, vol. 8, no. 3, pp. 348–358, Aug. 2020, doi: 10.17645/mac.v8i3.3172.
- [20] C. Beckett and M. Deuze, "On the Role of Emotion in the Future of Journalism," *Social Media and Society*, vol. 2, no. 3, p. 205630511666239, Jul. 2016, doi: 10.1177/2056305116662395.
- [21] Papacharissi, "A networked self," Google Books. <http://books.google.be/books?id=qhwpKkoFrgsC>
- [22] C. Schwarzenegger, "Personal epistemologies of the media: Selective criticality, pragmatic trust, and competence-confidence in navigating media repertoires in the digital age," *New Media & Society*, vol. 22, no. 2, pp. 361–377, Jan. 2020, doi: 10.1177/1461444819856919.
- [23] M. Bulger, "The promises, challenges, and futures of media literacy," *Data & Society*, Feb. 21, 2018. <https://datasociety.net/library/the-promises-challenges-and-futures-of-media-literacy/>
- [24] M. Bedard and C. Schoenthaler, "Satire or Fake News," *Companion Proceedings of the Web Conference 2018*, Jan. 2018, doi: 10.1145/3184558.3188732.
- [25] S. Edgerly, "Red media, blue media, and purple media: news repertoires in the colorful media landscape," *Journal of Broadcasting & Electronic Media*, vol. 59, no. 1, pp. 1–21, Jan. 2015, doi: 10.1080/08838151.2014.998220.
- [26] B. J. Toff and R. Palmer, "Explaining the gender gap in news avoidance: 'News-Is-for-Men' perceptions and the burdens of caretaking," *Journalism Studies*, vol. 20, no. 11, pp. 1563–1579, Oct. 2018, doi: 10.1080/1461670x.2018.1528882.
- [27] H. Gezduci and L. D'Haenens, "Culture-specific features as determinants of news media use," *Communications*, vol. 32, no. 2, pp. 193–222, Jun. 2007, doi: 10.1515/commun.2007.012.
- [28] A. Kalogeropoulos and R. K. Nielsen, "Social inequalities in news consumption," *Social Science Research Network*, Oct. 2018, [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3270975
- [29] R. Fletcher, A. Kalogeropoulos, and R. K. Nielsen, "More diverse, more politically varied: How social media, search engines and aggregators shape news repertoires in the United Kingdom," *New Media & Society*, vol. 25, no. 8, pp. 2118–2139, Jul. 2021, doi: 10.1177/14614448211027393.
- [30] A. Kalogeropoulos, J. Suiter, L. Udriș, and M. Eisenegger, "News media trust and news consumption: Factors related to trust in news in 35 countries," *International Journal of Communication*, vol. 13, p. 22, Aug. 2019, doi: 10.5167/uzh-175863.
- [31] T.-T. Lee, "Why they don't trust the media: An examination of factors Predicting trust," *American Behavioral Scientist*, vol. 54, no. 1, pp. 8–21, Sep. 2010, doi: 10.1177/0002764210376308.
- [32] N. Verma, K. R. Fleischmann, and K. Koltai, "Demographic factors and trust in different news sources," *Proceedings of the Association for Information Science and Technology*, vol. 55, no. 1, pp. 524–533, Jan. 2018, doi: 10.1002/pa2.2018.14505501057.
- [33] A. Brosius, J. Ohme, and C. H. De Vreese, "Generational Gaps in Media Trust and its Antecedents in Europe," *The International Journal of Press/Politics*, vol. 27, no. 3, pp. 648–667, Aug. 2021, doi: 10.1177/19401612211039440.

- [34] R. Karlsen, A. Beyer, and K. Steen-Johnsen, “Do High-Choice media environments Facilitate news avoidance? A Longitudinal Study 1997–2016,” *Journal of Broadcasting & Electronic Media*, vol. 64, no. 5, pp. 794–814, Nov. 2020, doi: 10.1080/08838151.2020.1835428.
- [35] M. Skovsgaard and K. Andersen, “Conceptualizing news avoidance: towards a shared understanding of different causes and potential solutions,” *Journalism Studies*, vol. 21, no. 4, pp. 459–476, Nov. 2019, doi: 10.1080/1461670x.2019.1686410.
- [36] J. Van Den Bulck, “Television news Avoidance: Exploratory Results from a One-Year Follow-Up Study,” *Journal of Broadcasting & Electronic Media*, vol. 50, no. 2, pp. 231–252, Jun. 2006, doi: 10.1207/s15506878jobem5002_4.
- [37] M. Villi et al., “Taking a Break from News: A Five-nation Study of News Avoidance in the Digital Era,” *Digital Journalism*, vol. 10, no. 1, pp. 148–164, May 2021, doi: 10.1080/21670811.2021.1904266.
- [38] M. Chan, F. Lee, and H.-T. Chen, “Avoid or Authenticate? A multilevel Cross-Country analysis of the roles of fake news concern and news fatigue on news avoidance and authentication,” *Digital Journalism*, pp. 1–20, Mar. 2022, doi: 10.1080/21670811.2021.2016060.
- [39] L. E. Mason, D. Krutka, and J. D. Stoddard, “Media literacy, democracy, and the challenge of fake news,” *The Journal of Media Literacy Education*, vol. 10, no. 2, pp. 1–10, Jan. 2018, doi: 10.23860/jmle-2018-10-2-1.
- [40] S. Morgan, “Fake news, disinformation, manipulation and online tactics to undermine democracy,” *Journal of Cyber Policy*, vol. 3, no. 1, pp. 39–43, Jan. 2018, doi: 10.1080/23738871.2018.1462395.
- [41] S. Waisbord, “Truth is What Happens to News,” *Journalism Studies*, vol. 19, no. 13, pp. 1866–1878, Jul. 2018, doi: 10.1080/1461670x.2018.1492881.
- [42] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.
- [43] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media,” *SIGKDD Explorations*, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.
- [44] G. Rampersad and T. Althiyabi, “Fake news: Acceptance by demographics and culture on social media,” *Journal of Information Technology & Politics*, vol. 17, no. 1, pp. 1–11, Oct. 2019, doi: 10.1080/19331681.2019.1686676.
- [45] T. Koch, L. Frischlich, and E. Lerner, “Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media,” *Journal of Applied Social Psychology*, vol. 53, no. 6, pp. 495–507, Jan. 2023, doi: 10.1111/jasp.12959.
- [46] K. P. Arin, D. Mazrekaj, and M. Thum, “Ability of detecting and willingness to share fake news,” *Scientific Reports*, vol. 13, no. 1, May 2023, doi: 10.1038/s41598-023-34402-6.
- [47] Y. J. Park and S. M. Jones-Jang, “Surveillance, security, and AI as technological acceptance,” *AI & Society*, Jan. 2022, doi: 10.1007/s00146-021-01331-9.
- [48] A. L. Antes, S. Burrous, B. A. Sisk, M. J. Schuelke, J. D. Keune, and J. M. DuBois, “Exploring perceptions of healthcare technologies enabled by artificial intelligence: an online, scenario-based survey,” *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, Jul. 2021, doi: 10.1186/s12911-021-01586-8.
- [49] A. Følstad et al., “Future directions for chatbot research: an interdisciplinary research agenda,” *Computing*, vol. 103, no. 12, pp. 2915–2942, Oct. 2021, doi: 10.1007/s00607-021-01016-7.

3 PSYCHOLOGICAL/MOTIVATIONAL/COGNITIVE MECHANISMS

3.1 SOCIAL COGNITIVE THEORY (SCT): THE BROAD PSYCHOLOGICAL FRAMEWORK

Social cognitive theory is a well-known framework for understanding how individuals interact with media and how this interaction may influence their behavior. In particular, the Triadic Reciprocal Determinism model [1], which is a fundamental component of social cognitive theory, suggests that personal factors, environmental factors, and individual behavior interact with each other to shape an individual's behavior in a given situation (Fig. 1). The interactionist approach is particularly relevant when dealing with complex behaviors, such as the relationship between the individual and misinformation. In fact, in order to better understand this problem, to evaluate it as comprehensively as possible, and thus know what skills to act on to promote active and critical thinking about misinformation, it is necessary to always keep in mind how individuals are influenced by multiple factors that contribute to their credulity and propensity to share misleading information. A meaningful understanding of these factors, therefore, is only possible by having this continuous interaction well in mind [2]. In other words, triadic reciprocal determinism describes how an individual adapts his or her actions in relation to a given situation to achieve desired outcomes [3], and thus only by considering this interaction it is possible identify self-regulatory capacities that can help individuals relate critically and actively to misinformation.

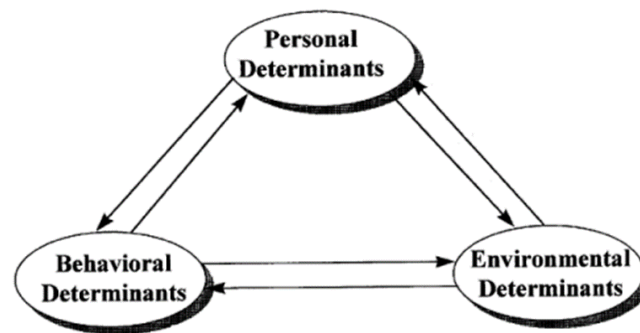


Figure 3.1: Reciprocal Triadic Determinism [1] : A. Bandura, "Human agency in social cognitive theory," *American Psychologist*, vol. 44, pp. 1175–1184, 1989, doi: 10.1037/0003-066X.44.9.1175.

3.1.1 Personal factors

Personal factors include all the different individual components that make up a person's self, such as the cognitive and motivational aspects that can explain specific goals and beliefs related to the particular behavior. This also includes an individual's ability to discriminate true from false news, or the set of cognitive biases that make a person inclined to classify a piece of information as true or false. It is possible, for example, to identify a noncritical and passive mode of information processing, so-called "lazy thinking," which has a low cognitive cost yet is reactive and fast, that has been shown to increase people's susceptibility to misinformation by reducing their ability to distinguish true from false news [4]. Partisan bias can also increase people's susceptibility to fake news, inducing a response bias to accept information congruent with one's ideological beliefs and discard incongruent information, regardless of truthfulness [5]. At the opposite end of the spectrum, an individual propensity to think thoroughly and carefully through a problem, resisting the urge to rush through impulsive tendencies, is indicative a more controlled, slower and less reactive modality of information processing in which the person can deal with the challenges posed by online misinformation with a reflective approach.

3.1.2 Situational and contextual factors

Considering the conditions under which the behavior occurs, i.e., the situational-contextual aspect, it is necessary to consider, for example, the influence played by different news content and verbal-linguistic modality on the regulation of the behavior of the person facing the piece of misinformation. In fact, it is known how certain specific issues and news contents can activate differently certain people with specific value configurations [6], prompting in certain cases the person's cognitive system to engage in a more intuitive and

rapid processing that leads to greater credulity, greater emotional activation and to a greater sharing of misinformation [7]. The same observation can be made about another contextual condition, namely the linguistic-verbal specificity of the news itself. Indeed, the formal and stylistic aspect of the piece of news (such as the use of exaggerations, linguistic biases, subjectivity, and hedged language) has the same effect of promoting the bypass of the proactive and reflective mode of information processing [8]. Additionally, the perceived novelty of news content also critically influences its effect and may act as a determining factor in significant differences in the propagation of misinformation and, consequently, its impact [7]. Another important contextual factor that cannot be ignored when considering the situational aspects is related to online conformity: due to the very nature of social networks, the effect on a person of seeing a piece of disinformation shared (and maybe liked) by a stranger or by an individual belonging to one's circle of reference figures is significantly different. It has been pointed out that certain types of news content may be more relevant to specific online communities, and affiliation with and loyalty to those communities are additional crucial contextual and environmental factors relevant to understanding the amplification of beliefs of specific online communities, and ultimately the spread of misinformation [7]. The same can be said regarding the difference in reception by the social network population, considering, for example, the number of likes or the amount and tone of comments below the news item. Moreover, not all social networks drive the same actions: misinformation is more prevalent on social platforms that allow for the creation and sharing of user-generated content (such as Facebook and Twitter) than on those that primarily feature professionally produced content or simply allow for its fruition [9].

3.1.3 Behavioral factors

Individual behavioral responses, such as liking, sharing, interacting, commenting, or reporting misinformation, are directly related to the degree of reactivity or, conversely, active reflection that individuals can engage in when they encounter misinformation. The passive and reactive mode of information processing, as described by Pennycook [4], leads by definition to a more impulsive and reactive response, whereas a more thoughtful and reflective behavioral response requires cognitive effort and a certain level of motivation. To understand whether responses to misinformation are more related to an impulsive or a thoughtful mode of information processing, it may be useful to examine them by considering three different dimensions, which are always in constant interaction with the personal and situational factors expressed above. The first dimension is involvement, which refers to the extent to which the individual feels the need or urge to engage with the disinformation article. This may depend on factors such as personal beliefs, interests, and values. The second dimension is responsiveness, which refers to the speed with which the person tends to react to the news item. This may be related to the level of emotional arousal triggered by the misinformation and the level of motivation to engage with the content. Finally, the third dimension is related to content, and relates to the specific topics and vocabulary used in the disinformation. This, as mentioned above, can influence the behavioral intention of the individual and the extent to which he or she is likely to engage with the content.

3.2 CRITICAL THINKING WITHIN AN AGENTIC PERSPECTIVE

In this theoretical framework, the person is viewed from an agentic perspective: the emphasis is on the active role of the individual as an agent of his or her own behavior, recognizing the person's ability to make decisions and become aware of information about the external world and himself in an active manner, constantly re-aligning his behavior on the basis of that information to move toward his or her goals [1]. Inherent in this agentic view of the person is the very definition of the capacity for self-regulation, that is, the ability to monitor and control one's thoughts, feelings and behaviors to achieve one's own goals; this includes adapting to changing circumstances and navigating complex social environments. In the context of misinformation, a person's ability to self-regulate thus appears to be a capacity closely related to the ability to exercise mature and autonomous critical thinking, providing the person with the ability to actively manage his or her own cognitive tools necessary to recognize fallacious information from truthful information, and resist the influence of information presented through misleading verbal and content modes.

Through a sense of personal agency and through their own self-regulation skills, individuals can develop the skills and strategies necessary to also critically evaluate their own level of specific knowledge related to the news topic and identify their own limitations and potential biases or prejudices. In fact, precisely because the active information-processing mode is able to "think about one's own thinking", thus performing the activity of metacognition, it also has the ability to question one's own views. This ability of the person to think critically about one's own beliefs and goals, as well as about one's own behavior and context, therefore, is in a condition of mutual exchange with the possibility of actively regulating oneself. It thus becomes possible to claim for oneself the power to change one's thinking and behavior.

For this reason, looking at the main theoretical and operational definitions of critical thinking in psychological literature, it can be seen that the role of a person's self-regulatory abilities is always placed at the forefront. In the most widely recognized definition, and most supported by empirical evidence, Facione [10] defines critical thinking as "the process of purposeful, self-regulatory judgment, which results in interpretation, analysis, evaluation, and inference". Based on a sequence of extensive research shared with the scientific community, the author reports the characterization of the ideal critical thinker obtained directly from the Delphi panel, stating how the ideal critical thinker should be "inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgments, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the subject and the circumstances of inquiry permit" [10]. Nosich's well-known definition of critical thinking [11] states clearly how "critical thinking is different from just thinking. It is metacognitive - it involves thinking about your thinking". To evaluate and understand critical thinking in a meaningful way, it is therefore necessary to recognize that, apart from the differences in formulations, the definitions agree that critical thinking involves self-regulatory judgments, reflection on a person's ways of thinking, and engagement in reflective skepticism to judge the rationality of one's thinking. The person capable of critical thinking is self-regulated, and the self-regulated person can think critically [12], [13].

3.2.1 Main cognitive skills of critical thinking

The 1990 APA Delphi Report on Critical Thinking [10], a comprehensive study that provided a thorough identification of the essential components that define critical thinking and that contribute to its development, identified six cognitive skills as central to its definition which include interpretation, analysis, evaluation, inference, explanation, and self-regulation. In particular, the first five are also the dimensions along which assessments that rely on this model are usually carried out, and several validated tools in this literature have been developed for this very purpose [14,17]. The first skill, interpretation, involves understanding and expressing the meaning of statements, further supported by sub-skills such as categorization, decoding the meaning, and clarifying the meaning. Analysis, the second skill, involves identifying the intended and actual relationships among different forms of statements, with sub-skills including examining ideas and identifying arguments. Evaluation, the third skill, is critical to assessing the credibility of claims and includes sub-skills such as evaluating claims and evaluating arguments. Inference, the fourth skill, involves identifying evidence for drawing reasonable conclusions and inferring consequences that can be derived from statements, with sub-skills including questioning evidence, hypothesizing alternatives, and drawing conclusions. Finally, explanation, the fifth skill, involves the appropriate statement and justification of the results of one's reasoning, with sub-skills such as stating results, justifying procedures, and presenting arguments. As said before, while these first five skills are commonly assessed through tests or questionnaires, the sixth skill, namely self-regulation, was not originally directly measured in the same way as the others because it involves the engagement of the higher-order personal resources that allow the other five skills and are mutually influenced by them. The Delphi Report explicitly specifies how self-regulation is fundamental to critical thinking because it involves consciously monitoring one's own cognitive activities, the elements utilized and the outcomes, while also agreeing to question, validate or correct the results of one's reasoning through the sub-skills of self-examination and self-correction.

3.2.2 Self-regulatory feedback loop for learning critical thinking

Building from that, relevant literature has already demonstrated that the subprocesses in the complex mechanism of self-regulation (which by definition unfold through personal, behavioral, and situational aspects) are closely related to the critical thinking processes identified by the APA Delphi Report (**Fig. 2**), and that the relationship between self-regulation and critical thinking is one of reciprocal reinforcement and mutual empowerment **[13]**.

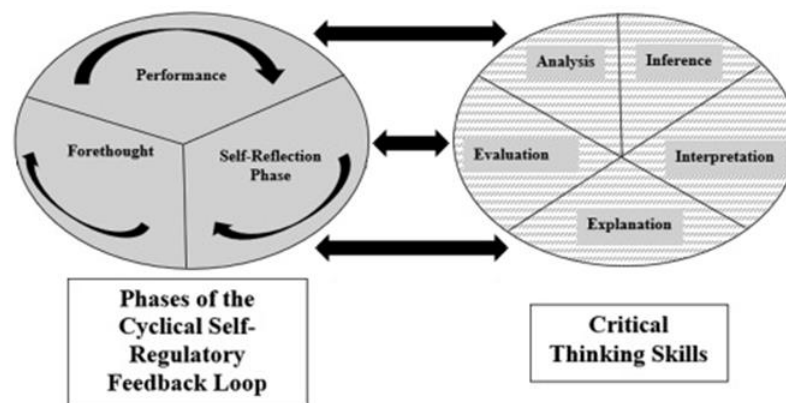


Figure 3.2: Relationship between critical thinking skills and the three-stage model of self-regulatory functioning, based on Kitsantas et al. [13]

Nevertheless, if self-regulation can be said to be a fundamental skill that is part of the psychological mechanisms related to critical thinking, it is also crucial to remember how from a social-cognitive point of view the role of self-regulation extends far beyond that of a simple sub-factor of critical thinking, going on to touch every aspect of the person. To explore this explicit connection between self-regulation and the development of critical thinking skills, it could be noticed how in Zimmerman's Self-Regulatory Feedback Loop model **[18]** three different cyclical phases of self-regulated learning are specified to highlight the role of metacognitive, motivational and behavioral processes in skill acquisition. This model specifically explains the dynamic and cyclical nature of self-regulation, and emphasizes the importance of metacognitive awareness, goal setting, strategy use, and self-assessment in effective self-regulated learning.

In detail, the forethought phase precedes behavior and involves the processes of planning and goal setting. The person sets a goal and plans his or her behavior to prepare to achieve it through task analysis and by setting attainable goals. In addition, the individual assesses his or her self-efficacy, that is, the belief that he or she is capable of performing the task successfully. In the performance phase, the person uses his or her own strategies and skills to engage in the learning task, monitor progress, and make necessary corrections. It also involves the use of metacognitive strategies, such as self-interrogation or self-explanation. Adjusting the motivation and affective states is also critical in this phase. The phase of self-reflection occurs after completion of the specific behavior and involves reflections on performance and evaluation of progress toward desired goals. Through the processes of self-evaluation, self-reinforcement and self-reaction, motivation and self-efficacy for future learning tasks are influenced. These three phases are interdependent, and occur physiologically during a dynamic and coordinated interaction, and the cycle of self-regulation is completed when the processes of self-reflection influence the subsequent beliefs associated with the forethought phase, and then the behaviors of the performance phase **[13]**.

3.3 SELF-REGULATED LEARNING AND CRITICAL THINKING AGAINST DISINFORMATION

From this perspective, intelligent systems could support the learning and acquisition of critical thinking skills against misinformation through adaptive environments with personalized support **[19]**. Considering the five sub-skills identified by the 1990 APA Delphi Report on Critical Thinking, one can recognize how interpreting statements encountered in the news, analyzing them, evaluating them, making meaningful inferences, and

explaining one's elaborations is a complex skill that requires practice, guidance, and repetition that cannot be separated from effective self-regulation abilities: “intelligent technology-based systems can augment one’s performance by facilitating self-regulation and critical thinking” [13]. In order to facilitate the process of shared construction of mature, self-regulated critical thinking skills by an intelligent system, it is possible to identify a set of open-ended questions that can be applied not to be answered precisely, but to show the person a common way of mental functioning that often occurs implicitly and can be intentionally exploited by disinformation outlets. Through open-ended questions, the person can be nudged to engage in slower self-regulated modes of information processing, which are analytical and less reactive, and from here the person’s engagement with the five mechanisms of critical thinking against misinformation could be enhanced [13].

System implementation	Importance for critical thinking skills	Assessment of relevant dimensions	AI-assisted support of critical thinking skills
(1) Forethought phase: self-motivational beliefs	Importance of the general effect of personal values orientations and cognitive processes in a given situation as possible antecedents of critical thinking.	<ul style="list-style-type: none"> -Does the user evaluate the personal salience of that specific piece of information? -Does the user consider the expectations of others before sharing news? -Does the user believe he or she is capable of distinguishing real news from fake news? 	<ul style="list-style-type: none"> -Evaluation of personal values and of their importance in discerning news -Evaluation of personal beliefs about the ability to realistically discriminate news.
(2) Executive phase: self-control and self-monitoring	Impact of self-observation and self-monitoring in the ability to critically judge news in different online contexts.	<ul style="list-style-type: none"> -Can the user manage the impulse to share? -Can the user notice the different affordances that different social contexts offer? 	<ul style="list-style-type: none"> -Assessment of the relative importance of the time factor in sharing fake news
(3) Self-reflective phase: self-judgment and self-reaction	Importance of attribution of the cause of failure in critical thinking skills, showing through interactive coaching the cognitive mechanisms in action so as to improve the future phases.	<ul style="list-style-type: none"> -Can the user self-assess his/her ability to discriminate true news from false news? -What are the causal attributions of errors in critical ability? -Is the user aware of the cognitive mechanisms that follow sharing? -Does the user consider the reaction of others? 	<ul style="list-style-type: none"> -Assessment of the importance of others' reactions on social media to news sharing.

Table 3.1: Relationship between the different stages of the self-regulated learning process, the importance for the person's critical thinking, and the assessment of relevant dimensions.

3.3.1 Critical thinking and individual vulnerability

Addressing the complex issue of enhancing critical thinking toward misinformation, particularly in relation to AI-based coaching, cannot be separated from a caveat. Care must be taken not to fall into the error of oversimplification, that is, not taking sufficiently into account different individual characteristics. For example, it has already been shown that for some people the personal and emotional cost of the risk of questioning their beliefs on certain specific issues may be too high. For individuals with low self-esteem, the sense of belonging that the community that identifies around certain polarizing topics provides is sufficient to motivate the person not to engage in the cognitive processes related to critical thinking [20]. These processes may directly undermine this sense of community connected to a specific misinformation issue. Because of individual vulnerability, for these people the fear of losing the social network related to certain disinformation issues is not counterbalanced by the perceived benefit of engaging in mature, self-regulated critical thinking [21]. To understand what psychological variables explain misinformation without falling into an oversimplification of the problem, therefore, it is crucial to consider how it may be difficult for certain people to have the resources necessary to choose and be motivated to implement critical thinking toward misinformation. For them, therefore, the problem would not be solved simply by improving their critical thinking skills, as these skills would not be found to be lacking or impaired [22]. It is very important, therefore, that AI-enhanced critical thinking tools for countering the spread and impact of disinformation should be designed based on the understanding that enhancing critical thinking processes in certain cases may be a necessary but not sufficient condition for addressing the problem and the impact of credulity and disinformation.

3.4 THE SOCRATIC METHOD AND CRITICAL THINKING ASSESSMENT

3.4.1 The Socratic method and Socratic AI: philosophical and educational approaches

The ancient Greek philosopher Socrates (469-399 BCE) emphasized the importance of the moral and intellectual qualities of human beings [23,24]. While Socrates did not leave behind any writings, his philosophical originality in method and speech primarily survives in the works of Xenophon and Plato, particularly in the early Platonic dialogues [25]. Socrates' method, known as the 'maieutic', involved the critical examination of people's unquestioned ideas, and the extraction of their latent ideas, thereby making them conscious of their own opinions and beliefs [26, 24, 27, 28,29]. The Socratic method had a significant influence, not only on classical thinking but also on contemporary educational practices [30]. In modern adaptations of the Socratic method, individuals are guided through systematic questioning [31,32, 26, 33, 34, 35], with the aim of facilitating the development of critical thinking skills [36,37]. The modern Socratic method encourages individuals to examine their beliefs, uncovering inherent and often unnoticed contradictions, biases, and stereotypes, in the context of various perspectives and viewpoints.

Moreover, Socratic dialogical reasoning has been integrated with educational practices [37, 30] offering a pathway to foster critical thinking, creative skepticism, self-reflection, and comprehension [32, 36]. Case methods in online and virtual educational approaches have indicated that Socratic dialogue promotes critical thinking and higher-order thinking skills [38]. The Socratic method has been applied in combination with educational technologies and artificial intelligence, frequently denoted as 'Socratic AI.' [39, 40, 41, 42]. A classification scheme has been proposed for technology-enhanced learning tools that utilize the Socratic method to support critical thinking [41]. This scheme includes the critical thinking framework [36], the classical Socratic method, the modern Socratic method, and the Neo-Socratic group discussion method [41]. Scholars have discussed and proposed formal models regarding the applicability and usefulness of the Socratic method, which involves individuals training with a chatbot to enhance their critical thinking skills [43]. In addition, the

concept of AI-framed Questioning has been introduced to convert AI-related information into questions, actively engaging users in thinking and offering support for their process of reasoning and critical thinking [44, 42] Socratic AI has also received attention in regard to its impact on human enhancement and the ethical empowerment of personal and group morality [39]. A human-centered perspective on AI [45] and narrative technologies for human interaction pave the way for assisting in productive, moral, and genuine collaboration among humans [46] in pursuit of critical thinking, collaborative inquiry [47], and collective argument mapping through computer-assisted tools.

3.4.2 Improving critical thinking with collective argument mapping in a collaborative module

The challenge of the collaborative module, proposed as “collaboration of the citizens” within TITAN, is to enhance a group's ability to critically examine a potentially disinformation-laden topic by involving different perspectives and potential disagreements in a productive discussion. While the classical Socratic method offers a way to conduct one-to-one dialogue, a collaborative discussion must allow for many to many interactions. This challenge calls for a clear structure of the group discussion while preserving the objective of increasing participants' critical thinking level. In accordance with the Socratic method, participants must be able to comprehend the claims, and critically examine their premises and counterarguments while having the potential to add a constructive contribution to the discussion. The most promising method for this task is computer-assisted argument mapping. Argument mapping is a visual representation or diagram that organizes and illustrates the structure of an argument, including its premises and conclusions, to facilitate understanding and analysis. It can enhance collective discussions by giving participants shared point of reference, visually clarifying the logical connections and reasoning within a current state of discussion, which promotes shared understanding and enables more effective collaboration.

Several studies have found that argument mapping is an effective method for improving critical thinking. This is demonstrated by comparing the size of the effect on critical thinking to other methods, measured by standard deviations (SD). For instance, the first year of college results in an increase of 0.16-0.3 SD [48] while a semester of college with a mixed philosophy and critical thinking course results in a 0.38 SD increase [49]. Conversely, the effect of a semester-long argument mapping course results in a much higher increase of 0.7-0.72 SD [50, 51, 52].

Other studies have found that argument mapping improves analytical reasoning [50], mathematical reasoning [50], and essay writing [51]. Additionally, computer-assisted argument mapping is more effective than the pen-and-paper method in recall, retention, and comprehension of information [53].

The proposed explanation for this phenomenon suggests that argument mapping decreases the cognitive load of the person trying to analyze new information. Cognitive load refers to the burden placed upon an individual in using and distributing working memory resources during cognitive activities such as learning and problem-solving [54]. The nature of a typical text article or speech is linear, sometimes making it difficult to decipher the reasoning and assimilate the information within a given narrative in which a variety of propositions are presented [55]. This places the cognitive load on the reader to analyze and compare claims that are not in direct vicinity of each other. Argument maps decrease this cognitive load by presenting information in a nonlinear and more structured manner, reducing the need for attention switching between different parts of the text or speech and allowing more mental processing to be devoted to understanding logical relationships [56].

Additionally, collaborative computer-assisted argument mapping frames discussions as precise, merit-based, and focused on the validity of reasoning itself, rather than on the persons formulating a claim. Moreover, qualitative analysis indicates that with argument mapping, students take a more deliberate and fair-minded approach to examining controversial arguments [57]. Therefore, we can expect the discussion to be more civil,

with no personal attacks or other unpleasant failure modes of traditional forums and social media. Due to all of the above reasons, it is a promising framework for a collaborative module in the TITAN tool.

3.5 REFERENCES

- [1] A. Bandura, "Human agency in social cognitive theory," *American Psychologist*, vol. 44, pp. 1175–1184, 1989, doi: 10.1037/0003-066X.44.9.1175.
- [2] A. Bandura, "Social Cognitive Theory of Mass Communication," in *Media Effects*, 3rd ed. Routledge, 2008.
- [3] A. Bandura, "Self-referent mechanisms in social learning theory," *American Psychologist*, vol. 34, pp. 439–441, 1979, doi: 10.1037/0003-066X.34.5.439.b.
- [4] G. Pennycook and D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, vol. 188, pp. 39–50, Jul. 2019, doi: 10.1016/j.cognition.2018.06.011.
- [5] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, Jun. 2014, doi: 10.1073/pnas.1320040111.
- [6] G. Feldman, "Personal Values and Moral Foundations: Examining Relations and Joint Prediction of Moral Variables," *Social Psychological and Personality Science*, vol. 12, no. 5, pp. 676–686, Jul. 2021, doi: 10.1177/1948550620933434.
- [7] F. D'Errico, G. Corbelli, C. Papapicco, and M. Paciello, "How Personal Values Count in Misleading News Sharing with Moral Content," *Behavioral Sciences*, vol. 12, no. 9, Art. no. 9, Sep. 2022, doi: 10.3390/bs12090302.
- [8] A. Staender, E. Humprecht, F. Esser, S. Morosoli, and P. Van Aelst, "Is Sensationalist Disinformation More Effective? Three Facilitating Factors at the National, Individual, and Situational Level," *Digital Journalism*, vol. 10, no. 6, pp. 976–996, Jul. 2022, doi: 10.1080/21670811.2021.1966315.
- [9] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018, doi: 10.1126/science.aap9559.
- [10] P. Facione, "Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report)." American Philosophical Association, 1990.
- [11] G. Nosich, "Learning to Think Things Through: A Guide to Critical Thinking Across the Curriculum," *Centers for Teaching Excellence - Book Library*, Jan. 2012, [Online]. Available: <https://digitalcommons.georgiasouthern.edu/ct2-library/197>
- [12] Critical thinking in psychology. in *Critical thinking in psychology*. New York, NY, US: Cambridge University Press, 2007, pp. xii, 340.
- [13] A. Kitsantas, A. L. Baylor, and S. E. Hiller, "Intelligent technologies to optimize performance: Augmenting cognitive capacity and supporting self-regulation of critical thinking skills in decision-making," *Cognitive Systems Research*, vol. 58, pp. 387–397, Dec. 2019, doi: 10.1016/j.cogsys.2019.09.003.
- [14] N. C. Facione, P. A. Facione, and C. A. Sanchez, "Critical Thinking Disposition as a Measure of Competent Clinical Judgment: The Development of the California Critical Thinking Disposition Inventory," *Journal of Nursing Education*, vol. 33, no. 8, pp. 345–350, Oct. 1994, doi: 10.3928/0148-4834-19941001-05.
- [15] G. G. Nair and L. L. Stamler, "A Conceptual Framework for Developing a Critical Thinking Self-Assessment Scale," *J Nurs Educ*, vol. 52, no. 3, pp. 131–138, Mar. 2013, doi: 10.3928/01484834-20120215-01.
- [16] G. G. Nair, L.-A. M. Hellsten, and L. L. Stamler, "Accumulation of Content Validation Evidence for the Critical Thinking Self-Assessment Scale," *J Nurs Meas*, vol. 25, no. 1, pp. 156–170, 2017, doi: 10.1891/1061-3749.25.1.156.
- [17] R. Payan-Carreira, A. Sacau-Fontenla, H. Rebelo, L. Sebastião, and D. Pnevmatikos, "Development and Validation of a Critical Thinking Assessment-Scale Short Form," *Education Sciences*, vol. 12, no. 12, p. 938, Dec. 2022, doi: 10.3390/educsci12120938.
- [18] B. J. Zimmerman, "Chapter 2 - Attaining Self-Regulation: A Social Cognitive Perspective," in *Handbook of Self-Regulation*, M. Boekaerts, P. R. Pintrich, and M. Zeidner, Eds., San Diego: Academic Press, 2000, pp. 13–39. doi: 10.1016/B978-012109890-2/50031-7.
- [19] P. J. Durlach and A. M. Lesgold, *Adaptive Technologies for Training and Education*. Cambridge University Press, 2012.
- [20] J.-W. Van Prooijen, "Sometimes inclusion breeds suspicion: Self-uncertainty and belongingness predict belief in conspiracy theories: Self-uncertainty and conspiracy beliefs," *Eur. J. Soc. Psychol.*, vol. 46, no. 3, pp. 267–279, Apr. 2016, doi: 10.1002/ejsp.2157.

- [21] A. Cichocka, M. Marchlewska, and A. G. De Zavala, "Does Self-Love or Self-Hate Predict Conspiracy Beliefs? Narcissism, Self-Esteem, and the Endorsement of Conspiracy Theories," *Social Psychological and Personality Science*, vol. 7, no. 2, pp. 157–166, Mar. 2016, doi: 10.1177/1948550615616170.
- [22] K. M. Douglas, R. M. Sutton, and A. Cichocka, "The Psychology of Conspiracy Theories," *Curr Dir Psychol Sci*, vol. 26, no. 6, pp. 538–542, Dec. 2017, doi: 10.1177/0963721417718261.
- [23] Morrison, D. (2010). *The Cambridge Companion to Socrates (Cambridge Companions to Philosophy)*. Cambridge: Cambridge University Press.
- [24] Wright, M. R. (2009). *Introducing Greek Philosophy*. Routledge.
- [25] Cooper, John (ed.) (1997). *Plato: Complete Works*. Indianapolis: Hackett.
- [26] Benson, H. (2010). "Socratic Method." In D. Morrison (Author), *The Cambridge Companion to Socrates (Cambridge Companions to Philosophy)*, pp. 179-200). Cambridge: Cambridge University Press.
- [27] Benson, H. (2000). *Socratic Wisdom*. Oxford: Oxford University Press.
- [28] Frede, M. "Plato's Arguments and the Dialogue Form." *Oxford Studies in Ancient Philosophy, Supplementary Volume 1992*, Oxford, 201–19.
- [29] Vlastos, Gr. (1983). "The Socratic Elenchus." *Oxford Studies in Ancient Philosophy, I*, Oxford 1983, 27–58.
- [30] Brickhouse, Th. C., and Nicholas D. Smith. (2010). "Socratic Teaching and Socratic Method." In Harvey Siegel (ed.), *The Oxford Handbook of Philosophy of Education, Oxford Handbooks (Oxford Academic)*.
- [31] Gogus, A. (2012). "Socratic Questioning." In: Seel, N. M. (eds) *Encyclopedia of the Sciences of Learning*. Springer, Boston.
- [32] Boghossian, P. (2012). "Socratic pedagogy: perplexity, humiliation, shame and a broken egg." *Edu. Philos. Theor.*, 44, 710–720.
- [33] Seeskin, K. (1987). *Dialogue and discovery: a study in Socratic method*, SUNY Press.
- [34] Overholser, J. C. (1993). *Elements of the Socratic method: I. Systematic questioning. Psychotherapy: Theory, Research, Practice, Training*, 30(1), 67.
- [35] Nelson, L. (1950). *Socratic Method and Critical Philosophy. Selected Essays Translated by Thomas K. Brown III*. Kessinger Publishing.
- [36] Paul, R., & Elder, L. (2006). "The Art of Socratic Questioning." *The Foundation for Critical Thinking*.
- [37] Shu, H., J. (1993). "Critical thinking: A Socratic model." *Argumentation*, 7(3), 291-311.
- [38] Brooke, S.L. (2006). "Using the case method to teach online classes: Promoting Socratic dialogue and critical thinking skills." *International Journal of Teaching and Learning in Higher Education*, 18(2), 142-49.
- [39] Lara, F., and Deckers, J. (2020). "Artificial Intelligence as a Socratic Assistant for Moral Enhancement." *Neuroethics*, 13, 275–287.
- [40] Kim, T. W. and Mejia, S. (2019). "From Artificial Intelligence to Artificial Wisdom: What Socrates Teaches Us." *Computer*, 52(10), 70-74.
- [41] Le N-T (2019). "How Do Technology-Enhanced Learning Tools Support Critical Thinking?" *Frontiers of Education*, 4:126, p. 3 ff.
- [42] Garito, M. A. (1991). *Artificial intelligence in education: evolution of the teaching—learning relationship. British journal of educational technology*, 22(1), 41-47.
- [43] Le, N-T and Huse, N. (2016). "Evaluation of the Formal Models for the Socratic Method." In *Proceedings of the 13th International Conference on Intelligent Tutoring Systems - Volume 9684 (ITS 2016)*. Springer-Verlag, Berlin, Heidelberg, 69–78.
- [44] Danry, V., Pataranutaporn, P., Mao, Y. and Maes, P. (2023). "Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 1–13
- [45] Shneiderman, B. (2022). *Human-Centered AI*. Oxford: Oxford University Press.
- [46] Coeckelbergh, M. (2022). *Self-improvement: Technologies of the Soul in the Age of Artificial Intelligence*. New York: Columbia University Press.
- [47] Ellerton, P. (2017). "On Critical Thinking and Collaborative Inquiry." *Education: Future Frontiers. Occasional paper series*, NSW Department of Education National Initiatives & Performance Directorate.
- [48] Huber and Kuncel (2016). "Does College Teach Critical Thinking? An Analysis", *Review of Educational Research*, 86(2): 431-68.

- [49] Abrami et. al (2015). "Strategies for Teaching Students to Think Critically: A Meta-Analysis", *Review of Educational Research*, 85(2): 275-314.
- [50] Cullen, S., Fan, J., van der Brugge, E. et al. Improving analytical reasoning and argument understanding: a quasi-experimental field study of argument visualization. *npj Science Learn* 3, 21 (2018). <https://doi.org/10.1038/s41539-018-0038-5>
- [51] Van Gelder, T.J. (2003). Enhancing deliberation through computer supported argument mapping. In Kirschner, P., Buckingham Shum, S., Carr, C. (eds.), *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, 97–115, London: Springer-Verlag.
- [52] Twardy, C. (2004). Argument maps improve critical thinking. *Teaching philosophy*, 27(2), 95-116.
- [50] Indrawatiningsih, N., Purwanto, P., As'ari, A. R., & Sa'dijah, C. (2020). Argument Mapping to Improve Student's Mathematical Argumentation Skills. *TEM Journal*, 9(3).
- [53] Robillos, R. J. (2021). Learners' Writing Skill and Self-Regulation of Learning Awareness Using Computer-Assisted Argument Mapping (CAAM). *Teaching English with Technology*, 21(4), 76-93.
- [54] Eftekhari, M., & Sotoudehnama, E. (2018, September). Effectiveness of Computer-Assisted Argument Mapping for Comprehension, Recall, and Retention. *ReCALL*, 40(3), 337-354. <https://doi.org/10.1017/S0958344017000337>
- [55] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257-285
- [56] Dwyer, Christopher P, Michael J Hogan, and Ian Stewart. "An evaluation of argument mapping as a method of enhancing critical thinking performance in e-learning environments." *Metacognition Learning* 7 (2012): 219-44.
- [57] Kaepfel, K. (2021, June). The Influence of Collaborative Argument Mapping on College Students' Critical Thinking About Contentious Arguments. *Thinking Skills and Creativity*, 40.

4 AI TECHNOLOGY TRENDS FOR MODELS, DATA AND RESOURCES TO FIGHT DISINFORMATION

The objectives of this subsection are: i) to review state-of-the-art procedures and tools for fact-checking, ii) to summarize key disinformation signals as identified in literature review, iii) to review state-of-the-art AI models, datasets and approaches used for detecting disinformation signals with a special focus on fact-checking, fallacies and hate speech, and iv) to conceptualize an AI coaching system, based on the Socratic method and experiential learning, with the aim of stimulating the critical thinking of a citizen to fight disinformation.

4.1 STATE-OF-THE-ART PROCEDURES AND TOOLS FOR FACT-CHECKING

In training an AI to engage in a Socratic dialogue with citizens that seek to fact-check information, we can learn from the procedures that fact-checkers have by now developed to monitor claims that are worth to fact-check. These include:

- CrowdTangle is a social media monitoring and analysis tool. Not all fact checkers use CrowdTangle this heavily. The reasons for this include lack of confidence or mastery, lack of local language capability (for example CrowdTangle not recognizing Turkish letters like ç), or seeing CrowdTangle as a less useful tool for identifying online misinformation in their country context.
- Manual searches - Many fact checkers also manually monitor, following specific people and pages, searching for links and misspellings of names, or following private Instagram accounts to monitor what content is appearing beyond CrowdTangle’s reach, which only covers public content. Some fact checkers also monitor topics to spot possible claims – for example a keyword search for “coronavirus” could highlight claims that the virus originated in a lab in Wuhan, or that 5G causes the virus.
- Reader suggestions - Many fact checkers encourage reader suggestions via WhatsApp, Messenger, email, Telegram and custom-built platforms.
- Facebook’s fact checking product - Facebook’s program provides partners with a tool to carry out their work: the “fact checking product”. This allows fact checkers to see a queue of user-submitted and AI-surfaced content which may be false or misleading (known colloquially as “the queue”). Ideas about how to funnel down the content in the queue included:
 - Connect Full Fact’s claim detection tool to the queue to sift checkable claims from general viral content. This uses a machine learning model to say whether or not text contains a factual claim, and could reduce the queue down and make it more manageable and useful.
 - Expand community reviewers to sift through the queue and narrow down what is there based on criteria agreed among fact checkers.
 - Feed whitelisted sites such as genuine news websites into the algorithm.
 - Integrate speech to text software for video content.
- While most fact checkers’ preferred monitoring tools are CrowdTangle and Facebook’s fact checking product, others include Google Alerts, Brand24, Twitter advanced search, Buzzsumo and (paid-for tool) Trendolizer.
- Claim selection - It is one thing to monitor what is being claimed on social media. Another thing is then to choose which claims should be checked by AI. The obvious answer is ‘the non-factual ones’. Still, some claims are more suitable to use AI on than others. We can learn from looking at how fact checkers select claims to investigate:
 - The first and most obvious consideration is whether a claim can be checked. Within Facebook’s fact checking product and among reader requests are a lot of opinions or commentary articles that fact checkers can’t verify. Sometimes data or evidence isn’t available for certain topics, or is of such low quality that it isn’t usable.

Selection criteria prioritize potential for harm. Questions fact checkers ask themselves about harm include:

- If someone believed a claim, what damage could it cause to their and others’ health, lives or finances?

- Could the claim threaten democratic processes or minority groups?
- What is the implication for public discourse and national security?
- Who and how will this claim hurt if people believe it?
- Is life at risk?
- Does the claim relate to an urgent situation (e.g. floods, bombings) and require a quick response to stop the misinformation from exacerbating that situation?

Virality and reach are important, but almost all the fact checkers we interviewed had a skeptical view of virality and how to define it. Here are a few of the ways fact checkers consider reach and virality as part of their selection criteria:

- Number and speed of shares on social platform
- A piece of content has reached a threshold of engagement (e.g. 3,000 retweets, 5,000 shares, 1,000+ reactions or comments)
- A claim is getting reported to the WhatsApp tip line multiple times
- The publisher (e.g. page or account) has lots of followers

Learning from fact-checkers to research, write, and review a fact check

4.1.1 Overview of the research, writing and review process

Every fact checker has a review process involving at least one other editor checking the quality of evidence, logic of argument, clarity of prose and political balance.

Organizations have different types and levels of checks in place, ranging from up to six layers of editing to a voting system where a minimum of four editors must approve a draft. Some organizations' directors are involved in the editing process daily, whereas in others the director is only involved for controversial topics or tricky fact checks. Usually, these steps are used to come to a fact-check:

- Identify Claim
- Check of existing fact-checks of the same claim
- Identify source of claim
- Consider motivation
- Attempt to contact claimant
- Look for evidence
- Assess quality of primary sources
- Contact press offices and data institutes
- Write draft
- First edit
- Editors read and vote on draft
- Second edit
- Writers and editors to discuss rating
- Final review

Extra steps which are taken when needed include, using freedom of information requests, contacting international colleagues, or seeking expert input. We can look at the process in a bit more detail:

4.1.2 Source of an online claim

Almost all fact checkers see identifying the source of a claim as the key to fact checking it and to understand intentions.

Intentions is hard to prove, but fact checkers do look at what motivation someone might have for sharing or creating a post.

Tracing the origin also reveals where a post has travelled and how it propagated, which can add useful context for both fact checkers and readers. Some fact checkers, such as Animal Politico and Rappler, follow the rabbit hole down to establish whether there is a coordinated network behind a piece of misinformation.

4.1.3 *Review process*

Senior or managing editors in most fact checking organizations review the credibility, quality and sufficiency of evidence used in a fact check. The editor looks at the draft from multiple perspectives and removes any bias they see, and considers possible misinterpretations to avoid backlash.

Editors also review language and writing style. Some, like Africa Check, have a style guide. Others make an effort to break away from journalistic norms within their country.

4.1.4 *Editorial Materials*

Many fact checkers have an established structure or template for drafting articles.

Fact Crescendo says, “We don’t want people left guessing or wrongly assuming that a claim is true, so we use an inverted pyramid, starting with what’s being spread, what is being claimed, and why it’s wrong. We keep our methodology simple, saying how we searched it on Google, and what keywords we used – then a clear conclusion.”

PesaCheck asks writers to answer six questions:

1. What is the claim?
2. Where was it published?
3. Who was it made by, or to whom do we attribute it?
4. Why is it deserving of a full fact check?
5. Is it something that could lead to real world harm that can be avoided by fact checking it? And what impact will fact checking have on public conversation – will it just create more buzz and confusion?
6. What is the verdict?

Chequeado has an eight-step methodology for fact checking misinformation online, developed in collaboration with other Latin American fact checkers and First Draft during a conference in 2019:

1. Select suspicious content from the social networks that are monitored
2. Weigh its relevance
3. Consult, when identifiable, the original source
4. Consult, if identifiable, those involved in/affected by the misinformation
5. Consult the official source
6. Consult alternative sources
7. Give context
8. Confirm or deny the content

PolitiFact has five standard questions that a fact checker answers as part of the research and writing process:

1. What is the claim?
2. Where was it published?
3. Who made the claim?
4. What’s the significance? (This covers potential for real world harm as well as the possible impact on public conversation of fact checking the claim – sometimes a fact check might just be adding to buzz and confusion.)

5. What is our verdict?

4.1.5 Skills

Social media claims vary in terms of topic, format and source. From one day to another, a fact checker might check text posts, videos, images or audio clips presenting fabricated quotes, bogus cures or overblown claims about the performance of the government. This requires knowledge of a wide range of tools and sources as well as human judgement, curiosity and determination.

In general, journalism schools do not teach students basic online verification techniques such as reverse image searching. Experienced journalists do not necessarily have the skills to do online fact checking, either.

Here's what many fact checkers will be able to do after several months on the job:

- Identify text-based and non text-based claims – e.g. which part of a meme is being checked.
- Spot when an image has been fabricated or manipulated and find the original.
- Identify edited videos.
- Search screenshots from videos to identify the original source.
- Construct effective keyword searches.
- Find and use basic statistics such as international population figures or voter registration data.
- Spot psychological tricks that attempt to elicit certain audience reactions.
- Look beyond individual claims to spot patterns and learn how misinformers operate.
- Quickly interpret new online environments and judge their credibility.

4.1.6 Training fact checkers to check online claims

Teyit's training for interns begins with reading translated research, Teyit's own reports, and experts' articles on information disorder and misinformation. Then, interns give a presentation on what they understand about misinformation, how Teyit tackles it, and their own ideas for tackling it. After a week, interns are assigned simple fact checks, graduating to harder and more complex claims which require more than simple reverse image searches. Teyit has a database of over 200 tools with descriptions of their uses which is shared with new staff and interns when they join, including internal training videos. Interns are taught how to use Wordpress and encouraged to try using a wide variety of tools. Teyit also encourages people to pick up the phone.

Fact Crescendo gives new editorial staff a fifteen day structured training induction. It covers:

- Tips on how to spot fake news: critical thinking, looking out for emotional appeals, incomplete details, etc.
- Brief introduction to International Fact-Checking Network's code of conduct and policies.
- Basic tools to analyze content: Simple tools such as reverse image, advanced Google Search, twitter search, translator, etc.
- Monitoring tools such as Crowd Tangle and Tweet Deck.
- Fact Crescendo maintains a list of tools which is regularly updated and is used by new and existing fact checkers.

Fatabyyano has five volunteer team members. People must enter a competition to join the team. Such competitions can involve up to 50 people checking information as quickly as possible. "They have to be able to read English, use basic tools such as reverse image search, and write a short draft of an article. Later we continue to train people depending on which team they join. But everyone joins with basic skills."

4.1.7 Evidence: Tools, websites, and access to data and institutions

Fact checkers mentioned a wide range of tools that they use primarily in research, rather than monitoring. These were the most frequently mentioned in interviews:

- Searching
 - Baidu search
 - Bing search
 - Google advanced search
 - Twitter advanced search
 - Facebook Graph Search (not currently operating - fact checkers want this to be reinstated)
- Video and image verification
 - Amnesty video verification tool
 - InVid
 - Google reverse image search
 - RevEye
 - Tin Eye
 - Yandex Reverse Images Search
 - Fotoforensics
- Archiving pages or locating previously-archived pages
 - Internet Archive
 - Archive.is
 - eyewitness
- Evaluating web pages
 - Website Informer
 - Who.is
 - CrowdTangle Chrome extension
- Other
 - Google Dataset search
 - Google Translate
 - News agencies
 - Newspaper archives

Fact checkers use a wide range of data to check online claims, for example international sources such as the World Bank and the World Health Organization, national statistics bureaus, data produced by NGOs, data obtained through freedom of information laws, archives and legal documents. Access to data, quality of data and publication formats vary from country to country.

4.1.8 Publication and distribution of online fact checks

Fact checkers distribute and publish online fact checks in multiple places: they publish fact checks on their own websites and promote them through their own social media channels.

Most also distribute their online fact checks more widely to reach a bigger audience, for example through media partnerships: sometimes media pay fees and others have arrangements where they republish and reuse fact checks free of charge.

Fact checkers sometimes run online advertisements, and many have received ad credits from internet companies, particularly during elections and during the coronavirus pandemic.

A third distribution method is technology set up by internet companies, such as Facebook's Third-Party Fact-Checking programme, which shows fact checks and fact checkers' ratings to Facebook users, and ClaimReview, which enables Google, YouTube, Bing and others to highlight fact checks in search results and in apps.

The main challenges fact checkers experience in this part of the process include:

- getting set up on new social media channels with limited staff resources,
- presentation of fact checks,
- knowing too little about audiences,
- managing media partnerships,
- internet censorship,
- online harassment,
- getting clear answers and support about how to use ClaimReview and how internet companies use it in their products.

4.1.9 Technology

The most popular proposition was a tool which identifies claims and provides virality metrics alongside them.

Based on interviews, the ideal – though perhaps unrealistic – monitoring tool would:

- Identify claims in a wide range of languages and alphabets.
- Take in data about previously-checked accounts and pages (helping to identify repeat offenders).
- Capture virality and predict the performance of a post.
- Work across Facebook, Twitter, Instagram and YouTube.
- Have video and image search functionality.
- Auto-generate keyword searches based on live data.
- Transcribe speech to text in a wide range of languages.
- Have detailed transparency documentation.

Beyond this, organizations' wish lists were more specific, either to their workflow or country:

- Improving natural language processing in specific languages, e.g. Arabic.
- "Editorial Checklist" WordPress plugin.
- Crowdsourcing platform for micro-research tasks (e.g. converting PDFs to raw data).
- PDF to Excel converter.
- Auto-generating parts of articles, e.g. the CMS suggests a link to frequently used dataset on the topic you are writing about, or auto-fills a sentence about the share count of the claim you're checking, based on the claim's URL.
- Software that flags whether videos are likely to have been altered.
- Instagram Stories monitoring tool.

Technology for automatic identification of claims, crowdsourcing reader tips, and search trends:

- Lead Stories' Trendolizer: identifying emerging viral posts and connecting the dots between known misinformers
- Rappler's shark tank: monitoring which ingests accounts previously identified as spreading misinformation
- Full Fact: claim detection and claim matching
- Tech and Check Cooperative: identifying claims and disseminating fact checks via an app
- Chequeado's Chequeabot: identifying claims in online media outlets
- Aos Fatos's Radar: disinformation monitoring in real time

- RMIT ABC Fact Check: identifying bushfires misinformation on Twitter
- Teyit's crowd-powered website: educating users as they participate in monitoring and research
- Africa Check's WhatsApp chat bot: crowdsourcing WhatsApp misinformation via reader request
- First Draft's coronavirus search trends: briefings on trending coronavirus searches
- Technology to help fact checkers with research
- Forensia: authenticity scores for audio files
- Maldita.es: superpowered community of experts to advise and contribute to research
- Full Fact's robochecking prototype
- Technology to help fact checkers publish and distribute their work:
- Aos Fatos' Fátima: replying to users in Facebook Messenger, and challenging sharing of false information on Twitter
- Coronavirus alliance: searchable global database of coronavirus fact checks

Challenges for technologists

- Social listening tools that combine virality with claim identification
- Claim spotting and matching
- YouTube monitoring tool
- Improving natural language processing in smaller languages
- Searchable image and video misinformation database
- Database for fact checks of claims that go across borders, with internal translation capability
- Speech-to-text transcription for YouTube content that can be connected with claim spotting tools

Summary of challenges

- Monitoring
- Volume and relevance
- Overemphasis on virality from social listening tools
- Inundation with audience requests
- No monitoring tool for YouTube
- Image and video searching
- Research
- Repetitive claims and time consuming or repetitive tasks
- Accessibility of information and transparency of authorities
- Training editorial staff
- Difficulty of finding a source for claims originating from closed platforms
- Publication and distribution:
- Setting up new social media channels
- Sustaining media partnerships
- Presenting fact checks with limited space and design resources
- Internet shutdowns
- Online harassment
- Working with internet companies:
- Financial dependency on internet company funding
- Transparency: both in terms of the full scope and nature of internet companies' responses to online misinformation, and of detailed impact metrics for partnerships with fact checkers or products powered by fact checks
- The need for more investment by more internet companies in partnerships and engagement with fact checkers
- Testing and feedback
- Variation in fact check data requirements of different internet companies products

4.2 RECOGNIZING DISINFORMATION SIGNALS

4.2.1 Key disinformation signals

Those who spread disinformation use several tricks to disseminate it in society. Identifying signals that relate to aspects of disinformation is therefore of high importance to combat its spread. A concise description of main disinformation signals is shown in Table 4.1

[1] provided a list of hints that should be taken into account by the users to fight fake text information. These include catchy headlines, sources that mimic authentic sites by making small changes to the link, stories that come from an unfamiliar organization, unusual formatting, event dates that have been altered, lack of evidence or reliance on unnamed experts, stories reported by only one source and cannot be cross-checked by multiple sources that a user trusts. Similarly, a news item should be examined from different aspects to verify its credibility such as the source, the coherency of the message (in terms of a compelling story without contradictions), the consistency of the information with prior knowledge, or the general acceptability of the message by others [2].

False information has lexical structures that are different from factual reports. For instance, Clickbait headlines – a form of misleading online content – tend to use stop or suspenseful words (e.g., “A 22-Year-Old Whose Husband And Baby Were Killed By A Drunk Driver Has Posted A Gut-Wrenching Facebook Plea”), hyperbolic or exaggerated words (e.g., Awe-inspiring, breathtakingly gut-wrenching, soul-stirring, etc.), Internet slang words (WOW, LOL, LMAO, AMA, AF), punctuation emphasis (!?, ..., ***, !!!), catchy phrases that blow up curiosity (“Will Blow Your Mind”, “You Won’t Believe”), extensive use of numerals (e.g., “Here’s What Happens When You Put A Few Little Kids In A Room With 2 Dolls In 2 Different Colors”), and provocative use of adjectives (“This Is The Most Inspiring Yet Depressing Yet Hilarious Yet Horrifying Yet Heartwarming Grad Speech”) [3,4,5].

Catchy headlines	Fallacies (see Table 4.2)
Unverified source	Bots/cyborgs
Clickbait headlines	Non cross-checked information
Emotional text	Incoherent text
Hate speech	Use of exaggerated, provocative words
Low lexical/content diversity	Low language complexity

Table 4.1: List of key disinformation signals

Stories that count on emotions instead of stating facts can be more persuasive than neutral content, attracting audience selection by producing more positive or negative sentiments, and finally distract readers from potentially more diagnostic cues, such as source credibility [6,7,8,9,10]. This may increase their vulnerability to disinformation. [11] found that the spread of fake news online was positively associated with the greater anger it carries. Studies have also noted that the reliance on emotion is associated with misinformation sharing [12] or believing in conspiracy theories [13].

Deceptive text messages are also associated with lower language complexity (average sentence length, average word length), lower lexical and content diversity (the ratio of unique words), informality (typographical error ratio) and verbal uncertainty [10].

Hate speech coupled with disinformation can lead to stigmatization, discrimination, and large-scale violence. Hate speech attacks a person or group based on attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity [14]. Identifying hate speech is a very challenging task⁷, as it is often related to the content of the text rather than the presence of hateful keywords in it [15]. For

⁷ <https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/>.

instance, slang such as “build that wall” literally means constructing a physical barrier, whereas, in a political context, it may be interpreted as a condemnation of immigrants.

4.2.2 Categories of disinformation signals

Disinformation signals can be categorized in two broad categories, based on the type of content they can be applied on:

- Disinformation signals derived from single items, such as a single news item. Disinformation signals in this category include: catchy/clickbait headlines, non-credible sources, unusual formatting, event dates that have been altered, lack of evidence or reliance on unnamed experts, stories reported by a single source, coherency of the message, the consistency of the information with prior knowledge, the presence of facts (that should be validated), the presence of argumentation (without fallacies), emotional manipulation, language complexity, hate speech, etc.
- Disinformation signals require the analysis of several items, like a set of news items, micro-posts, comments, etc. In this category, signals related to affective polarization, echo chambers, etc. can be considered. (These disinformation signals are harder to implement).

4.2.3 Identifying logical fallacies

Identifying logical fallacies is another method we can use to determine whether online information is valid. A fallacy is a statement or argument based on faulty reasoning; it may “seem to be valid but is not so” [16]. Since there are hundreds of fallacies⁸, we will provide brief explanations and examples for the most common fallacies used in persuasive speech (Table 4.2, (see Annex 1 for a comprehensive list). Information was retrieved from [16], and web resources^{9, 10, 11}

Fallacy	Description	Example
Ad Hominem	It over-simplifies or misrepresents the opponent’s argument by discrediting the person itself, making an irrelevant attack.	“Thompson’s suggestion for the wetlands should be rejected because last year she was arrested for hunting without a license.”
Straw man	It takes the opponent’s argument and creates an exaggerated version of the argument rather than your actual one.	Person 1: "I believe hunting for sport is immoral." Person 2: "Apparently, my opponent wants us all to be vegetarians because animals are more important than people."
Slippery slope	When someone claims that one step in a particular direction will culminate into	“If we allow the government to regulate the economy, then,

⁸ <https://iep.utm.edu/fallacy/#H2>.

⁹

<https://www.sjsu.edu/writingcenter/docs/handouts/Figuring%20out%20Fallacies%20and%20Fake%20News.pdf>.

¹⁰ <https://ggu.libguides.com/c.php?g=810754&p=5785291>.

¹¹ <https://blog.hubspot.com/marketing/common-logical-fallacies>.

	a (usually negative) result later and that it's impossible to stop midway.	eventually, we'll become socialists."
False Dichotomy	The arguer sets up the situation, so it looks like there are only two choices.	"We can either agree with Barbara's plan, or just let the project fail. There is no other option."
Red herring	It attempts to disrupt an opponent's argument by introducing a counter argument that isn't strictly related to the original.	Person 1: "More tax money should be used to build homeless shelters." Person 2: "No, because more homeless would mean more crime."
Post Hoc, Ergo Propter Hoc	It creates a false causal relationship between ideas/events.	"Every time I buy a good seat, our team wins."
Appeal to Authority	It attempts to prove an argument by saying that it is supported by some authority.	"A commercial claim that a specific brand of cereal is the best way to start the day because athlete Michael Jordan says that it is what he eats every day for breakfast."
Tu Quoque	It occurs when someone's argument is discredited solely based on the allegation that their past actions or words are not consistent with their current belief.	<i>Mother:</i> You should stop smoking. It's harmful to your health. <i>Daughter:</i> Why should I listen to you? You started smoking when you were 16!
Naturalistic	It argues that because something happens naturally or derives from nature, that therefore it ought to happen this way.	"Medications interfere with the natural course of events and so should be avoided. It is better to allow the body to heal itself all-naturally."
Hasty generalization	Jumping to conclusions on a topic using one or few examples rather than relying on more extensive research.	"I've met two people in Nicaragua, and they were both nice to me. So, Nigerians are nice people."
Bandwagon fallacy	Claiming that something is true because everyone believes it.	If you're shopping for a smartphone, go with the iPhone 13. It's the best-selling phone.
Appeal to pity	It manipulates someone's feelings of pity in order to get them to support a point of view.	"Commercials that show starving children in Africa

		before asking for donations to feed them.”
Appeal to ignorance	It occurs when a person mistakenly believes something to be true that is not, because he or she does not know enough about the subject.	“There must be fairies living in our attic because nobody’s ever proven that there aren’t fairies living in our attic.”

Table 4.2: Common fallacies identified in persuasive speech

4.2.4 State-of-the-art AI models, data and approaches for detecting fact-checking, fallacies and hate speech

A detailed list of fact-checking, fallacies and hate speech datasets along with adopted AI approaches and codes are listed in the supplementary **Annex 1**.

4.3 REFERENCES

- [1] Facebook, 2018. Tips to spot false news. <https://www.facebook.com/help/188118808357379>.
- [2] S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, “Misinformation and Its Correction: Continued Influence and Successful Debiasing,” *Psychol. Sci. Public Interes. Suppl.*, vol. 13, no. 3, pp. 106–131, 2012, doi: 10.1177/1529100612451018.
- [3] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, “Stop Clickbait: Detecting and preventing clickbaits in online news media,” *Proc. 2016 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2016*, pp. 9–16, 2016, doi: 10.1109/ASONAM.2016.7752207.
- [4] Y. Chen, N. J. Conroy, and V. L. Rubin, “Misleading online content: Recognizing clickbait as ‘false news,’” *WMDD 2015 - Proc. ACM Work. Multimodal Decept. Detect. co-located with ICMI 2015*, no. November, pp. 15–19, 2015, doi: 10.1145/2823465.2823467.
- [5] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” *EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 2931–2937, 2017, doi: 10.18653/v1/d17-1317.
- [6] J. Berger and K. L. Milkman, “What makes online content viral?,” *J. Mark. Res.*, vol. 49, no. 2, pp. 192–205, 2012, doi: 10.1509/jmr.10.0353.
- [7] U. K. H. Ecker *et al.*, “The psychological drivers of misinformation belief and its resistance to correction,” *Nat. Rev. Psychol.*, vol. 1, no. 1, pp. 13–29, 2022, doi: 10.1038/s44159-021-00006-y.
- [8] H. S. Kim, “Attracting Views and Going Viral: How Message Features and News-Sharing Channels Affect Health News Diffusion,” *J. Commun.*, vol. 65, no. 3, pp. 512–534, 2015, doi: 10.1111/jcom.12160.
- [9] J. Paschen, “Investigating the emotional appeal of fake news using artificial intelligence and human contributions,” *J. Prod. Brand Manag.*, vol. 29, no. 2, pp. 223–233, 2020, doi: 10.1108/JBPM-12-2018-2179.
- [10] L. Zhou and D. Zhang, “Following linguistic footprints: Automatic deception detection in online communication,” *Commun. ACM*, vol. 51, no. 9, pp. 119–122, 2008, doi: 10.1145/1378727.1389972.
- [11] Y. Chuai and J. Zhao, “Anger can make fake news viral online,” *Front. Phys.*, vol. 10, no. August, pp. 1–12, 2022, doi: 10.3389/fphy.2022.970174.
- [12] B. E. Weeks, “Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation,” *J. Commun.*, vol. 65, no. 4, pp. 699–719, 2015, doi: 10.1111/jcom.12164.
- [13] C. Martel, G. Pennycook, and D. G. Rand, “Reliance on emotion promotes belief in fake news,” *Cogn. Res. Princ. Implic.*, vol. 5, no. 1, 2020, doi: 10.1186/s41235-020-00252-3.
- [14] S. Wermiel, “The Ongoing Challenge to Define Free Speech,” *Law Rev. Other Acad. Journals*, vol. 43, no. 4, p. 82, 2018.
- [15] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, “Hate speech detection: Challenges and solutions,” *PLoS One*, vol. 14, no. 8, pp. 1–16, 2019, doi: 10.1371/journal.pone.0221152.
- [16] Tindale, C. W. (2007). *Fallacies and argument appraisal*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511806544>

5 LEGAL AND ETHICAL DIMENSIONS OF TRUSTWORTHY AI TO FIGHT DISINFORMATION

AI plays a role in spreading disinformation, used in social media platforms to categorize content and even create new content. As an example, new digital engagement tools, such as ‘clickbait’ – which refers to “[w]eb content designed to generate attention and online advertising revenue at the expense of quality or accuracy, relying on sensationalist headlines or eye-catching pictures to attract click-throughs and shares” [27] – have facilitated the spread of disinformation by increasing interaction. Although discouraging such practices, social media through algorithmic prioritization of highly engaging and viral content tend to challenge the veracity of the news. In this regard, Nelson and Taneja [28] demonstrated how the visitors to fake news sites mainly come from social media networks and how online media increases the exposure to disinformation among the population.

The European Union promotes a Digital Agenda for the decade 2020-2030 [29] based on these pillars: (1) Build a technology that works for people (2) a fair and competitive digital economy, and (3) an open, democratic and sustainable society (4) Europe as a global leader for the digital economy (5) A technology with the aim of benefiting citizens, business and the environment. The digital landscape continues to change rapidly, and the EU has launched new regulatory actions to limit the negative impact of digital technologies, to build a safer digital landscape, focusing on protecting people’s fundamental rights. Most relevant to the TITAN project are the newly proposed AI Act [30] and the Digital Services Act [31]. In this section, we offer an overview of the most pressing ethical issues and the legal dimensions, following EU’s digital agenda, for the creation of trustworthy AI to fight disinformation.

5.1 THE CONCEPT OF TRUSTWORTHY AI

The concept of Trustworthy AI was developed in 2019 by the High-Level Expert Group on AI (AI HLEG)[32] in the Ethics Guidelines for Trustworthy Artificial Intelligence [33]. They propose three key requirements for AI to be trustworthy, that should be met throughout the entire system life cycle: (1) lawful, (2) ethical and (3) robust; the latter meaning that it should also be technically strong as AI systems can cause unintentional harm. In these guidelines, the concept of Trustworthy AI is defined as follows:

“Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the system’s life cycle” [page 38, AI HLEG Ethics guidelines].

The guidelines are the basis for further regulatory action that is being developed, such as the AI Act. In practice, however, the expert group also points out that tensions can occur between law and ethical norms and that should be balanced. Trustworthy AI is based, according to the group on four main principles, which translate into seven key requirements: (1) *human agency and oversight*, (2) *technical robustness and safety*, (3) *privacy and data governance* (4) *transparency* (5) *diversity, non-discrimination and fairness* (6) *societal and environmental well-being* (7) *accountability*. The AI HLEG translated these requirements into a detailed Assessment List [34] (in a question-based template) and developed a prototype web-based tool [35], to guide developers and deployers of AI through an accessible and dynamic checklist.

5.2 ETHICAL DIMENSIONS

We focus on the AI HLEG Ethics Guidelines for Trustworthy Artificial Intelligence, where the fundamental principles for a trustworthy AI are envisioned. However, the AI HLEG is not the only group pursuing ethical principles of AI. This is a field of study in ethics of technology, and a sub-field of applied ethics, that has grown especially during the past decade, and has developed different frameworks that offer different values or

requirements. For instance, the SIENNA project has proposed a framework for an ethical guideline on AI and Robotics [36]. The project SHERPA [37] in 2020 also produced guidelines for an ethical use of AI with seven similar requirements: (1) Human agency, liberty, and dignity; (2) Technical robustness and safety; (3) Privacy and data governance; (4) Transparency; (5) Diversity, non-discrimination, and fairness; (6) Individual, societal, and environmental wellbeing; (7) Accountability. SHERPA developed an AI Impact Assessment Tool that focuses on risks but also on ethical issues [38].

Recently, NATO’s Data and Artificial Intelligence Review Board (DARB) developed certification standards [39] for a responsible AI to increase trust in the sector from different stakeholders, particularly focusing on six principles (1) lawfulness (2) responsibility and accountability (3) explainability and traceability (4) reliability (5) governability and (6) bias mitigation.

There are ethics guidelines, such as the publicly available AI Ethics tools [40], that go beyond AI HLEG trustworthy principles, although similar, expanding different methodologies available in AI Ethics that extend the principles and values to build an Ethical AI. In 2022, UNESCO published a recommendation on the Ethics of Artificial Intelligence [41] based on principles such as: (1) Proportionality and no harm (2) fairness and non-discrimination (3) safety and security (4) Sustainability (5) Right to privacy and data protection (6) Human oversight and determination (7) Transparency and explainability (8) Responsibility and accountability (9) Awareness and literacy (10) Multi-stakeholder and adaptive governance and collaboration.

In the AI HLEG Ethics Guidelines, there is no specific mention of ethical frameworks, the expert group selected four main (ethical) principles and that were the basis for seven requirements (see Image 1). For an AI to be ethical (but differs from being trustworthy), it should follow these four ethical principles that are, according to the AI HLEG, intrinsically linked to and grounded on fundamental rights. In chapter 1 of the guidelines [32], the AI HLEG describes the four ethical principles that should lead to an ethical AI as follows: (1) respect for human autonomy, (2) prevention of harm, (3) fairness and (4) explicability. Next, the AI HLEG translates these principles into 7 key requirements [33] (chapter 2) and the assessment list [34] aims to help with the application of each key requirement (chapter 3):

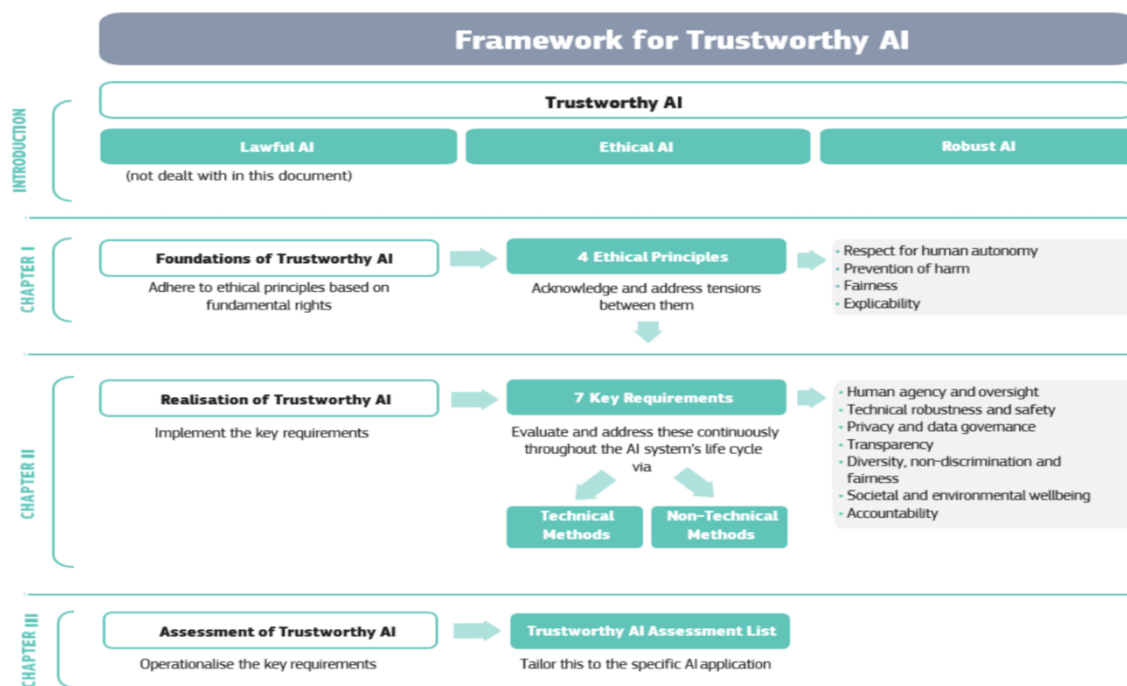


Figure 5.1: Ethics Guidelines for Trustworthy AI [Source: HLEG, A. (2019). Ethics Guidelines for Trustworthy Artificial Intelligence. High-Level Expert Group on Artificial Intelligence]

The expert group further mentions that for an AI to be ethical, it must also acknowledge and address potential tensions between ethical principles, and the experts highlight the importance to pay attention to the most vulnerable groups such as children, or at risk of exclusion, and to asymmetries of power or information e.g., between employers and workers, or business and consumers. Particularly for TITAN, considerations of the vulnerability of citizens with the introduction of tools that aim to guide their knowledge will be considered. AI HLEG also highlights the necessity to consider those risks that are difficult to anticipate and thus aim for the flexibility of the system which will require adopting measures and mitigating risks.

In the following paragraph, the four ethical principles and their relation to the protection of the EU fundamental rights (hereafter The Charter) [42] are summarized:

(1) Respect for human autonomy: there are two fundamental rights that play a role – the right to freedom of expression and information (Art. 11, The Charter) and the right to autonomy (Art. 3, The Charter). The expert group highlights that “[h]umans interacting with AI systems must be able to keep full and effective self-determination over themselves and be able to partake in the democratic process”. The AI should not manipulate or coerce, and it should be designed to encourage positive attributes such as cognitive and cultural skills e.g., meaningful work. It should also allow for a “meaningful opportunity for human choice”, and this is related to, as the AI HLEG argues, the importance of securing human oversight within AI systems.

(2) Prevention of harm: related to the fundamental right of human dignity (Art.1, The Charter) and the mental and physical integrity (Art.3). The AI systems must always be safe and secure, preventing wrong use or intent. Particularly, the system must be technically robust. AI systems must prevent harm by technical means, but also must pay attention to the social construct such as power asymmetries (e.g., between companies and users).

(3) Fairness: here there is a distinction between two dimensions of fairness: substantive and procedural. The substantive refers to “ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation” (Art. 21 non-discrimination, The Charter). Thus, AI systems must be free from unfair biases. However, it also considers the principle of proportionality, a procedural understanding of fairness, to balance the interests of the parties involved. According to the expert group, fairness relates to the decision-making process, thus important for this principle is the necessity to identify within an AI system the entity accountable for the decision-making process.

(4) Explicability: relevant for maintaining trust. It means that AI processes need to be transparent, informative, where the purposes and decisions of the AI system are communicated. Here it is important to consider the problem of so-called ‘black boxes’. The HLEG argues that “In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required [...] the degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate”.

More recently, in 2021, the Commission also drafted a non-binding Ethics by Design and Ethics of Use Approaches for Artificial Intelligence [43] specifically regarding research that uses AI systems. This guideline offers a more practical approach to an ethical AI and allows to operationalize the seven requirements. It follows the work of the HLEG and other EU projects, SHERPA and SIENNA. It is divided into three parts, which we will summarize (and changed the order) for our own purposes below:

5.2.1 Section 1– How to apply ethics by design in AI Development: practical steps

This section is dedicated to preventing ethical issues from arising before the system is put into place. Ethics by design is created in a 5-layer model:

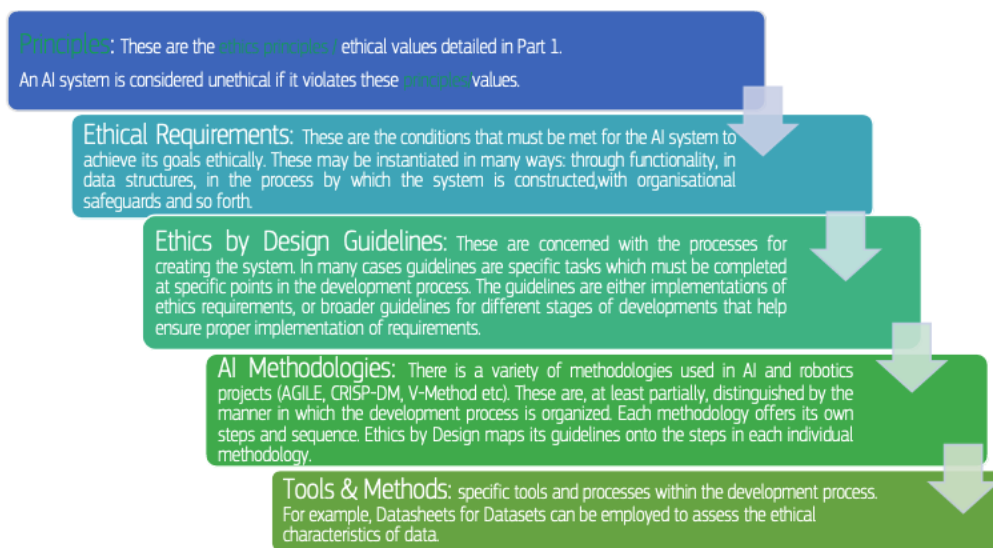


Figure 5.2: Ethics by design Steps [Retrieved from: Ethics by design and ethics of use approaches for artificial intelligence. 2021. Page 11]

This section follows the AI-HLEG seven requirements. It summarizes them into 6 main ethical principles that an AI must preserve and that should be considered in Step 1 of the 5-layer model:

<p>Respect for Human Agency</p> <p>It involves 3 specific principles</p>	<p>Autonomy -> AI based systems should avoid restricting unreasonably individual decision-making context, and to respect people to think for themselves.</p> <p>Dignity -></p> <p>Freedom-></p>	<p><i>Human Agency: General Ethical Requirements</i></p> <ul style="list-style-type: none"> - End-users and others affected by the AI system MUST NOT be deprived of abilities to make basic decisions about their own lives or have basic freedoms taken away from them. - It MUST be ensured that AI applications do not autonomously and without human oversight and possibilities for redress make decisions: about fundamental personal issues (e.g. affecting directly private or professional life, health, well-being or individual rights), that are normally decided by humans by means of free personal choices; or about fundamental economic, social and political issues, that are normally decided by collective deliberations, or similarly significantly affects individuals. - End-users and others affected by the AI system MUST NOT be in any way subordinated, coerced, deceived, manipulated, objectified or dehumanized. - Attachment or addiction to the system and its operations MUST not be purposely stimulated. This should not happen through direct operations and actions of the system. It also should be prevented, as much as possible, that systems can be used for these purposes. - AI applications should be designed to give system operators and, as much as possible, end-users the ability to control, direct and intervene in basic operations of the system. - End-users and others affected by the AI system MUST receive comprehensible information about the logic involved by the AI, as well as the significance and the envisaged consequences for them. <p>[Retrieved from: Ethics by design and ethics of use approaches for artificial intelligence. 2021. Page 6]</p>
<p>Privacy and Data governance</p>	<p>This involves complying with data protection requirements expelled in the EU’s General Data Protection Regulation (see the section below on ‘Legal Dimensions’) of data protection by design and by default.</p>	

	<p>The guidelines rightly highlight that “ethical issues can arise not only when processing personal data but also when the AI system uses <i>non-personal data</i> (e.g. racial bias)”. And advises considering further requirements on ‘data trust’ that could help research participants negotiate data use, especially if there are different types of data. To protect and manage personal data and the way systems use it. Appropriate data protection from the beginning allows for trust in the system.</p>	<p><i>Privacy and Data Governance: General Ethical Requirements</i></p> <ul style="list-style-type: none"> - The AI systems MUST process personal data in a lawful, fair and transparent manner. - The principles of data minimisation and data protection by design and by default MUST be integrated in the AI data governance models. - Appropriate technical and organisational measures MUST be set in place to safeguard the rights and freedoms of data subjects (e.g. appointment of data protection officer, anonymization, pseudonymisation, encryption, aggregation). Strong security measures MUST be set in place to prevent data breaches and leakages. Compliance with the Cybersecurity Act⁴ and international security standards may offer a safe pathway for adherence to the ethical principles. - Data should be acquired, stored and used in a manner which can be audited by humans. All EU funded research must comply with relevant legislation and the highest ethics standards. This means that all Horizon Europe beneficiaries must apply the principles enshrined in the GDPR. <p>[Retrieved from: Ethics by design and ethics of use approaches for artificial intelligence. 2021. Page 7]</p>
<p>Fairness</p>	<p>Referred to here as non-discrimination of one’s own identity (such as discrimination based on race, gender, sexual orientation, national origin, religion, health, and disability). The guidelines mentioned how the AI systems must be designed in a way that does not disadvantage specific groups or individuals, regarding two parts: procedural and substantive fairness. Procedural fairness means and substantive fairness:</p>	<p><i>Fairness: General Ethical Requirements</i></p> <ul style="list-style-type: none"> - <i>Avoidance of algorithmic bias:</i> AI systems should be designed to avoid bias in input data, modelling and algorithm design. Algorithmic bias is a specific concern which requires specific mitigation techniques. Research proposals MUST specify the steps which will be taken to ensure data about people is representative of the target population and reflects their diversity or is sufficiently neutral. - Similarly, research proposals should explicitly document how bias in input data and in the algorithmic design, which could cause certain groups of people to be represented incorrectly or unfairly, will be identified and avoided. This necessitates considering the inferences drawn by the system which have the potential to unfairly exclude or in other ways disadvantage certain groups of people or single individuals. - <i>Universal accessibility:</i> AI systems (whenever relevant) should be designed to be usable by different types of end-users with different abilities. Research proposals are encouraged to explain how this will be achieved, such as by compliance with relevant accessibility guidelines. To the extent possible, AI systems should avoid functional bias by offering the same level of functionality and benefits to end-users with different abilities, beliefs, preferences, and interests. - <i>Fair impacts:</i> Possible negative social impacts on certain groups, including impacts other than those resulting from algorithmic bias or lack of universal accessibility, may occur in the short, medium and longer term especially if the AI is diverted from its original purpose. This MUST be mitigated. The AI system MUST ensure that it does not affect the interests of relevant groups in a negative way. Methods to identify and mitigate negative social impacts in the medium and longer term should be well documented in the research proposal. <p>[Retrieved from: Ethics by design and ethics of use approaches for artificial intelligence. 2021. Page 8]</p>
<p>Individual, Social and Environmental Well-being</p>		<p><i>Well-being: General Ethical Requirements</i></p> <ul style="list-style-type: none"> - AI systems MUST take into account all end-users and stakeholders and must not unduly or unfairly reduce their psychological and emotional well-being.

		<ul style="list-style-type: none"> - AI systems should empower and to advance the interests and well-being of as many individuals as possible - AI development MUST be mindful of principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects. Whenever relevant, there should be documented efforts to consider the overall environmental impact of the system and the Sustainable Development Objectives, where needed, steps to mitigate it. In the case of embedded AI this must include the materials used and decommissioning procedures. - AI systems that can be applied in the area of media, communications, politics, social analytics, behavioural analytics online communities and services MUST be assessed for their potential to negatively impact the quality of communication, social interaction, information, democratic processes, and social relations (for example by supporting uncivil discourse, sustaining or amplifying fake news and deepfakes, segregating people into filter bubbles and echo chambers, creating asymmetric relations of power and dependence, and enabling political manipulation of the electorate). Mitigating actions must be taken to reduce the risk of such harms. - AI and robotics systems MUST not reduce safety in the workplace. Whenever relevant, the application should demonstrate consideration of possible impact on workplace safety, employee integrity and compliance standards, such as with IEEE P1228 (Standard for Software Safety). <p>[Retrieved from: Ethics by design and ethics of use approaches for artificial intelligence. 2021. Page 8-9]</p>
<p>Transparen cy & explainabi ty and objection</p>		<p><i>Transparency: General Ethical Requirements</i></p> <ul style="list-style-type: none"> - It MUST be made clear to end-users that they are interacting with an AI system (especially for systems that simulate human communication, such as chatbots). - The purpose, capabilities, limitations, benefits, and risks of the AI system and of the decisions conveyed by it MUST be openly communicated to end-users and other stakeholders, including instructions on how to use the system properly. - When building an AI solution, one MUST consider what measures will enable the traceability of the AI system during its entire lifecycle, from initial design to post-deployment evaluation and audit or in case its use is contested. - Whenever relevant, the research proposal should offer details about how decisions made by the system will be explainable to users. Where possible this should include the reasons why the system made a particular decision. Explainability is a particularly relevant requirement for systems that make decisions or recommendations or perform actions that can cause significant harm, affect individual rights, or significantly affect individual or collective interests. - The design and development processes MUST address all the relevant ethical issues, such as the removal of bias from a dataset. The development processes (methods and tools) MUST keep records of all relevant decisions in this context to allow tracing how ethical requirements have been met. <p>[Retrieved from: Ethics by design and ethics of use approaches for artificial intelligence. 2021. Page 9-10]</p>
<p>Accountabil ity and Oversight</p>	<p>Taking responsibility for the way that these applications function and for the resulting consequences.</p>	<p><i>Accountability and Oversight: General Ethical Requirements</i></p> <ul style="list-style-type: none"> - It MUST be documented how possible ethically and socially undesirable effects (e.g. discriminatory outcomes, lack of transparency) of the system will be detected, stopped, and prevented from reoccurring. - AI systems MUST allow for human oversight and control over the decision cycles and operation, unless compelling reasons can be provided which demonstrate such oversight is not required. Such a justification should explain how humans will be able to understand the decisions made by the system and what mechanisms will exist for humans to override them. - To a degree matching the type of research being proposed (e.g. basic or precompetitive) and as appropriate, the research proposal should include an evaluation of the possible ethics risks related to the proposed AI system. This should include also the risk assessment procedures and the mitigation measures after deployment. - Whenever relevant, it should be considered how end-users, data subjects and other third parties will be able to report complaints, ethical concerns, or adverse events and how these will be evaluated, addressed and communicated back to the concerned parties. - As a general principle, all AI systems should be auditable by independent third parties (e.g. the procedures and tools available under the XAI approach² support best practice in this regard). This is not limited to auditing the decisions of the system itself, but covers also the procedures and tools used during the development process. Where relevant, the system should generate human accessible logs of the AI system's internal processes.

Table 5.1: Layer model of ethics by design [Retrieved from: Ethics by design and ethics of use approaches for artificial intelligence. 2021. Page 12]

Ethics by Design Guidelines: this level refers to the development process for AI and robotic systems, and it is described as divided into 6 phases that incorporate Ethics by Design into the development process. These do not intend to be in order, but rather can be an iterative process:



Figure 1: The Generic Model for AI Development

Figure 5.3: Generic Model for AI Development [Retrieved from: *Ethics by design and ethics of use approaches for artificial intelligence*. 2021. Page 12]

The six tasks in the generic model are:

1. Specification of objectives: The determination of what the system is for and what it should be capable of doing. Here the system's objectives are evaluated against the ethical principles and requirements that were mentioned in 1.1 and 1.2.
2. Specification of requirements: Development of technical and non-technical requirements for building the system, including initial determination of required resources, together with an initial risk assessment and cost-benefit analysis, resulting in a design plan.
3. High-level design: Development of a high-level architecture, where "Ethical requirements should be treated just the same as any other requirements for the system" (p.15), such as transparency and oversight will probably require additional features beyond what is needed to achieve the system's aim. In page 16 are shown 21 points that could be applied to the high-level design face:
4. Data collection and preparation: Collection, verification, cleaning and integration of data: in this section it should be assumed that any data gathered is biased. Fairness and accuracy are the primary concerns here. In general, data gathered from human activity may reflect the biases in that society. It must therefore be actively demonstrated that data is accurate, representative or neutral before it can be trusted as such. Steps should be taken to ensure that testing learning and/or algorithmic manipulation do not introduce new biases or give rise to other ethical issues (such as de-anonymization).
5. Detailed design and development: This refers to the *actions* which will incorporate the ethical requirements that are added to the various tasks in the detailed design process, as well as to the development infrastructure (tools, methodologies, procedures, and anything else that may affect exactly how something is built).
6. Testing and evaluation: Testing and evaluation of the system: As part of the testing and evaluation phase, here it is use the project's ethical requirements checklist to design a testing regime which can check the system's ethical compliance (that will be done in **TITAN's D.1.3: Ethical and Legal Impact Assessment**).

AI Methodologies: Related to the previous Section 1.3, here it refers to project management, and monitoring, where external ethical boards come into play.

5.3 LEGAL DIMENSIONS

The HLEG expert group, and additional ethical guidelines, are non-binding instruments and do not deal with legislation but refer to the pertinent regulations. In this section, we summarize the most prominent regulations that must be considered for TITAN in relation to AI development:

5.3.1 General Data Protection Regulation (GDPR)

The **General Data Protection Regulation (GDPR) of 2018 [44]**, related to the ethical principle of data governance, ensures the protection of personal data, for any system that gathers personal information from individuals. To ensure the fair processing of personal data, TITAN will implement the Data Management Plan and the (Ethical) and Legal Impact Assessment, where TITAN will incorporate the Data Protection Impact Assessment, for the sensible processing of data such as religious beliefs, which will take measures to safeguard personal data gathered by the TITAN system. The GDPR is especially relevant, as TITAN will potentially use sensitive data from individuals, considered ‘special categories’ of data, such as political inclinations or religious beliefs. The GDPR also sets standards for data protection by design and by default, such as anonymization techniques that will be considered. In this regard, the GDPR clearly indicates (art. 35) the importance to conduct a Data Protection Impact Assessment (DPIA) if the processing is likely to result in a high risk to the rights and freedoms of individuals.

The GDPR also safeguards the right not to be subject to a decision based solely on automated means, including profiling (Art. 22, Recital 71). This article is important in relation to explainability of AI systems. In this regard, Art.13 and 14 list the information that needs to be provided when personal data is collected and specifies (Art.13 (f)) that it must be communicated the existence of any automated decision-making, the logic involved and the consequences of such processing. In addition, Art.14 stipulates that, when the data has not been obtained from the data subject her/himself, it need to provide additional information to ensure fair and transparent processing, and even more, the existence of the right to withdraw consent at any time (Art 14(2d)).

The European Data Protection Board (EDPB), responsible for the application of the GDPR, issues guidelines that clarify the law and are applicable and complement the GDPR. Recently, in February 2023 the EDPB issued guidelines on ‘Deceptive Design Patterns in Social Media Platforms’ [45], that aim to limit the nudging of people. The guidelines offer practical recommendations for the designers and users of the platforms on how to assess and avoid deceptive design patterns that go against the principles of GDPR, thus unlawful. In addition, some clarifications were added, for example on how to integrate it into the design process (Annex 2 of the guidelines). Further relevant documents of the EDPB are shown in the table below:

Name and Specifications	Link to EDPB document
Guidelines 03/2022 on deceptive design patterns in social media platform interfaces: how to recognize and avoid them	https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en
Guidelines 01/2022 on data subject rights - Right of access	https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-012022-data-subject-rights-right_en
Guidelines 07/2020 on the concepts of controller and processor in the GDPR	https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-072020-concepts-controller-and-processor-gdpr_en
Guidelines 02/2021 on virtual voice assistants	https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-022021-virtual-voice-assistants_en
Guidelines 4/2019 on Article 25 Data Protection by Design and by Default	https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-42019-article-25-data-protection-design-and_en

EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)	https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-52021-proposal_en
--	---

Table 5.2: Relevant EDPB guidelines on artificial intelligence

5.3.2 Artificial Intelligence Act

In 2021, the European Commission released a *proposal* for the **Artificial Intelligence Act (AI Act)** [46] enshrined in EU law, and following the guidelines of AI HLEG, adopting a set of rules for Artificial Intelligence systems tailored to fundamental rights and following a risk-based approach, that aims to address the challenges of these new systems [47].

The difference with Ethics Guidelines for Trustworthy AI is that the AI Act is meant to be binding, whereas the guidelines are not. However, in 2021, following recent scandals in AI, and especially post-Covid increase in tracking technologies and disinformation online [48], the EU shifted towards a legislative approach, known as the new EU Digital Strategy, with a set of new regulations [49]. Specifically for TITAN, the AI Act is very relevant, although it is still being discussed, as it sets out the rules for AI systems. The AI Act adds to the GDPR, which only partially covers the regulation of AI systems, mainly through processing personal data and protecting data subjects' rights.

The AI Act is still under revision, but we can point out several characteristics that we consider to be essential for the legal foundations of TITAN – the risk-based approach. The proposed AI Act will introduce a risk-based approach to regulate AI systems, with this solution, the legal obligations are tailored to different risk levels, distinguishing between 1) unacceptable risk, 2) high risk 3) limited risk 4) low and minimal risk. **At the time of writing this deliverable, the AI Act considered chatbots under limited risk; which was further amended with the rapid spread and evolution of Large Language Models (such as open AI GPT-4, Google's Bar, ChatGPT) in 2023. Currently, there are "high-risk" chatbots. However, other systems that fall under limited risks are AI systems that manipulate image, audio or video content, emotion recognition systems, video games or spam filters. etc. These systems relate, to a certain extent, to fake news.** Disinformation is not merely, as it has been shown, the fault of AI systems (see the introduction of the deliverable, subsection definitions) but has been present in society in different forms. However, certain AI-enabled means could lead to the propagation of disinformation. Whereas the AI Act considers it limited risk at the first instance, and was further amended, TITAN takes this issue very seriously and we will implement measures to reduce risks in our system.

The AI systems related to TITAN are considered 'limited risks', which contain *ad hoc* transparency requirements in the AI Act. In addition, the creation of *optional* codes of conduct to encourage providers of non-high-risk AI systems to voluntarily apply the requirements for high-risk AI systems (Title IX, AI Act). Chatbots are thus encouraged to adopt codes of conduct and *obliged to* a set of limited transparency obligations, that are reduced to article 52, specified below:

"Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use" [Article 52:1 AI Act].

Regulatory Sandboxes are also part of the AI Act, which falls under the scope of TITAN (Article 53-55, AI Act) that allows the lawful creation of "a controlled environment that facilitates the development, testing, and validation of innovative AI systems for a limited time before their placement on the market". However, the AI Act is still being developed and does not specify what would the sandboxes require and the involvement of the Member States in the decision.

Turning to other relevant EU policy, the new EU Digital Strategy proposes a broad set of regulations for the digital age [50] that combines: Digital Service Act (DSA) [31], Digital Markets Act (DMA), Data Act (DA), Data Governance Act (DGA) and the already mentioned AI ACT, among others. Particularly relevant for this project is the DSA that focuses on social networks, it will be summarized below.

5.3.3 Digital Services Act (DSA)

DSA sets new standards for a more transparent and safer online environment, especially for children and illegal content, that aim to regulate Big Tech in the digital landscape. It does not directly apply to TITAN, as the DSA covers large platforms (named gatekeepers) – intermediaries that connect consumer goods, services and content, such as online marketplaces and others – and imposes the obligation to *self-assess*, on a yearly basis, the ‘systemic risks’ arising from the operation of their services and take appropriate mitigation measures. Thus the DSA regulates how companies – online platforms, search engines, online marketplaces, and every other significant provider of digital services – moderate and manage content (including illegal content, hate speech and disinformation) and has several implications for children. The DSA is thus an approved regulatory action (different from the AI Act), which is available since October 2022 and it is expected to be applicable from January 2024 [51]. In this regard, there are emerging scholarly studies that aim to summarize and explain this novel regulation [52], in an already complex digital landscape. Furthermore, other studies attempt to summarize the provisions specifically for content-sharing platforms both for the DSA and DMA as these are interconnected [53].

It introduces principles for content moderation, responsible design and algorithmic accountability, to limit the business incentives of websites that use disinformation and clickbait to attract advertising revenues. These principles can still guide TITAN’s development.

We have summarized [54] some of the obligations that could be relevant for TITAN:

Obligation DSA imposes to platforms	Short summary
Measures against illegal content online, including illegal goods and services	The DSA imposes new mechanisms allowing users to flag illegal content online, and for platforms to cooperate with specialised ‘trusted flaggers’ to identify and remove illegal content;
New rules to trace sellers on online market places	Sustained efforts to enhance the traceability of products through advanced technological solutions
New safeguards for users	Including the possibility to challenge platforms’ content moderation decisions based on a new obligatory information to users when their content gets removed or restricted
Transparency measures for online platforms	Such as transparency on the algorithms used for recommending content or products to users
New obligations on the protection of minors on any platform in the EU	
Obligations for very large online platforms and search engines to prevent abuse of their systems by taking risk-based action	Including oversight through independent audits of their risk management measures. Platforms must mitigate against risks such as disinformation or election manipulation, cyber violence against women, or harms to minors online. These measures must be carefully balanced against restrictions of freedom of expression, and are subject to independent audits;

Bans on targeted advertising on online platforms	Such as profiling children or based on special categories of personal data such as ethnicity, political views or sexual orientation, enhanced transparency for all advertising on online platforms and influencers’ commercial communications
Ban on ‘dark patterns’	Referring to misleading tricks that manipulate users into choices they do not intend to make;
New provisions to allow access to data to researchers of key platforms	In order to scrutinize how platforms work and how online risks evolve;
Users will have new rights, including a right to complain to the platform, seek out-of-court settlements, complain to their national authority in their own language, or seek compensation for breaches of the rules	

Table 5.2 [Summary on the basis of: Questions and Answers, Digital Service Act; European Commission; 14 November 2020: https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348]

For TITAN, which is not a communication platform but rather a system that will be deployed with or along large online platforms, it is still necessary to analyze how a system like TITAN could benefit from the obligations required about content in social networks (e.g., Section 3 DSA) and the reduction of risks. This leaves the question: Can TITAN be implemented within online platforms to help mitigate some risks, such as dark patterns and illegal content in the social network?

5.3.4 Fundamental rights

The spread of disinformation not only affects secondary law but also fundamental (human) rights. As previously exposed, it is especially related to the right to data protection (Art.8 EU Charter), and the right to non-discrimination (Art. 21 EU Charter), but also linked to the freedom of expression and information (Art. 11 EU Charter) [42]. Studies have shown how there is a fine line between freedom of expression – meaning to protect people’s freedom to engage, comment or post online – and the restriction of a certain amount of content that is considered misinformative or malign, and the need to balance between these two rights. For example, recent studies have shown how digital platforms are accountable for the spread of gender disinformation that affects women, where harmful narratives are promoted through algorithmic decisions that consider this type of content viral or engaging [55], arguing how toxic masculinity is spreading online even more than in real life, naming it ‘manosphere’. This study argues that gender disinformation deserves specific attention in social networks, where social media platforms should develop tools to combat this issue, such as hate campaigns or online harassment of women’s politicians. And shows throughout the case studies how gender disinformation in social networks e.g., through recommender systems and algorithms (AI) that are trained to maximize attention, influences politics by reducing women’s voices and therefore having a negative effect on democratic values such as women’s rights [56].

The study has argued how the right to freedom of expression needs to be balanced with the negative effects of disinformation that propagate harmful content. The same study concludes how the DSA has tackled some of these issues, by imposing obligations to platforms to conduct risk assessments at least once a year and improve content moderation activities, such as manipulation with bots or fake accounts. Nonetheless, it also points out how the DSA lacks remedies for gender-based harms online and does not tackle this issue in detail.

5.3.5 International and national developments and best practices in AI

There are international and national developments and best practices in AI that are worthy of mentioning. Regarding international developments, in 2019 the Organization for Economic Co-Operation and Development (OECD) released also principles for a trustworthy AI [57] and the respect for human rights and democratic values, which are similar to the HLEG on AI. It also relates to The Global Partnership on Artificial Intelligence (GPAI) on Climate Change and AI, which have given recommendations for Government Action [58]. The OECD guidelines refer to:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
- Organizations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

In June 2020, the Ad Hoc Committee on AI (CAHAI) at the Council of Europe, released AI Ethical Guidelines [59] focusing on privacy among other principles – transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trustworthiness, sustainability, dignity, solidarity. In December 2021, CAHAI also released the possible elements of a legal framework on AI, based on the Council of Europe standards on human rights [59].

Similarly, in 2021 UNESCO released recommendations on Artificial Intelligence [41].

Particularly relevant for disinformation and hate in online contexts, and not focusing on the issue from a Trustworthy AI principle, there are national policies addressing the issue of disinformation. The study from 2021 conducted by the Global Disinformation Index [60], a non-profit group, provides a thorough analysis of disinformation practices in a few countries. Also relevant is the study by the same group on 44 tech companies’ publisher policies [61] (Figure 5.4) concluding that companies’ policies are not sufficient nor comprehensive enough, and do not address the different types of disinformation.

Figure 1. Sample of publisher policy coverage on six adversarial narrative topics

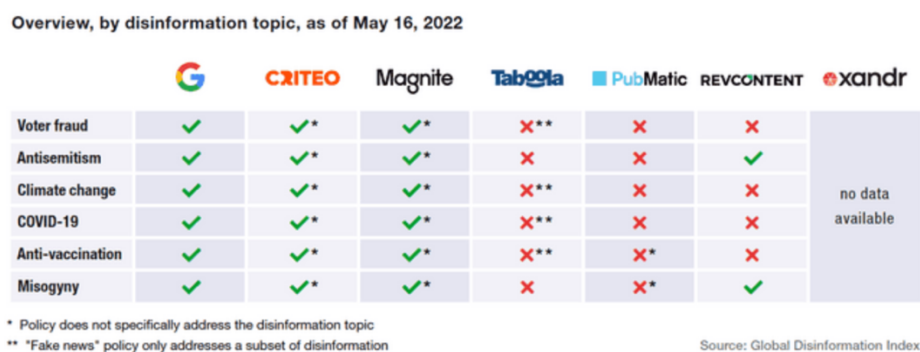


Figure 5.4 [Retrieved from: Ad Tech Policy and Enforcement Gaps: Challenges and Solutions. The Global Disinformation Index. 2022]

Especially relevant for TITAN are their guidelines for developers of the conversational AI from 2018 [62] however outdated these may seem, they did not update them with recent changes as the principles remain the same:

1. Explain the purpose of your bot and take special care if your bot will support consequential use cases.
2. Be transparent about the fact that you use bots as part of your product or service.
3. Ensure a seamless hand-off to a human where the human-bot exchange leads to interactions that exceed the bot's competence.
4. Design your bot so that it respects relevant cultural norms and guards against misuse.
5. Ensure your bot is reliable.
6. Ensure your bot treats people fairly.
7. Ensure your bot respects user privacy.
8. Ensure your bot handles data securely.
9. Ensure your bot is accessible.
10. Accept responsibility.

Furthermore, this year 2023 will bring numerous initiatives from companies, such as Microsoft (investor and partner in OpenAI and ChatGPT, who has also coined the toolkit 'Responsible AI Principles' [63]) by bringing ethical guidelines that guide their products, which are the following: Fairness, Reliability and Safety, Privacy and Security, Inclusiveness, Transparency, and Accountability, as well in the same report they propose guidelines to develop AI responsibly.

The OECD has published a Catalogue of Tools and Metrics for Trustworthy AI, particularly relevant to the AI Act: <https://oecd.ai/en/catalogue/tools>

Social networks and AI algorithms facilitate the interaction, creation, and categorization of content. In this respect, AI used in social networks influences how citizens consume content and news. However, these technologies; especially the facilitation and rapid spread of disinformation (e.g., clickbait and fake news), also have negative consequences, which TITAN seeks to mitigate.

The EU is taking initiative and giving importance to these new issues brought (or increased) by the use of AI in digital platforms. The EU pays particular attention to the human-centric aspect of AI development, increasing focus on standardization levels, the aim is to develop responsible AI systems and creates policies for good practices and standards that ensure an ethical and legal framework for AI systems. TITAN must account for these standards, which are founded in ethical principles, to create a tool that are trustworthy and free from biases and reduces risks.

We start from the AI High-Level Expert Group (AI HLEG)'s Ethical Guidelines that set the standards for trustworthy AI systems to support TITAN's development, with the seven principles: (1) Human agency and oversight (2) Technical Robustness and safety (3) Privacy and data governance (4) Transparency (5) Diversity, non-discrimination, and fairness (6) Societal and environmental well-being (7) Accountability, We also specifically account for EU new legislation that should be seen in combination with other fields, such as the GDPR, and are especially relevant to TITAN: the DSA and the proposed AI Act.

5.4 NEXT STEPS: THE PROCESS FOR OBTAINING THE ETHICAL AND LEGAL REQUIREMENTS OF TITAN

This section (5) sets the basis for **D1.3** Template and Guidance for the ethical Impact Assessment (Self-Assessment Template). This section analyses the legal and ethical dimensions of trustworthy AI to fight disinformation, following the standards of (new) regulations that account for Impact Assessment of technologies, especially Artificial Intelligence. These regulations apply to TITAN to assess the trustworthiness

of the system: **both GDPR (with its Data Protection Impact Assessment) and the AI Act (with the trustworthy AI assessment)**. We consider that there are two different impact assessments that aim at balancing ethical principles of trustworthy AI and misuse of data. The data protection and trustworthy AI impact assessments are complementary, trustworthy AI principles also account for the importance of privacy and data protection principles, and we consider both necessary to prevent the negative effects (risks) to the rights of individuals. Given the sensitive nature of TITAN to provide guidance on thinking, whereby personal information can be of sensitive nature, we will conduct both assessments, which will guide TITAN’s legal and ethical development.

5.5 REFERENCES

- [1] <https://plus.google.com/+UNESCO>, “Journalism, ‘Fake News’ and Disinformation: A Handbook for Journalism Education and Training,” *UNESCO*, Sep. 03, 2018. <https://en.unesco.org/fightfakenews> (accessed Mar. 10, 2023).
- [2] “Final report of the High Level Expert Group on Fake News and Online Disinformation | Shaping Europe’s digital future.” <https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation> (accessed Mar. 24, 2023).
- [3] “Automated tackling of disinformation-Major challenges ahead | Panel for the Future of Science and Technology (STOA) | European Parliament.” [https://www.europarl.europa.eu/stoa/en/document/EPRS_STU\(2019\)624278](https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624278) (accessed Mar. 24, 2023).
- [4] “Regulating disinformation with artificial intelligence | Panel for the Future of Science and Technology (STOA) | European Parliament.” [https://www.europarl.europa.eu/stoa/en/document/EPRS_STU\(2019\)624279](https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624279) (accessed Mar. 24, 2023).
- [5] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018, doi: 10.1126/science.aap9559.
- [6] V. Bakir and A. McStay, *Optimising Emotions, Incubating Falsehoods: How to Protect the Global Civic Body from Disinformation and Misinformation*. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-031-13551-4.
- [7] J. L. Juul and J. Ugander, “Comparing information diffusion mechanisms by matching on cascade size,” *Proc. Natl. Acad. Sci.*, vol. 118, no. 46, p. e2100786118, Nov. 2021, doi: 10.1073/pnas.2100786118.
- [8] “Zimmermann and Kohring - 2020 - Mistrust, Disinforming News, and Vote Choice A Pa.pdf.”
- [9] E. C. Tandoc, Z. W. Lim, and R. Ling, “Defining ‘Fake News’: A typology of scholarly definitions,” *Digit. Journal.*, vol. 6, no. 2, pp. 137–153, Feb. 2018, doi: 10.1080/21670811.2017.1360143.
- [10] N. Van Raemdonck and T. Meyer, “Why Disinformation is Here to Stay. A Socio-technical Analysis of Disinformation as a Hybrid Threat,” in *Addressing Hybrid Threats: European Law and Policies*, L. Lonardo, Ed., Edward Elgar, 2022.
- [11] W. L. Bennett and S. Livingston, “The disinformation order: Disruptive communication and the decline of democratic institutions,” *Eur. J. Commun.*, vol. 33, no. 2, pp. 122–139, Apr. 2018, doi: 10.1177/0267323118760317.
- [12] F. Zimmermann and M. Kohring, “Mistrust, Disinforming News, and Vote Choice: A Panel Survey on the Origins and Consequences of Believing Disinformation in the 2017 German Parliamentary Election,” *Polit. Commun.*, vol. 37, no. 2, pp. 215–237, Mar. 2020, doi: 10.1080/10584609.2019.1686095.
- [13] M. Himelein-Wachowiak *et al.*, “Bots and Misinformation Spread on Social Media: Implications for COVID-19,” *J. Med. Internet Res.*, vol. 23, no. 5, p. e26933, May 2021, doi: 10.2196/26933.
- [14] A. Hern, “AI bot ChatGPT stuns academics with essay-writing skills and usability,” *The Guardian*, Dec. 04, 2022. Accessed: Mar. 09, 2023. [Online]. Available: <https://www.theguardian.com/technology/2022/dec/04/ai-bot-chatgpt-stuns-academics-with-essay-writing-skills-and-usability>
- [15] “Pause Giant AI Experiments: An Open Letter - Future of Life Institute.” <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> (accessed Apr. 07, 2023).
- [16] N. Chomsky, I. Roberts, and J. Watumull, “Opinion | Noam Chomsky: The False Promise of ChatGPT,” *The New York Times*, Mar. 08, 2023. Accessed: Mar. 13, 2023. [Online]. Available: <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- [17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, in *FACCT ’21*. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

- [18] P. Marx, “Artificial intelligence’s dirty secret,” *Business Insider*. <https://www.businessinsider.com/chatgpt-ai-will-not-take-jobs-create-future-work-opportunities-2023-2> (accessed Mar. 14, 2023).
- [19] “Art and artificial intelligence collide in landmark legal dispute | Financial Times.” <https://www.ft.com/content/d691d599-3cdb-48d8-9824-9b2784a17d90> (accessed Mar. 09, 2023).
- [20] “Artificial intelligence is reaching behind newspaper paywalls,” *The Economist*. Accessed: Mar. 09, 2023. [Online]. Available: <https://www.economist.com/business/2023/03/02/artificial-intelligence-is-reaching-behind-newspaper-paywalls>
- [21] D. Susser, B. Roessler, and H. Nissenbaum, “Technology, autonomy, and manipulation,” *Internet Policy Rev.*, vol. 8, no. 2, Jun. 2019, doi: 10.14763/2019.2.1410.
- [22] J. L. Egelhofer and S. Lecheler, “Fake news as a two-dimensional phenomenon: a framework and research agenda,” *Ann. Int. Commun. Assoc.*, vol. 43, no. 2, pp. 97–116, Apr. 2019, doi: 10.1080/23808985.2019.1602782.
- [23] Y. Directorate-General for Education, *Guidelines for teachers and educators on tackling disinformation and promoting digital literacy through education and training*. LU: Publications Office of the European Union, 2022. Accessed: Jan. 30, 2023. [Online]. Available: <https://data.europa.eu/doi/10.2766/28248>
- [24] J. L. Egelhofer and S. Lecheler, “Fake news as a two-dimensional phenomenon: a framework and research agenda,” *Ann. Int. Commun. Assoc.*, vol. 43, no. 2, pp. 97–116, Apr. 2019, doi: 10.1080/23808985.2019.1602782.
- [25] “COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Artificial Intelligence for Europe (2018).” Accessed: Jan. 30, 2023. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN>
- [26] “What’s in a name?,” *CEPS*, Feb. 22, 2023. <https://www.ceps.eu/ceps-publications/whats-in-a-name/> (accessed Apr. 17, 2023).
- [27] V. Bakir and A. McStay, “Fake News and The Economy of Emotions,” *Digit. Journal.*, vol. 6, no. 2, pp. 154–175, Feb. 2018, doi: 10.1080/21670811.2017.1345645.
- [28] J. L. Nelson and H. Taneja, “The small, disloyal fake news audience: The role of audience availability in fake news consumption,” *New Media Soc.*, vol. 20, no. 10, pp. 3720–3737, Oct. 2018, doi: 10.1177/1461444818758715.
- [29] “Shaping Europe’s digital future.” https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/shaping-europes-digital-future_en (accessed Jan. 30, 2023).
- [30] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. 2021. Accessed: Mar. 01, 2022. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>
- [31] “The Digital Services Act: ensuring a safe and accountable online environment,” *European Commission - European Commission*. https://ec.europa.eu/info/digital-services-act-ensuring-safe-and-accountable-online-environment_en (accessed Mar. 01, 2022).
- [32] “High-level expert group on artificial intelligence | Shaping Europe’s digital future,” Jan. 27, 2023. <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai> (accessed Jan. 30, 2023).
- [33] “Ethics guidelines for trustworthy AI | Shaping Europe’s digital future,” Apr. 08, 2019. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (accessed Jan. 30, 2023).
- [34] “THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI) for self assessment.” Accessed: Jan. 30, 2023. [Online]. Available: https://www.google.com/url?q=https://ec.europa.eu/newsroom/dae/document.cfm?doc_id%3D68342&sa=D&source=docs&ust=1675097583629073&usg=AOvVaw30jmGA4s4P9QykFy_beYhS
- [35] “Home page - ALTAI.” <https://altai.insight-centre.org/> (accessed Jan. 30, 2023).
- [36] J. Fernow, “Ethical framework - SIENNA.” <https://www.sienna-project.eu/robotics/ethical-framework/> (accessed Jan. 30, 2023).
- [37] “Workbook – Project Sherpa.” <https://www.project-sherpa.eu/workbook/> (accessed Jan. 30, 2023).
- [38] “AI Impact Assessment – Project Sherpa,” Dec. 08, 2020. <https://www.project-sherpa.eu/ai-impact-assessment/> (accessed Jan. 30, 2023).
- [39] NATO, “NATO starts work on Artificial Intelligence certification standard,” *NATO*. https://www.nato.int/cps/en/natohq/news_211498.htm (accessed Mar. 15, 2023).
- [40] “Applied AI Ethics Typology,” *Google Docs*. https://docs.google.com/document/d/1h6nK9K7qspG74_HyVITOLx97URM0dRoGbJ3ivPxMhaE/edit?usp=embed_facebook (accessed Jan. 30, 2023).

- [41] “Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library.” <https://unesdoc.unesco.org/ark:/48223/pf0000381137> (accessed Mar. 14, 2023).
- [42] “EU Charter of Fundamental Rights.” https://commission.europa.eu/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en (accessed Mar. 14, 2023).
- [43] “ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf.” Accessed: Apr. 18, 2023. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf
- [44] “GDPR EUR-Lex - 32016R0679 - EN - EUR-Lex.” <https://eur-lex.europa.eu/eli/reg/2016/679/oj> (accessed Mar. 01, 2022).
- [45] “Guidelines 03/2022 on deceptive design patterns in social media platform interfaces: how to recognise and avoid them | European Data Protection Board.” https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en (accessed Mar. 15, 2023).
- [46] “ai_hleg_ethics_guidelines_for_trustworthy_ai-en_87F84A41-A6E8-F38C-BFF661481B40077B_60419 (5).pdf.”
- [47] “Communication on Fostering a European approach to Artificial Intelligence | Shaping Europe’s digital future,” Apr. 21, 2021. <https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence> (accessed Mar. 14, 2023).
- [48] “The challenges of regulating disinformation with artificial intelligence,” *Epthinktank*, May 03, 2019. <https://epthinktank.eu/2019/05/03/the-challenges-of-regulating-disinformation-with-artificial-intelligence/> (accessed Mar. 15, 2023).
- [49] J. Meszaros, J. Minari, and I. Huys, “The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union,” *Front. Genet.*, vol. 13, 2022, Accessed: Mar. 15, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fgene.2022.927721>
- [50] “A Europe fit for the digital age,” Feb. 19, 2020. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age_en (accessed Jan. 30, 2023).
- [51] “The Digital Services Act package | Shaping Europe’s digital future,” Mar. 10, 2023. <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package> (accessed Mar. 15, 2023).
- [52] “Buri-Van-Hoboken-DSA-discussion-paper-Version-28_10_21.pdf.” Accessed: Mar. 15, 2023. [Online]. Available: https://dsa-observatory.eu/wp-content/uploads/2021/11/Buri-Van-Hoboken-DSA-discussion-paper-Version-28_10_21.pdf
- [53] S. L. School, “No. 72: An Analysis of New Obligations for Online Content Sharing Platforms in the Media Environment under the DSA and DMA,” *Stanford Law School*. <https://law.stanford.edu/publications/no-72-an-analysis-of-new-obligations-for-online-content-sharing-platforms-in-the-media-environment-under-the-dsa-and-dma/> (accessed Mar. 15, 2023).
- [54] “Questions and Answers: Digital Services Act,” *European Commission - European Commission*. https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348 (accessed Apr. 20, 2023).
- [55] L. D. Meco, “Gendered Disinformation and the Undermining of Women’s Rights and Democracy Globally”, Accessed: Mar. 24, 2023. [Online]. Available: https://she-persisted.org/wp-content/uploads/2023/02/ShePersisted_MonetizingMisogyny.pdf
- [56] M. Bladini *et al.*, “Anti-Gender Hate Speech in Populist Right-Wing Social Media Communication,” *This Rep. Result Eur. Proj. Titled “Hate Speech Gend. Soc. Netw. Polit. Parties” GENHA Ref. 875388*, [Online]. Available: http://genha.eu/sites/default/files/pdf/Anti-Gender%20Hate%20Speech%20in%20Populist%20Right-Wing%20Social%20Media%20Communication_0.pdf
- [57] “The OECD Artificial Intelligence (AI) Principles.” <https://oecd.ai/en/ai-principles> (accessed Mar. 15, 2023).
- [58] “CLIMATE CHANGE AND AI Recommendations for Government Action.” Accessed: Apr. 20, 2023. [Online]. Available: <https://www.gpai.ai/projects/climate-change-and-ai.pdf>
- [59] “CAHAI - Ad hoc Committee on Artificial Intelligence - Artificial Intelligence - publi.coe.int,” *Artificial Intelligence*. <https://www.coe.int/en/web/artificial-intelligence/cahai> (accessed Mar. 15, 2023).
- [60] The Global Disinformation Index., “Disrupting Disinformation: A Global Snapshot of Government Initiatives.” <https://www.disinformationindex.org/> (accessed Mar. 24, 2023).
- [61] The Global Disinformation Index., “Ad Tech Policy and Enforcement Gaps: Challenges and Solutions.” <https://www.disinformationindex.org/research/2022-09-07-ad-tech-policy-and-enforcement-gaps-challenges-and-solutions/> (accessed Mar. 24, 2023).
- [62] “A guide to anti-misinformation actions around the world – Poynter.” <https://www.poynter.org/ifcn/anti-misinformation-actions/#uk> (accessed Mar. 24, 2023).

[63] “Responsible AI - Cloud Adoption Framework | Microsoft Learn.” <https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/strategy/responsible-ai?toc=https%3A%2F%2Fdocs.microsoft.com%2Fen-us%2Fazure%2Farchitecture%2Ftoc.json&bc=https%3A%2F%2Fdocs.microsoft.com%2Fen-us%2Fazure%2Farchitecture%2Fbread%2Ftoc.json> (accessed Mar. 24, 2023).

6 MAPPING EXISTING USER TACTICS TO COPE WITH DISINFORMATION

Understanding disinformation from the perspective of everyday news use [1] is key, both to advance our fundamental understanding of the issue and to devise solutions for it. The European Commission's Communication on Disinformation [2] focuses on strengthening media literacy amongst citizens and investing in fact-checking and in-depth journalism. While this approach is valuable and necessary to increase citizens' informational skills, it also poses challenges as to how citizens can incorporate these skills in their daily news routines. We can see citizens developing their own strategies to cope with disinformation in everyday news, like unfollowing people that spread unreliable information or turning to friends and family to discuss the news. A deeper understanding of citizens' coping practices can help policy makers and media educators to develop initiatives aimed at strengthening current coping practices and that will win citizens' acceptance. This section will further elaborate on the existing coping tactics adopted by citizens to tackle disinformation.

6.1 AUTHENTICATION

The first coping tactic people can adopt is based on the identification and recognition of fake news [3]. This process of authentication, or fact-checking, refers to the way in which people seek out input from other sources to further verify the authenticity of a (news) source [4]. As pointed out by Edgerly and colleagues [5] and York and colleagues [4] this audience authentication or verification is a key skill since it should safeguard people against disinformation and thus holds a normative promise. However, whether this promise is fulfilled depends on the response of audiences whenever they come across (potential) disinformation and fake news. This will be discussed further. Tandoc and colleagues [6] identify a two-step authentication process: internal and external.

6.1.1 Internal authentication

Internal authentication is the first step and is first and foremost based on someone's existing knowledge and their intuition or, as pointed out by Swart [7], people's gut feeling or common sense. The study of Kow and colleagues [8] states that people generally feel they have sufficient common sense and logic to decide whether a source or message is valid and thus the information is trustworthy. Similar reasonings were given by participants in the study of Wagner and Boczkowski [9].

The following element is that of text-based characteristics and the tone used in the message that will signal accuracy or inaccuracy [6][10]. This ranges from the inclusion of visuals and the use of scientific data to a consistency between the headline and content and a credible source. These observations also have been pointed out by Edgerly and colleagues [5], Malita and Grosbeck [11] and Lehaff [12] whose studies highlight how source credibility directly impacts the verification process. Interestingly, research by Tsang [13] reveals that the source credibility has little influence on whether a news message is perceived to be fake. For example, even messages without a source could be perceived as legitimate. In general, the idea of source credibility is relative to the individual's interpretation of the source and text. Someone might consider a problematic source to be credible, and vice versa [14].

The last part of internal authentication refers to people using quantitative elements such as the amount of likes and comments as guidance to determine the trustworthiness of a news message. This is a problematic tactic, since quantitative elements reflect the popularity of a message and not accuracy or trustworthiness [6]. This was also voiced by respondents in Lehaff's study [12, p. 943] who expressed "strong reservations and downright derision of the value of comment sections". Overall, internal authentication is a passive process since it heavily relies on what the person already knows and their own point of view, therefore not requiring that much effort. Moreover, people are required to have more knowledge to fully determine the accuracy of information [13].

6.1.2 External authentication

Whenever doubts arise during the process of internal authentication, the individual might take it a step further and go over to the second step of external authentication. In this scenario, they will seek out support from other people (interpersonal authentication strategies) or institutions (institutional authentication strategies) to further authenticate the news message. This cross-checking with external sources is either incidental and passive or intentional and active [6].

In the case of interpersonal authentication strategies, someone will verify a message with their own network. They can approach this in an incidental manner by simply waiting for confirmation by their friends who might post about a news event on their social media. At the same time, they can also seek out confirmation and intentionally participate in discussions about the news message. This can either be in real life with friends and family or through social media (e.g., being part of a group chat) [6]. It is clear that people put a lot of trust in their own social network. For example, they will assume that others in their environment, whether that is in real life or digital communities, are trustworthy and able to differentiate between what is true and false information. Thus, there is the underlying belief that information coming out of their social circles is accurate by default and therefore, they do not feel the need to (constantly) verify [10].

Finally, people are skeptical towards those who have conflicting worldviews and almost expect them to spread inaccurate information. Whilst the political orientation or familiarity to a person is at the bottom of the evaluation tactics, it is still used to credit or discredit information regardless of other textual clues [10]. The idea of depending on people that have been labelled as ‘credible’ and avoiding those who might be ideologically different is also voiced by participants in the study of Wenzel [15]. Interestingly, research of Wagner and Boczkowski [17] shows that some participants do value hearing both sides of the political aisle and will seek out different types of news media. However, this is not the case for everyone and quickly becomes tiring [15].

With institutional strategies, people will visit other news media or use the internet to search for further information [3][6]. Once again, they can adopt a passive approach. For example, by waiting for governments or other media to confirm a news event. Interestingly, if more than one source talks about the event, people see this as trustworthy by default [6]. On the other hand, people also can be more intentional and thus look for confirmation themselves via these sources. For example, research by Wenzel [15] shows how several participants simply use Google to verify information they found elsewhere. The search engine was also mentioned by participants in the study of Wagner and Boczkowski [16]. This is also voiced by the study of Fletcher and Nielsen [17] whose respondents state using multiple sources depending on their purpose. Interestingly, participants in Wenzel’s study [15] felt that the authentication process is their own responsibility and were not impressed by fact-checking initiatives offered by social media platforms and tools like Snopes.

Whilst the previously mentioned studies show how people adopt certain tactics to verify information, this is not a given. The abundance of information in the digital environment and the blurred lines between professional and non-professional sources make it increasingly challenging for people to verify news messages [6][7][18]. For example, the study of Swart [7] shows how these challenges result in cynicism and apathy with young people who will limit their efforts to verify information. Surprisingly, Edgerly and colleagues [5] found that respondents are more likely to start the authentication process when they already find the source credible. The passive approach to disinformation is also noticeable in terms of the potential correction tactics.

6.2 CORRECTION

6.2.1 Intrapersonal tactics

Once someone realizes they have been confronted with fake news, there are three possible tactics: intrapersonal, interpersonal and interventional strategies. First, someone might adopt intrapersonal strategies

or what is known as ‘passive disengagement’ [3][19]. This tactic, as the term suggests, does not require a lot of effort. Someone might choose to disregard the post altogether or step away from specific news topics and sources [3][20]. Thus, this raises the question of what kind of sources people are eliminating from their news repertoire and the consequences this has for their news repertoire.

Furthermore, people will ignore certain interpersonal sources like those who they do not agree with, share fake news or information that is seen as offensive to the reader. For example, unfollowing or blocking someone on social media [15]. In general, as pointed out by Wagner and Bockzkowski [16], people tend to be careful about who they follow and will use their inner circle as a filter. People who adopt these types of tactics can be classified as ‘evaders’ or ‘do-nothings’ and are the majority [3][20]. This is especially true in the case of social media platforms that allow users to curate their feed to show the content that is particularly of interest to them [21].

Last, some people will avoid news altogether as a tactic to cope with disinformation and fake news and are known as ‘news avoiders’ [19][22]. As established under ‘Zooming in on news avoidance and disengagement’, one of the drivers of news avoidance is the perception of being overloaded with news [23]. This is also pointed out by participants of Wenzel’s study [15] who express frustration because of the abundance of information. This goes back to news avoidance as a tactic that is used by individuals to protect themselves from negative reactions such as fear and distrust [12][23]. The sentiment of distrust is particularly interesting in relation to coping tactics and fake news. Participants in the study of Wenzel [15] referenced to fake news as one of the main causes for distrusting media, thus resulting in news avoidance and rearranging how they use media in general. Whilst distrust can result in people avoiding news, it also possibly leads people to non-mainstream media [24]. One of the participants in the study of Wenzel [15] specifically mentions the platform 4chan as his main source of information. Whilst the participant is aware of the problematic content on the platform, he still feels that 4chan is his only option to find news because of his distrust in mainstream media and platforms.

6.2.2 *Interpersonal and interventional tactics*

Then there is the ability to take on a more active engagement approach or adopt interpersonal and interventional coping tactics [3][19]. These tactics require a more active approach from the individual. In the case of interpersonal tactics, someone will warn their friends and family of false information. Thus, the individual sets out to protect their own network and make sure they are not affected. Interventional tactics go a step further and indicate that someone will actively intervene with the false information. For example, by leaving comments that say the news is fake, posting corrections themselves or asking the source to delete it [3]. These individuals can be identified as ‘the righteous’ and ‘clan protectors’ and are the minority [3] [20]. However, as established, this is not a given. Before someone adopts either interpersonal or interventional tactics, they must identify the news message as harmful. According to Chang [3], this is based on the prevalence and severity of the false information as well as people’s own susceptibility to the false information. This perception influences what type of coping tactic is adopted.

Similarly, the study of Tandoc and colleagues [20] identifies three actors that determine whether people will act when they are confronted with fake news or will choose to ignore it: (1) relevance of the issue, (2) their relationship with the source and (3) the perceived personal efficacy. First, depending on how relevant the issue is to the individual determines whether they will react to fake news. Participants in this study made a distinction between minor and major fake news. Under minor fake news they classified ‘harmless posts’ that tackle topics such as celebrity news since they feel this is trivial. Major fake news, however, is interpreted as the type of news that has consequences for their own lives or that of their friends or family.

Second, participants were more likely to correct false information if it was posted or shared by a family member or friends. Because of the close relationship, they felt the need to protect others in their environment and felt that it was easier to correct them than a stranger on the internet since that might lead to an argument. Last, correcting fake news is not easy nor is it always successful. Participants of this study felt that it is difficult

to change someone's opinion and thus, do not feel that they would be successful in correcting the false information. Furthermore, some of them also believed that fake news is still an expression of someone's opinion and are therefore not always comfortable to correct since they do not want to impose their own opinion.

6.3 CONCLUSION

The aim of this section of the report is to gain insight in the existing tactics that users adopt to cope with false information. Whilst the available data at this point in time is still limited, existing research identifies two important approaches. The first main coping tactic is based on the idea of authentication or people's ability to identify and recognise fake news. Here, we elaborated on the processes of internal and external authentication. Internal authentication requires little effort of the individual since it is based on their own knowledge. Interestingly, different studies indicate that **people believe they have sufficient common sense to determine whether a message is trustworthy**. They achieve this by looking at the text-based characteristics and the tone that was used. Moreover, source credibility is oftentimes mentioned simultaneously with authentication coping tactics. Whilst at first glance it is a good thing people look at the source of a news message, **not everyone will have the same interpretation of what is fake and what is not**.

While internal authentication is done individually, external authentication includes someone's personal network (e.g., family and friends) or institutions (e.g., other media and governmental information). Overall, it is clear that **the concept of familiarity plays a big part in authenticating information**, whether that is with their social network or sources. More often than not, the social network is referred to as one of the filters to authenticate information, with the underlying conviction that this information is accurate by default. In combination with the insight that research finds some news users explicitly question fact-checking initiatives, **the TITAN chat bot will need to consider how it will position itself in relation to users' friends and families**. Maybe a chat bot, more than a Google page or news medium, can enter that sphere of familiarity through its dialogic affordance. But even then, users may put forward incorrect but highly trusted information coming from their network in the dialogue, to which the chatbot will need to respond tactfully.

The second main coping tactic of correcting false information includes intrapersonal, interpersonal and interventional tactics. Intrapersonal tactics show methods such as **ignoring false information or unfollowing sources** with whom they do not agree. This part in particular highlights some of the **perverse effects of coping tactics** such as people avoiding news altogether and depending on alternative, biased news sources. In sum, whilst someone might identify fake news through authentication, this does not mean further correction. The majority of people will adopt a passive approach and deflect the false information one way or another. Finally, **actively correcting false information is not a given** and people experience multiple barriers in doing so.

Clearly, coping tactics such as diversifying news sources or proactively seeking extra confirmation of news events might strengthen citizens' engagement with the public sphere. At the same time, tactics such as news avoidance might weaken it. Therefore, the coping tactics people adopt might not always lead to the favourable, nor the expected outcomes.

6.4 REFERENCES

- [1] D. Boyd, "Did media literacy backfire?," *Journal of Applied Youth Studies*, vol. 1, no. 4, pp. 83–89, 2017.
- [2] European Commission, "Tackling online disinformation: a European approach," COM(2018) 236 final, Apr. 2018.
- [3] C. Chang, "Fake news: audience perceptions and concerted coping strategies," *Digital Journalism*, vol. 9, no. 5, pp. 636–659, May 2021, doi: 10.1080/21670811.2021.1923403.
- [4] C. York, J. D. Ponder, Z. Humphries, C. E. Goodall, M. A. Beam, and C. Winters, "Effects of Fact-Checking Political Misinformation on Perceptual Accuracy and Epistemic Political Efficacy," *Journalism & Mass Communication Quarterly*, vol. 97, no. 4, pp. 958–980, 2020, doi: 10.1177/1077699019890119.

- [5] S. Edgerly, R. R. Mourão, E. Thorson, and S. M. Tham, "When Do Audiences Verify? How Perceptions About Message and Source Influence Audience Verification of News Headlines," *Journalism & Mass Communication Quarterly*, vol. 97, no. 1, pp. 52–71, 2020, doi: 10.1177/1077699019864680.
- [6] E. C. Tandoc, R. Ling, O. Westlund, A. Duffy, D. Goh, and L. Z. Wei, "Audiences' acts of authentication in the age of fake news: A conceptual framework," *New Media & Society*, vol. 20, no. 8, pp. 2745–2763, 2018, doi: 10.1177/1461444817731756.
- [7] J. Swart, "Tactics of news literacy: How young people access, evaluate, and engage with news on social media," *New Media & Society*, vol. 25, no. 3, pp. 505–521, 2023, doi: 10.1177/14614448211011447.
- [8] Y. M. Kow, Y. Kou, X. Zhu, and W. H. Sy, "'Just My Intuition': Awareness of Versus Acting on Political News Misinformation," in *Information in Contemporary Society, IConference 2019.*, N. Taylor, C. Christian-Lamb, M. Martin, and B. Nardi, Eds. Springer, 2019, pp. 469–480. doi: 10.1007/978-3-030-15742-5_45.
- [9] M. C. Wagner and P. J. Boczkowski, "The Reception of Fake News: The Interpretations and Practices That Shape the Consumption of Perceived Misinformation," *Digital Journalism*, vol. 7, no. 7, pp. 870–885, 2019, doi: 10.1080/21670811.2019.1653208.
- [10] S. Koçer, B. Öz, G. Okçuoğlu, and F. Tapramaz, "Folk theories of false information: A mixed-methods study in the context of Covid-19 in Turkey," *New Media & Society*, pp. 1–21, 2022, doi: 10.1177/14614448221142310.
- [11] L. Malita and G. Grosseck, *Tackling Fake News in a Digital Literacy Curriculum*, vol. 4. "Carol I" National Defence University, 2018. doi: 10.12753/2066-026X-18-000.
- [12] J. Lehaff, "When News Use Feels Wrong: Four Reactions to Misalignments Between Feeling Rules and Feeling Responses," *Journalism Studies*, vol. 23, no. 8, pp. 932–950, 2022, doi: 10.1080/1461670X.2022.2059695.
- [13] S. J. Tsang, "Motivated fake news perception: The impact of news sources and policy support on audiences' assessment of news fakeness," *Journalism & Mass Communication Quarterly*, vol. 98, no. 4, pp. 1059–1077, 2021, doi: 10.1177/1077699020952129.
- [14] E. C. Tandoc, Z. W. Lim, and R. Ling, "Defining 'Fake news,'" *Digital Journalism*, vol. 6, no. 2, pp. 137–153, 2017, doi: 10.1080/21670811.2017.1360143.
- [15] A. Wenzel, "To Verify or to Disengage: Coping with 'Fake News' and Ambiguity," *International Journal of Communication*, vol. 13, pp. 1977–1995, 2019, [Online]. Available: <http://ijoc.org/index.php/ijoc/article/view/10025/2636>
- [16] M. C. Wagner and P. J. Boczkowski, "The Reception of Fake News: The Interpretations and Practices That Shape the Consumption of Perceived Misinformation," *Digital Journalism*, vol. 7, no. 7, pp. 870–885, 2019, doi: 10.1080/21670811.2019.1653208.
- [17] R. Fletcher and R. K. Nielsen, "Generalised scepticism: how people navigate news on social media," *Information, Communication & Society*, vol. 22, no. 12, pp. 1751–1769, 2019, doi: 10.1080/1369118x.2018.1450887.
- [18] M. Chan, "News literacy, fake news recognition, and authentication behaviors after exposure to fake news on social media," *New Media & Society*, p. 146144482211276, Oct. 2022, doi: 10.1177/14614448221127675.
- [19] M. F. Scheier, C. S. Carver, and J. K. Weintraub, "Assessing coping strategies: A theoretically based approach.," *Journal of Personality and Social Psychology*, vol. 56, no. 2, pp. 267–283, 1989, doi: 10.1037/0022-3514.56.2.267.
- [20] E. C. Tandoc, D. J. Lim, and R. Ling, "Diffusion of disinformation: How social media users respond to fake news and why," *Journalism: Theory, Practice & Criticism*, vol. 21, no. 3, pp. 381–398, 2020, doi: 10.1177/1464884919868325.
- [21] B. Ytre-Arne and H. Moe, "Approximately Informed, Occasionally Monitorial? Reconsidering Normative Citizen Ideals," *The International Journal of Press/Politics*, vol. 23, no. 2, pp. 227–246, Apr. 2018, doi: 10.1177/1940161218771903.
- [22] S. Edgerly, "Red media, blue media, and purple media: news repertoires in the colorful media landscape," *Journal of Broadcasting & Electronic Media*, vol. 59, no. 1, pp. 1–21, 2015, doi: 10.1080/08838151.2014.998220.
- [23] M. Villi et al., "Taking a Break from News: A Five-nation Study of News Avoidance in the Digital Era," *Digital Journalism*, vol. 10, no. 1, pp. 148–164, 2022, doi: 10.1080/21670811.2021.1904266.
- [24] Y. Tsfati, "Online News Exposure and Trust in the Mainstream Media: Exploring Possible Associations," *American Behavioral Scientist*, vol. 54, no. 1, pp. 22–42, 2010, doi: 10.1177/0002764210376309.

7 CONCLUSION PART 1: A SET OF REQUIREMENTS FOR TITAN'S SOCIO-TECHNICAL FRAMEWORK

This deliverable aimed at building the socio-technical groundwork of TITAN, with the intent to create a trustworthy, ethical and legal conversational agent to fight disinformation. TITAN brings the user to the center, considering the issues of individuals and society related to news consumption – by arguing that fact-checking should be supplemented with reason-checking: individually evaluating whether the argument exposed in the news is acceptable, relevant, or sufficient.

The deliverable is divided into five parts:

- 1) Systematic literature review of socio-political factors affecting citizen trust and engagement
- 2) Empirical Investigations of psychological/motivational/cognitive mechanisms
- 3) AI technology trends for models, data and resources to fight disinformation
- 4) Legal and ethical dimensions of trustworthy AI to fight disinformation
- 5) Mapping existing user tactics to cope with disinformation

The study of socio-political factors to improve users' critical thinking (2) describes the importance of having *trust* in media as well as critical skills to assess that information. This trust is built on professionalism and ethics, yet conflicts with critical thinking's questioning nature. The relationship between trust and distrust is intricate; excessive trust can lead to vulnerability, while balanced skepticism is crucial. Promoting critical thinking can inadvertently foster reliance on biases, risking misinformation. The key insight is that **both conventional and alternative thinkers perceive their reasoning as a form of critical thinking, even though their ways of accepting information differ**. Projects like Titan must navigate this challenge by addressing these biases. Additionally, news consumption often involves emotions, influencing trust. Emotional connections can lead to trusting untruthful sources, seen in alternative media's popularity. Recognizing emotions' role is vital for platforms like Titan that encourage critical thinking within the emotionally charged digital realm.

Looking at the role of socio-demographic factors, we see that they do matter in predicting news and information consumption, with a strong effect on an individual's **socioeconomic position**. Intuitively, one would expect these same socio-demographic factors to also matter in predicting media trust, but empirically we see that the significant effect of these socio-demographic factors disappears almost completely when we also control for political trust and political orientation. In other words, **TITAN should primarily consider socio-political factors (political trust and orientation) to understand media trust, while socio-demographics play much less of a role**. In addition, it is also recommended that Titan factors in that people from different political backgrounds might engage quite differently with the technology.

Building on part 1, section 2, **the study of psychological factors (3)** argues for the necessity to consider social cognitive theory, specifically Reciprocal Triadic Determinism (RTD) in the TITAN ecosystem. This theory puts forward different factors related to the spread of disinformation and individuals that should be considered: (1) Personal factors such as user's capabilities and biases (2) Situational and contextual factors, that are determined by the digital ecosystems such as affordances in social networks that amplify or reduce certain content (3) Behavioural factors, which refer to the intention of the individual to interact with disinformation news. In addition, we introduce the California Critical Thinking Test (CCTS) as a way to assess cognitive capacities of individuals, which focuses on six main cognitive skills that define critical thinking: interpretation, analysis, evaluation, inference, explanation and self-regulation. In addition, scientific studies are also introduced in favor of the AI Socratic method (AI frame questioning) and collective argument mapping to enhance group's ability to be more critical and minimize disinformation.

The study of the AI trends for models, data and resources to fight disinformation (4) gives an overview of the disinformation phenomenon, particularly of structural characteristics of fake news. It shows how disinformation can refer to different features (e.g., language styles that are too emotional or extreme, can

also contain visual-based features, or even can contain user-based features that resemble the users). In addition, this section highlights the difficulty of locating fake news, as it is a phenomenon constantly evolving, which finds new ways to be less predictive; for example, by deducing false inferences (fake) from facts (truth), such as generalizing facts that are not accurate, also named “defective reasoning”. The study supports one of TITAN’s premises that considers that fact-checking (e.g., monitoring by analytic websites that find patterns of disinformation) is not sufficient to combat disinformation and that it should be supported by reasoning. TITAN will thus focus on AI-driven Socratic dialogues to improve citizens’ critical thinking. Complementary to part 1, section 3, this section shows that we must take into account that there are different individuals with different critical thinking skills. It argues how the *AI formal argumentation theory structures* will guide the generation of Socratic dialogues, by evaluating citizens’ responses. It also shows the self-searching steps that individuals can do to fight disinformation: learning more about the source, check if an image is used with the correct content, getting the story behind provocative headlines, checking the author etc., in addition to the main characteristics to fact-checking sites.

The legal and ethical dimensions of trustworthy AI to fight disinformation (5) sets the basis for the ethical and legal approach of TITAN. This will lead the basis for the creation of a template on the Impact Assessment (D.1.3) & the self-assessment method (D.1.4) that will guide TITAN’s lifecycle. The section also accounts for EU new legislation that should be seen in combination with other fields, such as the GDPR and the proposed AI Act that are especially relevant to TITAN. We follow the High-Level Expert Group on AI (HLEG) of the European Union that has set standards for a trustworthy (and ethical) AI, respecting values and the regulatory landscape. For an AI to be trustworthy it is expected to be: ethical, legal and robust; that is to follow seven principles according to HLEG: (1) Human agency and oversight (2) Technical Robustness and safety (3) Privacy and data governance (4) Transparency (5) Diversity, non-discrimination, and fairness (6) Societal and environmental well-being (7) Accountability.

Finally, **mapping the existing user tactics to cope with disinformation (6)** details the two main steps citizens possibly go through whenever they are confronted with disinformation and fake news. First, they will have to authenticate the news message. Key insights show that **citizens prefer to consult their inner circle to authenticate information**, as well as search engines such as Google and other media. It is crucial for projects such as Titan to take this into consideration to make sure citizens know of its existence in the first place and establish a trusting relationship. Second, citizens will find themselves at a crossroad where they either choose to correct the false information or let it pass by. Here it is important to emphasize that most people take on a **passive approach**, with a tendency to ignore disinformation. Thus, recognizing existing user tactics and attitudes is important for projects like Titan to possibly fill in the gaps, while at the same time being aware that this passive attitude might make the use of Titan in authenticating information an important hurdle that many citizens might not be willing to take.

PART 2: USER NEEDS ANALYSIS

In an era characterized by an overwhelming influx of information and the rapid spread of disinformation, the need for effective strategies to combat falsehoods has never been more pressing. Disinformation, intentionally misleading or false information, poses a significant threat to individuals, communities, and societies at large. It erodes trust, distorts public discourse, and undermines the pursuit of truth. Addressing this challenge requires innovative approaches that empower individuals with the knowledge and tools necessary to discern fact from fiction.

One promising solution lies in the realm of conversational agents or chatbots. These artificial intelligence-powered systems have gained immense popularity and have become increasingly sophisticated in recent years. By leveraging natural language processing and machine learning techniques, chatbots can engage in interactive conversations, providing tailored information and guidance to users. In the context of combating disinformation, chatbots have the potential to play a crucial role in equipping users with the skills and resources to identify, verify, and counteract false information.

However, developing an effective Intelligent Coaching Environment that guides users in the fight against disinformation requires a comprehensive understanding of user requirements. User requirements analysis is a crucial initial step in system development, aimed at identifying the needs, expectations, and limitations of the target user base. By conducting a thorough investigation into the user landscape, including their knowledge gaps, information-seeking behaviors, and challenges encountered when encountering disinformation, researchers and developers can design chatbot systems that are tailored to the users' specific needs.

In TITAN the aim is to address the pivotal aspect of user requirements analysis in the development of a chatbot that will help the users to increase their critical thinking level to combat disinformation. By examining existing studies, surveys, and user feedback, we seek to identify key user requirements that such a system should fulfill. We will explore the desired functionalities, features, and characteristics of a chatbot that can effectively guide users in navigating the complex world of disinformation. Moreover, we will investigate the potential barriers and challenges users face when seeking information, enabling us to develop strategies to overcome these obstacles and enhance the chatbot's usability and user experience.

Ultimately, this research aims to contribute to the development of an intelligent chatbot that empowers users with the knowledge and tools needed to critically evaluate information, discern credible sources, and effectively combat disinformation. By focusing on user requirements analysis, we strive to create a system that aligns with users' expectations, integrates seamlessly into their information-seeking processes, and fosters a more informed and resilient society.

In the second part of this report, we analyze the methodology that we followed to collect specific user needs from the several available sources we had in the quiver. Then for each of these sources we thoroughly discuss how we end up to specific user needs. Before we delve deeper into user needs collection, we briefly discuss the innovative two step approach that we will follow in TITAN. Creating a chatbot to fight disinformation is a noble and challenging endeavor. Disinformation is a pervasive problem and addressing it effectively through a chatbot requires careful planning and consideration of various challenges like the collection of data. Ensuring the quality and accuracy of the training data is vital to the chatbot's performance and thus, labeling data correctly and dealing with biased or incomplete data is a challenging task. In part 2, section 8 below we describe these two steps in our approach.

8 TITAN COACHING SYSTEM: A TWO STEP APPROACH

The proliferation of disinformation has become a significant challenge in today's digital age. As disinformation spreads rapidly across various online platforms, combating its effects and ensuring the dissemination of accurate information has become an urgent priority. In response to this pressing issue, TITAN aims to deliver a chatbot specifically designed to counter disinformation by increasing the critical thinking level of the users. We have devised a two-step approach to tackle this complex problem, which involves the creation of a rule-based chatbot followed by the integration of a large language model. This section provides a comprehensive explanation of why this approach was chosen and how it is expected to enhance the effectiveness of our chatbot solution.

8.1 STEP 1: RULE-BASED CHATBOT DEVELOPMENT

To lay a solid foundation for our disinformation-fighting chatbot, we have opted for a rule-based system as the initial step. This decision has led to the analysis of 10 articles (see Annex 3) provided by the pilots, and by carefully examining these articles, our team has identified patterns, logical fallacies, and other indicators of disinformation. This knowledge forms the foundation of our rule-based system, enabling the chatbot to detect and respond to disinformation effectively. To maximize the impact of the rule-based system, we have developed Socratic dialogues based on the identified disinformation signals. These dialogues serve a dual purpose. First, they educate users on recognizing and understanding disinformation techniques, thereby increasing their critical thinking skills. Second, they provide valuable input for the rule-based system, enabling the chatbot to generate relevant responses and counteract disinformation effectively. By leveraging the power of dialogue, we aim to empower users with the tools needed to evaluate and debunk disinformation.

The advantages of a rule-based chatbot are manifold. Firstly, it allows for explicit control over the bot's responses, ensuring that disinformation is promptly identified and corrected. Through a carefully curated set of rules, our chatbot will be equipped with the ability to recognize and address various disinformation techniques, such as false statistics, misleading claims, or cherry-picked evidence. Additionally, the rule-based approach enables a high level of transparency, as the rationale behind the bot's responses can be clearly explained and justified. Moreover, by involving users in the utilization of the rule-based chatbot, we can gather valuable data on disinformation patterns and user interactions. These data, collected during real-world usage, will serve as a valuable resource for training and improving our chatbot's performance in the second step.

With the aim of tailoring the rule-based chatbot, building on the theoretical scaffolding identified we are structuring a prototype of a new procedure focused specifically on assessing critical thinking in the context of online disinformation, focused on commonly encountered Internet hoaxes. This procedure operates across the five dimensions of critical thinking outlined in the theoretical groundwork (see Section 3.2.1). This multidimensional approach will enrich our rule-based system, offering a granular level of assessment that enables the chatbot to assign tailored educational modules depending on the user's competency in each of the five critical thinking dimensions. By identifying the specific dimensions in which a user's critical thinking may be lacking, we will direct the chatbot to engage in targeted Socratic dialogues. This enhances not only the immediate user experience, but also serves our broader aim of enhancing the user's skills in the peculiar dimension of critical thinking where deficiencies have been detected.

Next, we give an overview of the methodologies supporting the first release of the TITAN AI assistant employing the Socratic method and the Microlessons and the respective methodologies we will follow towards the first platform release in M18.

8.1.1 The Socratic method

The Socratic method is an “inquiry” method, a dialogue between a teacher and students, aiming to explore the underlying beliefs that shape the students’ views and opinions. Political Science professor Rob Reich¹² suggests that “is better used to demonstrate complexity, difficulty, and uncertainty than to elicit facts about the world” and that “is characterized by ‘productive discomfort,’ not intimidation” and “the questioning proceeds open-ended with no pre-determined goal”. According to Reich, “the focus is not on the participants’ statements but on the value system that underpins their beliefs, actions, and decisions”, which in TITAN suggests that *the focus is on the critical thinking capacity*.

The Socratic method involves several types of questions, that aim at a different kind of understanding, and that are frequently organized as a series of steps¹³, as depicted in Figure 8.1¹⁴.

Question type	Examples
1. Clarification	What do you mean when you say X? How does this relate to Y? What is the main issue here? What do we already know about X? Could you explain that point further? Why are you stating that? Can you provide an example?
2. Challenging assumptions	What assumptions are we making here? How did you choose those assumptions? Why would someone assume that X? What could we assume instead? How can you verify or disprove? What would happen if...? Are you saying that...?
3. Evidence and reasoning	Can you provide an example that supports what you are saying? What led you to that belief? Upon what assumption is this belief based? What evidence is there that supports...? What evidence is there to support this argument? Can we validate that evidence? Are there any reasons to doubt the evidence? Do we have all the information we need? How might it be refuted?
4. Alternative viewpoints	Are there alternative viewpoints? What are alternative ways of looking at this? What are strengths/weaknesses of...?

¹² The Socratic Method: Fostering Critical Thinking: <https://tilt.colostate.edu/the-socratic-method/>.

¹⁵ Examples of Socratic Questioning: <https://positivepsychology.com/socratic-questioning/#examples>.

¹⁴ Figure from: <https://medium.com/@jdunns4/how-to-use-education-technology-with-the-socratic-method-af408b8de0a9>.

	What might someone who thought X think? How could someone else respond, and why?
5. Implications and consequences	How would this affect someone? What are the consequences of this assumption? How does ... affect ...? What are the long-term implications of this? How does ... fit with what our experience tells us?
6. Challenging the question	What do you think was important about that question? Why do you think I asked the question? What would have been a better question to ask?

Table 4.3 Socratic method steps and sample questions per step.

The aforementioned steps can be employed in cycles, usually until all parties accept the (restated) belief.



Figure 8.1: The Socratic Process^[1]

8.1.2 *Compilation of micro-lessons*

8.1.2.1 *What are micro-lessons?*

Micro-lessons are an educational concept: “Targeted lessons for students based on their knowledge of the subject matter rather than grade or age.”¹⁵ They tend to be short and focused¹⁶, usually about a single task or activity or about a simple and narrow topic¹⁷, under the assumption that they can be assimilated easily¹⁸, due to their shortness and focused nature.

From the DOA:

- AI will apply an experiential learning approach to personalize media literacy **micro-lessons** according to the citizen’s critical thinking capacity.
- Incorporation of **micro-lessons** that contain media literacy material on fact-checking processes and use of corresponding tools, including, among other signals, learning more about the source, whether various types of logical fallacies are present, checking if an image is used with the correct content, searching for ‘click-baits’, checking the author, seeing how (and if) different news agencies have reported on the same event, searching in real-world checks from trusted sources around the world, verifying location of event by using public tools such as Google Maps, Google Earth, or Google Street View, among others.
- Only when the critical thinking level of the citizen has been evaluated, appropriate Socratic dialogues embedding suitable **micro-lessons** can be compiled and delivered.
- Acting more as a facilitator than a tutor, the focus is mainly on the diversity and plethora of signals, enlarging the palette of “**lessons**” available for user engagement and signals that can be used as examples and triggers for dialogues.
- For this, the TITAN Socratic dialogues will also embed **micro-lessons** which incorporate the use of suitable fact-checking and media literacy tools and services according to the assessment of the citizen’s capacity in thinking critically. Due to the fact that each citizen has a different style to approach knowledge, Adaptive/Personalized Learning will be taken into account in the design and implementation of TITAN’s **micro-lessons**.
- TITAN will employ a methodology that combines well-known and widely applied learning theories, including the Constructivism and Experiential Learning theories along with the Socratic method in order to compile effective **micro-lessons**. Framed by theoretical models from learning, cognitive and behavioral sciences, the methodology will employ various learning paradigms, including learning by example, and inquiry-based learning following the Socratic questioning.

Summarizing the aforementioned excerpts, micro-lessons have the following properties:

1. They are personalized, according to citizen’s critical thinking capacity.
2. They contain media **literacy material**, and disinformation coping **tools**.
3. They are embedded in Socratic dialogues.
4. They employ well known learning theories (Constructivism, Experiential Learning, Socratic method, learning by example, inquiry-based learning) from learning, cognitive and behavioral sciences.

¹⁵ Micro Lessons: What They Are and How They Can Combat Learning Loss: <https://www.techlearning.com/news/micro-lessons-what-they-are-and-how-they-can-combat-learning-loss>.

¹⁶ Microlearning: A Complete Guide: <https://www.educationcorner.com/microlearning-guide.html>.

¹⁷ Everything you wanted to know about microlearning (but were afraid to ask): <https://www.efrontlearning.com/blog/2015/09/everything-you-wanted-to-know-about-micro-learning-but-were-afraid-to-ask.html>.

¹⁸ What Is Microlearning? The Case For Shorter, Bite Sized Learning: <https://www.thinkific.com/blog/what-is-microlearning/>.

8.1.2.2 Contents of micro-lessons

Since micro-lessons contain both literacy material and tools, a micro-lesson can include¹⁹:

- A video, or a tutorial
- A visual presentation, such as an infographic
- An activity, a small task to be completed
- A reflection prompt, a small dialogue
- A game (or a gamified activity), a quiz, a puzzle
- A discussion
- Any type of learning object

From [here](#)²⁰: “Learners often stay engaged longer when the learning activities are **highly visual** and vary in their **cognitive demand**.”²¹

8.1.2.3 Compilation of microlessons for AI coaching systems

The AI coacher has some requirements in order to be engaged with a citizen in a dialogue:

1. A form of “*trigger*”, that signals the initiation of a dialog session. This signal will come from the citizen, either because a) the citizen may want to ask the coacher a question, either a generic one, or about specific content (providing a URL, or pasting some text); or b) because the AI coacher has “flagged” some elements inside the citizen browser, and the user is curious to know more about the visual cues added by the coacher.
2. Some contextual information, like a) the critical thinking level assessment; b) contextual factors related to the trigger, if available.

For the time being, the browsing history is not required (or useful) to the AI coach (but it may be required for the critical thinking level assessment, this has to be determined).

8.1.2.3.1 Trigger A: The user asks a question

In this scenario, the user wants to pose a question. Yet, it is not known what the user may want to ask (this question should be asked during co-creation to the users). Possible questions may be:

- I want to know about a disinformation aspect (e.g., what is fake news, what is misinformation/hoax/conspiracy theory). A possible answer is to direct the user to literacy material.
- I want to know more about an issue, whether it is true or not. Reaction needs to be defined for this case.
- I want to know more about how to tackle disinformation (resources, tools, etc.).
- I am reading this content, and I want help. (This may be the case we can help more!) If the user provides a URL/some copy-pasted content, we can analyze through disinformation signals. Then the user query, disinformation signals, the content, the context, and the critical thinking level assessment can be used to compile micro-lessons (plan a dialogue).
- Other possible question types (from co-creation).

8.1.2.3.2 Trigger B: The user responds to visual stimuli added by the AI assistant

In this scenario, the AI coach may analyze the content currently visited by the citizen inside a browser. (Other content channels may be desirable, e.g., mobile phones, but feasibility must be assessed, as typical usage scenarios include applications, where integration may not be possible). Based on the critical thinking level

¹⁹ What Is Microlearning? The Case For Shorter, Bite Sized Learning: <https://www.thinkific.com/blog/what-is-microlearning/>.

²⁰ What Is Microlearning? The Case For Shorter, Bite Sized Learning: <https://www.thinkific.com/blog/what-is-microlearning/>.

²¹ Using Cognitive Load Theory To Reduce Learning Barriers For Students: <https://www.thinkific.com/blog/cognitive-load-theory-explained/>.

assessment, the content/context available by the browser (automatically, through a plugin) and the analysis results, the AI coach may add visual stimuli (highlighted text, badges, etc.) to the content the user sees. Then the user may decide to click on this visual stimulus to start a discussion.

8.1.2.3.3 Methodology for compiling micro-lessons

In an ideal for TITAN scenario, both triggers are very similar: both triggers provide very similar information to the AI coacher as its starting point. Available information includes the critical thinking level assessment, the content, the disinformation signals derived from the content, and possibly some contextual information. For trigger A we also have a question, which is the main difference between the two triggers, as it represents the aim of the citizen. In trigger B, where no question is available, the coach may decide the topic based on the confidence of the disinformation signal (select the one with the best confidence score), or add an additional utterance, asking for a “topic”.

Thus, the methodology for compiling a micro-lesson based on a trigger (A or B), is:

1. **Trigger:** The AI coach accepts a user trigger
 - a. Trigger A has a user query (question). The query is analyzed with respect to whether a dialogue should start, or if the citizen is directed to literacy material. If the query should start a dialogue, the AI coach should ask for content (URL, text to analyze) if not already provided.
 - b. Trigger B is missing a user query, but has content, and context, and the visual stimulus the user has clicked.
2. **Signals Extraction:** The AI coach analyses the content (and uses the context) to extract disinformation signals through signal-specific classifiers (e.g., hate speech, fallacies, etc.).
3. **Critical thinking Level Assessment:** The AI coach examines the current critical thinking assessment for the user. This assessment should provide information like a) the types of disinformation signals the user is aware of; b) the degree of the knowledge of the citizen for each signal; c) the degree of each signal being operational by the user (i.e., whether he/she applies this knowledge to the content at hand); d) the degree of “reasoning” about each signal; e) the ability to conclude about a signal. This methodology assumes that the disinformation signals in TITAN are representative of the disinformation types TITAN wants to address (and that raising the critical thinking level for these signals also raises the ability of the user to resist disinformation). Critical thinking level assessment may happen as a one-off test, taken by the user when starting to use the AI coach, which may be repeated to assess if the user’s critical thinking level has been modified.
4. **Disinformation Signals Assessment:** The AI coach assesses the disinformation signals along with the critical thinking level assessment. Trying to maintain an equilibrium between what the user does not know and the confidence in each signal, the coach ranks disinformation signals and selects the top 2 ranked as the topics for guiding the dialogue with the user. In case of trigger A, the 2 signals most related to the user question will be selected.
5. **Goal Planning:** The AI coach now has the discussion topics (2 disinformation signals), the segments of where each signal has been extracted (and can be used as an example for experiential learning). The output of this step should include: a) the goal, what to discuss with the user (e.g., if the topic is click-bait title, whether the goal is awareness, detection aspects, prioritization over other signals (importance), or examples on improving operation²²; b) the way, what is the best approach to discuss the topic/goal/detection aspects/experience. The way must take engagement incentives also into consideration. (A list of goal/communication way pairs is more probable as the output of this step).
6. **Dialog Implementation:** For each goal/way pair, the AI coach will employ Socratic method/experiential learning to dialog with the citizen. Compiling a pool of “initial” Socratic questions for each signal may be beneficial for starting this process. This step may be an iterative process, both within each goal/way pair (depending on the dialog with the user), and on discussing all goal/way pairs

²² These goals are loosely modelled after the four-component model of moral behaviour (Rest and Narvaez, 1995), interpreted as four “stages”: perception, judgement, prioritisation, action. Perhaps a better theory is required.

on the list of things to discuss (the input to the step). The user must be able to terminate interaction at any time.

Critical Thinking Level Assessment: We may want to assess the progress after the interaction has ended (or simply consider some things discussed, and simply append user history).

8.2 STEP 2: INTEGRATION OF A LARGE LANGUAGE MODEL

Building upon the rule-based system, our project's second step involves the integration of a large language model using the data collected during user interactions with the chatbot. The rule-based system will be deployed to users, allowing them to engage with the chatbot and receive accurate responses based on the initial set of rules and Socratic dialogues. This interaction data, including user queries and chatbot responses, will be stored and utilized for training a more advanced chatbot assisted by a large language model.

By training a language model with real-world data obtained from user interactions, we can enhance the chatbot's understanding, adaptability, and response generation capabilities. The large language model, built on a vast corpus of text, can leverage its contextual understanding to provide nuanced and informed responses to users' queries. This training process, fueled by user data, helps refine the chatbot's ability to identify patterns and contextual cues that might not be explicitly covered by the rule-based system. The large language model's training will empower it with a deeper understanding of the task at hand, enabling it to generate accurate and informative responses in real-time and cater to a wider range of user needs.

Employing an agile training process involving several training cycles will lead to an enhanced large language model through continuous learning and improvement. As the chatbot interacts with more users and encounters new types of dialogues, it can refine its responses, adapt to evolving disinformation coping tactics, and enhance its overall performance over time. This learning process ensures that our chatbot remains up-to-date and resilient against emerging disinformation threats.

In conclusion, our two-step approach to combating disinformation through the development of a rule-based chatbot followed by the integration of a large language model presents a robust and comprehensive solution. By leveraging the strengths of both systems, we can effectively detect, address, and prevent the spread of disinformation by increasing the critical thinking of the users. The rule-based chatbot provides a solid foundation for immediate response and transparency, while the integration of a large language model enhances the chatbot's ability to tackle complex and engaging dialogues with the users. This approach ensures that our chatbot solution remains adaptable, efficient, and continually evolves to stay ahead in the battle against disinformation.

9 USER NEEDS COLLECTION METHODOLOGY

9.1 APPROACH

To carry out the process of requirements identification and analysis for TITAN, a user-generated approach has been followed, with active participation and contribution from end users in the design process. To achieve this, users have been encouraged to provide input throughout the process in the form of needs, ideas, and feedback. Their perspectives, needs, and preferences are considered crucial inputs for creating TITAN as a user-centered solution. This approach is an ongoing process that takes various forms and has been applied to different stages of the design process. To facilitate the requirements identification and analysis for TITAN, a variety of tools were used in a complementary manner, including:

- **Literature Review:** We conducted an extensive review of existing literature, including scholarly articles, research papers, and studies that explored user behaviors, coping mechanisms, and challenges related to disinformation. This literature review provided valuable insights into user requirements, highlighting the factors that influence individuals' interactions with disinformation and their information-seeking behaviors.
- **User Stories and Pilot Users:** To gather user needs and understand their perspectives on combating disinformation, we collected user stories from pilot users. Through interviews, surveys, and feedback sessions, we obtained firsthand accounts of their experiences, challenges, and expectations when encountering and combating disinformation. These user stories played a vital role in identifying specific user requirements and shaping the overall design of the chatbot.
- **Co-Creation Workshops:** To foster active user participation and co-design, we organized co-creation workshops involving citizens from diverse backgrounds. These workshops served as collaborative platforms where participants, including digital literacy experts, journalists, and concerned citizens, shared their insights, ideas, and suggestions. By engaging in group discussions, brainstorming sessions, and interactive exercises, we gathered a range of perspectives, ensuring that the TITAN system reflected the needs and aspirations of its intended users.
- **Focus groups with partners:** After the collection of the requirements as described in the above steps, we organized focus groups with partners of the consortium and revised the requirements from a technical, business, psychological, etc.) point of view to incorporate domain expertise and ensure the relevance of the system, we consulted task owners and subject matter experts who possessed in-depth knowledge of disinformation, digital literacy, and information verification. Through expert interviews, discussions, and consultations, we gained valuable insights into the practical considerations, challenges, and emerging trends in combatting disinformation. Their expertise helped validate and refine the user requirements, ensuring that the TITAN system addressed the complexities of the disinformation landscape.
- **Legal and Ethical Considerations:** Recognizing the importance of legal and ethical requirements, we engaged legal experts to analyze and assess the legal implications and ethical considerations associated with the development of the TITAN system (see details in Section 5). By reviewing relevant regulations, guidelines, and ethical frameworks, we integrated legal and ethical requirements into the user requirements, ensuring compliance and safeguarding user privacy, security, and rights.
- **Iterative Design Process:** To create an intuitive and user-friendly chatbot, we followed an iterative design process. This involved developing mockups, and interactive demonstrations based on the identified user requirements. Through feedback collection by the partners, and continuous refinement, we iteratively improved the design, functionality, and user experience of the chatbot,

ensuring that it will effectively meet the identified user requirements and will be aligned with their expectations. In the living labs that will follow these mock-ups will be shared to the users for their feedback and further refinements will follow if needed.

A combination of the aforementioned methods was employed throughout the first period of the project to elicit user requirements from diverse groups with varying backgrounds and geographical locations. The implementation details and results of these efforts are described in the following sections. This methodology enabled us to capture the diverse needs, expectations, and challenges of users, resulting in a tailored and user-centric solution that empowers individuals to navigate the complex landscape of disinformation effectively.

9.2 EXPECTATIONS

A fundamental challenge for technology-oriented projects is to meet user expectations when developing applications and services. To stay competitive, a successful approach involves designing applications based on the latest state-of-the-art practices. This challenge necessitates a meticulous approach that revolves around designing applications that not only leverage cutting-edge technology but also align closely with and fulfill user needs. User requirements analysis forms the bedrock of this user-centered approach, enabling the creation of applications that seamlessly merge technical capabilities with user expectations.

User requirements analysis serves as the compass that provides precise descriptions of the content, functionality, and quality that prospective users demand. To identify these user needs, it is imperative to adopt a user perspective. This approach ensures that the development process remains tightly bound to user expectations, ultimately resulting in applications that strike the optimal balance between technical innovation and user satisfaction.

User requirements analysis provides precise descriptions of the content, functionality, and quality that prospective users demand. To identify user needs, it is essential to adopt a user perspective, resulting in the following:

- **Functional requirements** (What the users want the system to do):
 - This involves identifying desired user activities, constraints, and preferences, as well as understanding how users prioritize different software capabilities (e.g., speed of execution versus scope and complexity).
 - Goals that users aim to achieve and the tasks they intend to perform with the new software or system must be determined.
 - It is important to note that functional requirements specify "WHAT is wanted" rather than "HOW it will be delivered."
 - Additional aspects include user categorization, description of user characteristics (e.g., prior knowledge and experiences), consideration of special needs for each user category, subjective preferences, and description of the user environment where the product or service will be utilized.
- **Non-functional requirements** (Restrictions on the types of solutions meeting functional requirements):
 - Non-functional requirements focus on aspects such as performance, response time, fault recovery, scalability, security, and other criteria for evaluating system operation.
 - These requirements are also referred to as qualities of the system and can be categorized as follows:
 - Execution qualities, such as security and usability, which are observable at runtime.
 - Evolution qualities, such as testability, maintainability, extensibility, and scalability, are embodied in the system architecture.
 - Legal considerations, intellectual property rights, security, and privacy requirements are also important factors to address.

9.3 TARGET USER GROUPS

In the user requirements process of the project, we recognized the paramount importance of actively involving all potential user groups outlined in the project description. Our overarching goal was to gather a wide array of opinions, perspectives, and suggestions, fostering inclusivity and comprehensiveness in our approach to tackling disinformation. To achieve this ambitious objective, we embarked on an extensive journey of engagement with a diverse range of participants. To achieve this, three distinct target groups have been meticulously identified: students and journalists, NGOs and activists, as well as migrants and refugees. Furthermore, co-creation sessions were conducted with diverse citizen participants, as well as researchers and technology experts from the consortium who contributed their refined expertise to further expand the understanding of user needs. This holistic approach is essential for creating a truly user-centric solution that can effectively address the unique requirements and challenges faced by each group.

Target Group 1: Students and Journalists

Students: Higher education institutions play a pivotal role in preparing the citizens of the future who must actively participate in the fight against disinformation. These institutions bear the responsibility of nurturing critical thinking and fostering a 'fact-checking state of mind' among their students. These students represent a concerned group of citizens who are deeply involved in the ever-present information environment. Therefore, they should possess high levels of critical thinking and (news) information literacy. Their diverse perspectives are invaluable as they collectively shape the future landscape of information consumers and creators. By engaging with students, we gain insight into the evolving nature of disinformation, how it is encountered across different academic disciplines, and the developmental stages of critical thinking skills. These insights are pivotal in ensuring that the chatbot caters to a wide spectrum of users, from those embarking on their educational journeys to those already deeply immersed in academia.

Journalists: Journalists are often on the front lines of information dissemination, where the consequences of disinformation can be particularly damaging. Their insights into the challenges of verifying facts and navigating the complex web of disinformation are of utmost importance. By engaging with this group, we aim to understand the intricacies of their profession and the tools they require to combat disinformation effectively. This includes considerations of speed and accuracy in verifying information sources, as well as the ethical dimensions of fact-checking in real-time.

Target Group 2: NGOs and Activists

This group represents organizations and individuals passionately dedicated to specific causes, often in the face of concerted disinformation campaigns designed to undermine their efforts. Members of non-governmental organizations (NGOs) and activists are often highly engaged in exposing malpractices from both governments and organizations. Particularly in the current era of global disinformation and propaganda campaigns from authoritarian states, their role has become increasingly critical. However, to fulfill their mission effectively, they need to meticulously verify information to avoid inadvertently spreading disinformation themselves. Understanding their needs for robust verification and fact-checking tools, as well as their insights into the dynamics of disinformation as it pertains to advocacy, is pivotal. By collaborating with members of NGOs and activists, we seek to empower them with a tool that not only helps them counter false narratives but also amplifies their messages in the crowded digital landscape. Insights from this group will inform the development of features and functionalities tailored to the unique requirements of advocacy and activism in the age of digital disinformation.

Migrants and Refugees:

In an era marked by significant migrant and refugee flows, both internally within the European Union and externally due to economic instability and socio-economic factors in the greater region, addressing the information needs of these vulnerable groups is paramount. The public perception of the "refugee crisis" has been heavily influenced by media framing, which has shifted from tolerance to solidarity and, at times, to securitization and fear. Information manipulation campaigns have further exacerbated hostility against minorities, directly impacting the fundamental right to human dignity. As a result, they often find themselves in particularly vulnerable positions, susceptible to misinformation and disinformation that can significantly impact their lives. Their unique challenges include navigating unfamiliar legal systems, understanding their rights, and ensuring their safety. Engaging with migrants and refugees is critical to understanding their unique information needs and challenges. It enables us to design a tool that not only provides accurate information but also empowers these individuals to navigate the complex information landscape, make informed decisions, and safeguard their rights. Their diverse linguistic backgrounds, varying levels of digital literacy, and specific information needs necessitate a thoughtful and inclusive approach to user requirements collection.

Co-creation Sessions with Diverse Citizen Participants

The co-creation sessions will follow a human-centered approach and will actively involve a wide spectrum of citizens. This inclusivity extends to individuals from different age groups, cultural backgrounds, educational levels, and varying degrees of digital literacy. The overarching goal of involving these diverse representatives was not limited to mere tokenism or data collection; it was to genuinely listen, understand, and respect their opinions and requirements and tap into the collective wisdom and diverse experiences of the broader community. By tapping into their insights, we ensure that the chatbot tool is not only accessible but also user-friendly for individuals from all walks of life. It is imperative that the tool resonates with and serves the needs of the larger populace to maximize its impact in combating disinformation.

Contributions from Technology Experts

In addition to the invaluable input from the three primary target groups and the co-creation sessions, technology experts from the consortium will play a pivotal role in refining and expanding upon user requirements. These experts bring a wealth of technical knowledge and experience to the table, allowing them to bridge the gap between user needs and the technical capabilities of the chatbot. Their insights will be instrumental in ensuring the tool's feasibility, security, scalability, and overall effectiveness in combating disinformation. Their contributions are vital in translating user requirements into practical, innovative solutions that can navigate the complex digital landscape.

In conclusion, incorporating input from a diverse range of users and stakeholders, including students, journalists, NGOs, activists, migrants, refugees, citizens, and technology experts, is indispensable for the successful development of an online chatbot tool designed to counter disinformation. This holistic approach ensures that the final product is not only technologically sound but also deeply attuned to the unique requirements and challenges faced by each group. By considering a wide range of perspectives, the tool becomes a powerful instrument in the fight against disinformation and a catalyst for promoting critical thinking across diverse communities.

10 USER REQUIREMENTS RESULTING FROM USER TACTICS TO COPE WITH DISINFORMATION

In this section we provide the user requirements drawn from the literature review conducted in Section 6 of Part 1 on user tactics to cope with disinformation.

- The chatbot will need to be explicit about its premise i.e., an informed citizenship.

The chatbot should provide information about the concept of an informed populace, highlighting the importance of logical thinking, information-seeking, and political involvement among citizens in a democracy.

- The chatbot will need to be explicit about its premise i.e., the significance of critical thinking in media literacy education.

The chatbot should explain how questioning sources, considering credibility, and understanding own biases are essential aspects of evaluating information.

- The user may need to get a trust indicator for sources.

The chatbot should give an indication of trustworthiness of sources e.g., by stating the trust scores of brands in the Digital News Report.

- The chatbot will need to invite users to define their level of trust in provided sources.

When providing an indication of trustworthiness, the chatbot should invite the user to make their own assessment of trust in the source.

- The user may need to be challenged regarding the ways in which they assign trust to sources.

The chatbot should invite users to reflect on the assumptions underlying how they assign trust to sources, specifically helping them reflect on how emotional attachment, habits, and cognitive biases can overrule reliability and accuracy.

- The user may need to be made aware of the importance of emotions in how they deal with information.

The chatbot should highlight when the user is responding to emotions rather than ideas or facts. It should then prompt the user to reflect on how these non-rational factors may influence their trust in news and assessment of information. It should provide users with tools to distinguish between emotional appeals and evidence-based arguments and then with strategies to make more informed, balanced, and objective judgments.

- When presented with facts that are potentially discomforting or disturbing, the user may need to get a warning or accompanying text.

The chatbot will need to consider the impact of disturbing or discomforting factual information for the user e.g., confrontational facts about war, famine, global warming, injustice, etc. – ex ante.

- When presented with facts that are potentially discomforting or disturbing, the user may need to be invited to share their concerns or fears.

The chatbot will need to offer guidance through dialogue that can help the user cope with discomforting or disturbing facts – post hoc.

- The user may be experiencing how critical thinking about news sources leads to increased levels of distrust in those sources.

The chatbot should help users understand the balance between trust and skepticism in their engagement with information sources.

- The chatbot may need to inform the user about the potential weaponization of critical thinking.

When identifying an overtly critical or biased form of reasoning by the user, the chatbot should explain how critical thinking can be exploited by anti-democratic forces, by giving examples about the manipulation of information or the promotion of conspiracy theories. It should raise awareness of the dangers of falling into ideological echo chambers or being influenced by propaganda disguised as critical thinking.

- The chatbot should explore strategies or approaches to encourage critical thinking while maintaining trust in reliable news sources.

The chatbot should provide guidance on how citizens can navigate the complex interplay between trust and critical thinking in their news consumption habits. It should attend the user on how personal beliefs, political affiliations, religious beliefs, or following their gut feeling affects trust in media and scientific sources. It should prompt users to reflect on how this dynamic can inadvertently steer individuals towards untruthful sources aligned with their existing beliefs.

- The user may need to be attended to the role of biases and worldviews in critical thinking.

The chatbot will need to address how individuals' personal biases and worldviews influence their perception of what is right and wrong, true and false. It should highlight the possibility that critical thinking can be influenced more by personal ways of knowing rather than specific sources of information. When pointing out the biases in the user's reasoning, the chatbot should do so in a constructive and considerate manner. It can provide examples, such as alternative media, populist politicians, and fearmongers, who have mastered ways to gain people's trust despite disseminating inaccurate information.

- The chatbot should facilitate users' reflection on their own dispositions and biases.

The chatbot should prompt users to consider how their beliefs and preconceived notions may impact their evaluation of information. It should encourage users to be aware of their own cognitive biases and strive for balanced and objective assessments.

- When critical thinking is leading the user to dismiss factual information, the user may need to be encouraged to build its argument on factual sources.

The chatbot will need to be able to converse with a user that does not consider factual information that is offered to them. The conversation will need to invite the user to reflect on why they dismiss the facts – continue dialogue.

- When critical thinking continues to lead the user to dismiss factual information or to distrust the chatbot, the chatbot should explain that critical thinking ends where distrust begins.

The chatbot will need to confront the user with the conundrum of trust and critical thinking, explaining that a level of skepticism is encouraged but too much hinders Socratic/democratic dialogue.

- When critical thinking is consistently leading the user to dismiss factual information, the chatbot must stop the conversation.

When a user repeatedly dismisses factual information, the chatbot will need to stop the discussion and (re)state its underlying premise that factual information forms the basis of democratic dialogue.

11 USE CASES DEFINITION AND USER NEEDS ANALYSIS

In this section, we will define the three key use cases of TITAN, each designed to address specific challenges in combating disinformation. These use cases involve improving the critical thinking skills of higher education students, empowering NGOs to counter disinformation, and supporting migrants and refugees in integration processes. Additionally, we will present the user requirements collected from the specific actors within these use cases, ensuring our project meets their needs effectively. The defined use cases of our project highlight the diverse challenges posed by disinformation in different contexts.

11.1 USE CASE 1: HIGH EDUCATION FIGHT AGAINST DISINFORMATION

Use Case 1 is designed to improve the critical thinking skills of higher education students. By setting up a knowledge sharing between higher education institutions and the public broadcaster journalist, we want to help them develop a critical mindset that can help them further in the rest of their career.

Within the TITAN project we are developing a trajectory of several weeks that students can follow when they want to learn more about critical thinking and dealing with disinformation signals.

This trajectory starts with a workshop where the students learn more about critical thinking, where they take a test to measure their critical thinking skills and where the TITAN chatbot is introduced. Through this workshop, the students also come into contact with the journalists, who then gain insight into their living environment and the fake news the students are confronted with.

The students then test this chatbot for a number of weeks, by having conversations with the bot and by following the various microlessons, they improve their critical thinking skills. This trajectory is concluded with a closing event where knowledge and experiences are exchanged between the various parties.

Students who study at higher education institutions live in a time when they have a lot of information at their disposal via the internet that they can consult when they delve into different subjects. Thus much that they sometimes face an information overload, where it is not always easy to know which information is reliable and which is not. Higher education institutions play a crucial role in equipping students with the necessary tools to navigate disinformation. Institutions are responsible for fostering critical thinking and a 'fact-checking state of mind' among their students. VRT (Vlaamse Radio- en Televisieomroeporganisatie), as public broadcaster, can support them in doing so because it is also their task to inform and educate citizens.

To date, students, interns and teachers have access to unlimited sources of information, but it is becoming increasingly difficult to assess the reliability of information. A news organization like the VRT does not have the means to monitor all the news that circulates and to publish a fact check about it. It is not always clear to journalists which disinformation is spreading among students, on which platforms. The communication between the different parties is limited and can be improved to the benefit of all parties.

The TITAN tool provides an opportunity for a more unified approach towards citizens on acquiring critical reasoning skills, supporting learning fact-checking skills and on accessing verified information. All of this is done while connecting in this process with other important actors (peers, journalists, fact-checker communities, teachers). The aim is to help the higher education students acquire a 'fact-checking state of mind' effectively battle disinformation. In the larger context, these citizens can serve as valuable resources for journalists, acting as 'human fact-checking sensors.' They can help journalists identify disinformation more quickly, improve the quality of fact-checks, and effectively direct fact-checking efforts towards the most relevant publication platforms.

11.1.1 Use case 1 user needs analysis

In this section the user requirements collected by the target groups of use case 1 are listed and analyzed one by one to understand in depth the concerns of the users and be able to address their concerns on the TITAN platform. The first list of requirements comes from the students that participated in the user needs collection through user stories that they have created.

Students' needs

- As a user I want to be informed about the purpose of the tool

Informing users about the tool's purpose is crucial for setting expectations, building trust, and enabling effective utilization of its features.

- Users need info about the duration of one session (conversation)

Providing users with information about session duration allows for better time management and informed decision-making. Users can plan and allocate their time accordingly, deciding if they have enough time for a complete session or if they should engage in shorter sessions that fit their schedule better. Clear communication about session duration empowers users to make informed choices, optimizing their experience and fostering a sense of control over their time.

- Upload new articles (URL) to TITAN-tool

Enabling users to upload new article URLs to the TITAN-tool expands the content database, promotes continuous learning, and encourages user engagement in combating disinformation.

- Highlight text as part of micro-lesson

The ability to highlight text as part of a micro-lesson is valuable for emphasizing key points and facilitating active learning. By allowing users to highlight specific text, the tool promotes focus and helps users identify and retain important information. This feature enhances the micro-lesson experience by encouraging critical thinking and enabling users to engage more deeply with the content. Overall, the text highlighting capability serves as a valuable tool for reinforcing learning objectives and facilitating meaningful comprehension within the context of the micro-lessons.

- Can start a session (conversation), pause it and continue later

The ability to start a session, pause it, and continue later is essential for accommodating users' busy schedules and promoting flexibility in their engagement with the chatbot. This feature allows users to initiate a conversation at their convenience and pause it whenever necessary, without losing progress. By enabling sessions to be resumed later, users can seamlessly pick up where they left off, ensuring a continuous and uninterrupted learning experience. This functionality empowers users to manage their time effectively, aligning their engagement with the chatbot to fit their individual needs and preferences. Ultimately, the option to start, pause, and continue sessions enhances user satisfaction and encourages regular usage of the tool.

- Users can consume the micro-lessons at any time, and not only during a conversation

Allowing users to consume micro-lessons at any time, independent of a conversation, enhances the accessibility and flexibility of the learning experience. By decoupling micro-lessons from conversation sessions, users have the freedom to engage with the educational content at their convenience. This feature accommodates different learning preferences and schedules, enabling users to access and benefit from the micro-lessons whenever it suits them best. Whether they prefer to explore the lessons as standalone resources or reinforce their understanding after a conversation, this flexibility empowers users to learn and enhance their critical thinking skills on their own terms, fostering a self-paced and personalized learning journey.

- Users can filter articles based on disinformation signals, so they can exercise only on specific signals

The ability for users to filter articles based on specific disinformation signals enables targeted and focused practice in identifying and combating specific types of disinformation. By providing this filtering feature, users can customize their learning experience and concentrate on practicing critical thinking skills in areas that they find particularly relevant or challenging. This functionality allows users to develop a deeper understanding of specific disinformation signals, enhancing their ability to recognize and evaluate them in real-world situations. By empowering users to exercise their critical thinking skills in a more tailored manner, the tool promotes effective learning and equips users with the necessary tools to combat disinformation more effectively.

- Users can flag possible mistakes, made by the system

Enabling users to flag possible mistakes made by the system is important for fostering a collaborative and iterative learning process. By allowing users to provide feedback on potential errors, the tool can continuously improve its performance and accuracy. This feature promotes user engagement and participation, as users feel empowered to contribute to the refinement of the system. Moreover, by addressing flagged mistakes, the tool can enhance its overall reliability and credibility, ensuring a more reliable and effective learning experience for all users.

- Users expect recommendations of other relevant articles they can analyze (to start a dialogue with)

Meeting users' expectation of receiving recommendations for other relevant articles to analyze fosters continued learning and exploration. By providing suggested articles for analysis, the tool encourages users to delve deeper into the subject matter and engage in meaningful dialogue. These recommendations enhance user experience by offering a curated selection of content that aligns with their interests and expands their knowledge base.

- Users need rewards/achievements when finishing a conversation

Rewarding users with achievements upon completing a conversation serves as positive reinforcement and enhances user motivation and engagement. By acknowledging and celebrating their progress, users are encouraged to continue participating and actively engaging with the tool. These rewards provide a sense of accomplishment and recognition for their efforts, fostering a sense of satisfaction and incentivizing users to explore more conversations and further develop their critical thinking skills. The implementation of rewards and achievements adds a gamification element to the learning experience, making it more enjoyable and encouraging users to maintain their commitment to combating disinformation.

- Users want to share something they have learned

Allowing users to share something they have learned is valuable for promoting knowledge dissemination and encouraging active participation. By providing sharing functionality, users can pass on valuable information or insights they have gained from the tool to others, fostering a culture of knowledge sharing. This feature empowers users to become advocates for critical thinking and combating disinformation, as they can contribute to spreading awareness and promoting informed discussions. By facilitating sharing, the tool leverages the user's enthusiasm and encourages a ripple effect, amplifying the impact of their learning and contributing to a collective effort in addressing disinformation challenges.

- Users need encouragements

Providing users with encouragement is important to foster a positive and supportive learning environment. By offering words of encouragement, the tool can motivate users to persist in their efforts to develop critical thinking skills and combat disinformation. These encouragements can take various forms, such as positive feedback, affirmations, or reminders of progress made. By boosting user confidence and morale, the tool instills a sense of accomplishment and empowers users to continue their learning journey with enthusiasm and determination. Ultimately, the provision of encouragements promotes a constructive and uplifting user experience, enhancing user engagement and the overall effectiveness of the tool.

- After a conversation the user receives concrete tips & tricks to factcheck information by themselves
 Providing users with concrete tips and tricks to fact-check information by themselves after a conversation is instrumental in empowering them with practical skills and promoting independent critical thinking. These tips and tricks can include reliable fact-checking sources, critical evaluation techniques, or strategies for verifying information. By equipping users with these tools, the chatbot encourages self-reliance and enables users to become more discerning consumers of information in their daily lives. This feature enhances the long-term impact of the conversation, enabling users to apply fact-checking principles beyond the chatbot interaction and fostering a more informed and vigilant approach to combating disinformation.

11.1.1.1 Journalists' needs

In the following requirements we give an overview of the user requirements from the journalists' point of view.

- As a journalist I want to manually add articles to the sub-site via the CMS
 Allowing journalists to manually add articles to the sub-site via the CMS (Content Management System) provides them with direct control and flexibility over the content published. This feature enables journalists to contribute relevant and timely articles that align with the goals and focus of the sub-site. By granting them access to the CMS, journalists can ensure the accuracy and quality of the content they publish, maintaining the integrity of the platform. This functionality promotes journalistic independence and collaboration, empowering journalists to actively participate in combating disinformation by sharing reliable and verified information with the users of the sub-site.
- The selection of articles needs to be accessible by the students
 By making the selection of articles readily accessible, students can easily browse and engage with the content. This accessibility feature accommodates diverse learning styles and preferences, allowing students to explore and analyze articles at their own pace and convenience. By removing barriers to access, the tool fosters a more inclusive and student-centered learning environment, empowering all students to actively participate in critical thinking and combating disinformation.
- As a journalist I need to have the ability to label an article as fake (wrong) or real (wright) in the CMS
 Granting journalists the ability to label articles as either fake (wrong) or real (correct) in the CMS (Content Management System) facilitates effective information management and promotes accuracy in content categorization. This feature empowers journalists to exercise their professional judgment and expertise in evaluating the veracity of articles. By labeling articles accordingly, journalists contribute to the creation of a reliable and trustworthy database of content. This functionality enhances the credibility and integrity of the platform, assisting users in discerning between reliable and misleading information, and ultimately supporting the fight against disinformation.
- As a journalist I can change manually the detected signals of disinformation & arguments, so that I can correct mistakes the system made
 Providing journalists with the ability to manually change the detected signals of disinformation and arguments allows for human intervention and correction of any mistakes made by the system. This feature acknowledges the expertise and nuanced understanding that journalists bring to the analysis of information. By enabling manual adjustments, journalists can ensure the accuracy and precision of the detected signals, enhancing the overall reliability of the system's analysis. This functionality promotes collaboration between human expertise and automated systems, resulting in more accurate and insightful assessments of disinformation, and supporting the journalist's role in combating misleading information effectively.

- As a journalist I can moderate the incoming articles (delete, on hold, approve, etc.), so that content is adapted to target audience

Empowering journalists with the ability to moderate incoming articles, including the ability to delete, put on hold, or approve content, ensures the adaptation of the content to the target audience. This moderation feature allows journalists to maintain editorial control and uphold the standards and guidelines of the platform. By reviewing and curating incoming articles, journalists can ensure the relevance, accuracy, and suitability of the content for the intended audience. This functionality supports the integrity of the platform, promoting responsible information dissemination and fostering a trusted environment for users to engage with reliable and high-quality content.

- As a journalist I want to see what students share to the platform (trends of topics), so that I get insights in the lifeworld of youngsters

The ability for journalists to see what students share on the platform, including trends of topics, provides valuable insights into the lifeworld of youngsters. By monitoring student contributions, journalists gain a deeper understanding of the interests, concerns, and perspectives of the younger generation. This insight enables journalists to stay attuned to current trends and effectively address relevant issues in their reporting. By leveraging these insights, journalists can foster meaningful connections with the target audience and produce content that resonates with their experiences, fostering engagement and promoting a more inclusive representation of the youth's perspective.

11.2 USE CASE 2: NGO'S CITIZENS AS ACTIVISTS

Members of NGOs are often highly engaged and seek to expose malpractices from both governments and organizations and especially in the recent times of global spread of hostile disinformation and propaganda from authoritarian states. However, to do so, they need to carefully verify information, to avoid spreading disinformation themselves. As such, TITAN will provide these citizens with an ecosystem that verifies or counters specific disinformation. In the pilot, we collaborate with the SELDI initiative (i.e., Southeast Europe Leadership for Development and Integrity), a good governance and anti-corruption coalition in Southeast Europe (SEE), involving partners from fourteen countries. Some of SELDI partners already have dealt with the issues of disinformation and propaganda, including running fact-checking activities and engaging members and other citizens in these activities. Major issues of interest are the disinformation targeting core European values (democracy, human dignity and rights, equality, rule of law, and freedom of media) in the SEE countries.

When battling disinformation about the European Union core values and the respective national and European policies and institutions, activists currently are hindered by difficulties to assess the trustworthiness of information despite the vast availability of information sources and resources. The citizens and particularly, grass-root activists among them, lack a complete framework that could help them to identify, assess and tackle disinformation and propaganda, including engaging and empowering other citizens in these activities. Moreover, civil society activists, but also the NGOs, are confronted with disbelief coming from either like-minded citizen due to absence of critical thinking or antagonists due to ideological reasons. The erosion of citizens' trust towards institutions of democratic society is reinforced by conspiracy theories stemming from the current COVID-19 crisis, which has broken many of the social links and pushed further the use of the online environment as a substitute of in-person social life.

The TITAN solution will provide civil society activists but also their organisations with a comprehensive and unified framework that will increase their capacity for critical thinking and assessing the trustworthiness of the information they consume and communicate, enhance the fact-checking knowledge and skills. They will also help with (i) training and testing algorithms, and (ii) facilitating and improving capacity for fact-checking, including ensuring platform and framework for increasing the crowd-source fact-checking and tackling misinformation activities.

11.2.1 Use case 2 user needs analysis

In this section the user requirements collected by the target groups of use case 2 are listed and analyzed one by one to understand in depth the concerns of the users and be able to address their concerns in the TITAN platform.

- As a user, I want the tool to adapt and personalize my learning path based on my progress and preferences. Enabling the tool to adapt and personalize a user's learning path based on their progress and preferences is instrumental in enhancing the learning experience. This feature tailors the content and interactions to align with each user's unique journey and needs. By tracking progress and considering preferences, the tool can deliver more relevant and engaging content, promoting sustained motivation and skill development. This adaptability fosters a user-centric approach to combating disinformation, ensuring that users receive the support and resources they need to continuously improve their critical thinking skills effectively.

- As a user, I want the ability to collaborate (in real time) with other NGO experts on analysing articles and signals.

Enabling users to collaborate in real-time with other NGO experts on analyzing articles and signals fosters a collaborative and synergistic approach to combating disinformation. This feature allows experts to pool their collective knowledge and insights, resulting in more comprehensive and accurate assessments. Real-time collaboration enhances the efficiency of analysis, enabling faster responses to emerging disinformation threats. It also facilitates knowledge sharing and capacity building within the expert community, ultimately strengthening the tool's effectiveness in addressing the challenges of misleading information.

- As a user, I want to customize notification settings to receive updates on new content, discussions, and achievements.

Offering users the ability to customize notification settings for updates on new content, discussions, and achievements is vital for providing a personalized and user-centered experience. This feature empowers users to tailor their engagement with the tool according to their preferences and priorities. Users can stay informed about the aspects of the tool that matter most to them and receive notifications at the frequency of their preference.

- As a user, I want clear information about how my data is used and shared within the tool.

Providing users with clear information about how their data is used and shared within the tool is essential for transparency and trust. This feature ensures that users have a comprehensive understanding of how their information is handled, promoting a sense of control and data privacy. Clear data usage guidelines foster trust in the tool's operations, reassuring users that their data is handled responsibly and in alignment with their expectations.

- As a user, I want the learning experience to include gamified elements, such as challenges, rewards, and leaderboards.

Incorporating gamified elements like challenges, rewards, and leaderboards into the learning experience enhances user engagement and motivation. These features add a fun and competitive dimension to the tool, making the learning journey more enjoyable and dynamic.

- As a user, I want the option to download content for offline access and learning.

Offering users the option to download content for offline access and learning provides flexibility and convenience in their learning journey. This feature enables users to continue their critical thinking and disinformation combat education even when they have limited or no internet connectivity. It accommodates various learning environments, whether it's during a commute or in areas with unreliable internet access. By

allowing offline access, the tool ensures that users have uninterrupted access to valuable resources, fostering a continuous and self-paced learning experience that aligns with their individual needs and preferences.

- As a user, I want the ability to easily share interesting content or achievements on social media platforms. Providing users with the ability to effortlessly share interesting content or achievements on social media platforms enhances user engagement and encourages knowledge dissemination. This feature leverages the power of social networks to amplify the tool's impact by allowing users to share valuable insights, articles, or achievements with their online communities. It promotes a sense of accomplishment and recognition among users while contributing to a wider awareness of critical thinking and disinformation combat.
- As a user, I want the ability to create and share content, such as articles or analyses, with the community. Enabling users to create and share content, such as articles or analyses, within the community empowers them as active contributors and fosters a collaborative learning environment. This feature encourages users to share their knowledge and insights, enriching the collective understanding of critical thinking and disinformation combat. It promotes engagement, creativity, and knowledge sharing among users, reinforcing the tool's community-driven approach.
- As a user, I want the ability to provide feedback on the tool's functionality and content. Offering users the ability to provide feedback on the tool's functionality and content is crucial for continuous improvement and user-centered development. This feature acknowledges the importance of user input and insights in shaping the tool's effectiveness. User feedback serves as a valuable source of information for identifying areas of enhancement, ensuring that the tool evolves in response to user needs and preferences.

11.3 USE CASE 3: IMMIGRATION

The European Union is experiencing large migrant and refugee flows, both internal and external, due to the economic environment and socio-economic instability in the greater region, with large populations near or within war zones. The public perception of the “refugee crisis”, framed by media shifting from careful tolerance over summer 2015, to an outpouring of solidarity and humanitarianism in September 2015, and to a securitization of the debate and a narrative of fear starting in November 2015, has reasonably represented an obstacle and a threat to effective integration processes. Furthermore, information manipulation campaigns contributed to increasing hatred against minorities and hence they have a direct negative impact on the fundamental right to human dignity.

UNINETTUNO launched the University for Refugees initiative in 2016, providing migrants and refugees access to degree program scholarships provided by UNINETTUNO. Data collected shows the variety and richness of the backgrounds, knowledge and skills of the users enrolled in University for Refugees. Now they can work together with UNINETTUNO students for increasing their XXI century skills (collaboration, advanced digital skills, critical thinking) and fighting the fake narratives about migrations’ motivations and impact on EU. Students from Psychology, Communication, Law and Economics faculties (from 167 different countries) collaborate creating stories and narratives, enhancing the visibility of the initiative and above all impacting on the general perception of the migration phenomenon in EU.

TITAN tool will support migrants, refugees and students’ counter-narrative ideation and definition, providing support in verifying information, access to reliable sources, and describing how a fake news spread in social media. TITAN will support both individual and collaborative researchers, while migrants, refugees and students will benefit from the opportunity for acquiring critical reasoning skills, fact-checking skills in an inquiry-based (informal) learning setting. In the broader ecosystem, the contents produced by the participants with the support of TITAN tool will impact citizenship at large through social media and web-based channels,

allowing participants to act as “advanced fact-checking based storytellers” and the social media users to receive validated and reliable information about the actual migration process and its background motivations and impact in their countries.

11.3.1 Use case 3 user needs analysis

In this section the user requirements collected by the target groups of use case 2 are listed and analyzed one by one to understand in depth the concerns of the users and be able to address their concerns in the TITAN platform.

- Dialogue presentation should follow a mixed text-image approach.

Employing a mixed text-image approach in dialogue presentation enhances user engagement and comprehension. This feature combines textual information with relevant images to make the content more visually appealing and informative. Visual elements aid in conveying complex concepts, making the dialogue more accessible and engaging. By integrating images into the presentation, the tool promotes a well-rounded and immersive learning experience, catering to various learning styles and preferences.

- Dialogues should integrate standard pictograms to text

Integrating standard pictograms alongside text in dialogues is a valuable visual aid for enhancing user understanding and accessibility. These pictograms provide universally recognizable symbols that complement textual information, making the content more comprehensible, especially for migrants or those with varying levels of literacy.

- Dialogues should be presented in multiple languages.

Presenting dialogues in multiple languages is essential for reaching a global audience and fostering inclusivity. This feature ensures that migrants from diverse linguistic backgrounds can access and engage with the tool's content effectively. It promotes a more inclusive learning environment and supports migrants in their native languages, eliminating language barriers and making critical thinking education accessible to a broader audience

- Purpose of the app should be well introduced in a video

Introducing the purpose of the app through a video is an engaging and informative approach. This feature provides users with a dynamic and visual overview of the app's mission and functionalities, making it more accessible and relatable. A video introduction effectively communicates the tool's goals in a concise and compelling manner, capturing users' attention and clearly outlining the educational objectives and benefits.

- The key features/ services of the app should be demonstrated through a video or animated tutorial

Demonstrating the key features and services of the app through a video or animated tutorial is an effective way to onboard users and showcase the tool's capabilities. This feature provides users with a visual and interactive guide, ensuring they grasp the app's functionalities quickly and thoroughly. It simplifies the learning curve, enhances user confidence, and encourages active engagement with the tool's critical thinking exercises and disinformation combat services. By offering a comprehensive tutorial, the tool empowers users to make the most of its features, promoting a successful and impactful learning experience

- Dialogue interactions should be around 1-2 minutes each

Attention span in social media users and specifically in young generation is short; to be attractive and engaging we cannot imagine having long conversations (despite them being or not Socratic). Even if the dialogue itself could be structured on a wide conversation tree, interactions should be managed as a set of “episodes”

- Interaction modalities should be flexible

For example, in a dialogue users should have the opportunity to pick from a set of options (multiple choice) leading to different dialogue branches according to the answer of each user. Ideally: in some case managing open answers from the users.

- Dialogues should provide a results/ conclusion to the user.

Users, not only migrants, should capture the immediate pragmatic value of the tool: not only because they are in an educational setting, but also because they might get a reward/ tangible benefit (e.g., a badge with possibly gamification logic).

- The app should motivate the users to get engaged in several dialogues in order to assure critical thinking skills development.

Critical thinking skills development is assured by the involvement of the users in a series of interactions with the Titan tool through time. Users' motivation is one of the fundamental goals the app needs to accomplish. The design of a gamification strategy might be considered.

- App interface and interaction design should assure to meet usability, accessibility and universal design principles.

UC3 will involve users from a variety of EU and non-EU countries, with a wide spectrum of linguistic, cultural and technological literacy levels. Designing interactions with as much broad literacy issue as possible is also required to fight disinformation in vulnerable populations.

12 CO-CREATION OF THE REQUIREMENTS FOR THE TITAN AI SERVICE/PLATFORM

In February and March 2023, citizens from eight European countries actively participated in citizen co-creation workshops, which played a crucial role in shaping the development of our project. These workshops provided valuable insights that were processed by FONDEN TEKNOLOGIRADET and VRIJE UNIVERSITEIT BRUSSEL (see D3.1). As mandated by the Grant Agreement, these workshop findings formed the foundation for stakeholder involvement, ensuring their perspectives and concerns were integrated into the project.

To further involve stakeholders in the project's development, three separate two-hour workshops were conducted, focusing on key topics: "Informed consent and AI tools," "Data and Trust," and "AI learning tools and citizen's needs." The outcomes and conclusions drawn from these stakeholder workshops will be presented in the following pages as user requirements. These requirements serve as crucial guidelines for adapting the TITAN tool to address the concerns and meet the needs of citizens, ultimately striving to develop a trustworthy, ethical, and socially accepted tool.

The engagement of citizens and stakeholders in the co-creation process demonstrates our commitment to participatory decision-making and ensuring the project's outcomes align with societal expectations. By incorporating their insights, we aim to build a robust tool that not only addresses disinformation but also upholds principles of transparency, accountability, and inclusivity.

Through this comprehensive stakeholder involvement approach, we endeavor to create a collaborative environment where citizen voices are heard, concerns are addressed, and the final product reflects the collective efforts of a diverse range of stakeholders. The user requirements derived from these workshops will guide the ongoing development of the TITAN tool, ensuring its alignment with the values and expectations of both citizens and society as a whole.

12.1 CITIZEN INSIGHTS QUALIFIED BY EXPERTS - RECOMMENDATIONS FOR THE TITAN TOOL

1. Workshop on Informed Consent and AI Tools

Five stakeholders participated in the workshops and shared their expert knowledge on people-centered development of digital tools, social implications of AI, ethics of AI, validity of consent, legal standards for technology, and transparency, protection of personal data, and more.

The stakeholders were asked to consider these insights from the citizens: Citizens want to understand what they give consent to and why an AI tool needs access to their data. Furthermore, citizens are concerned about transparency and if they can trust an AI tool, which the stakeholders were also asked to consider.

2. Workshop on Data and Trust

Three stakeholders participated in the workshop and shared their expert knowledge on bottom-up empowerment infrastructures, a human-centric approach to the value of data, how individuals gain control of their own data, and more.

The stakeholders were asked to consider these insights from the citizens: Citizens want control with their data and transparency about how data is used. Citizens want a choice of what data to give to an AI system, to get access to specific functionalities, and have the option to not give any personal data. Additionally, the stakeholders were asked to consider citizens' concern about the security of how data is stored.

3. Workshop on AI learning tools and citizens' need

Four stakeholders participated in the workshops and shared their expert knowledge on the potential of AI and technology in teaching, what role technology should play in teaching, developing applications with societal impact, digital literacy among vulnerable youth, collaborative learning, and more.

The stakeholders were asked to consider these insights from the citizens: Citizens' fear AI oppressing free speech or creating an echo chambre of (dis)information as well as citizens wanting a tool that is not too time consuming, that is inclusive, non-invasive and where you can engage with the tool at different levels.

4. General recommendations from stakeholders

Across all three stakeholder workshops, the stakeholders said that TITAN must develop a tool that is close to people's need and have the project sufficiently disseminated after the ending of the project. In their experience, this type of project risks to create an output that is a state-of-the-art technology, but due to the lack of consideration of users' needs and insufficient dissemination, the technology does not create the impact and change that it aimed for.

12.2 USER REQUIREMENTS ON INFORMED CONSENT IN AI TOOLS

- The tool's focal point must be the user's interest.

The user's interest must be the focal point of the tool. Therefore, it is not sufficient to avoid doing evil, it must be an ambition to do good. Do good by avoiding third party cookies, selling user data, using data for commercial interest, using manipulative design e.g., tricking the user to give consent to cookies, and by avoiding surveillance of the user. Make sure there are no surprises to the user and that the use of their data will not have any cost or side effects for them.

- The user should only receive information that is necessary for decision making.

The information given to the user, for them to give consent, should be limited to avoid information overload. The user should know the information that makes a difference to their decision to give consent or not. Therefore, it is fundamental to identify what information is necessary and what is merely nice to know e.g., be specific about why the user needs to share specific kinds of data. Furthermore, it is necessary to paraphrase legal text in the consent form into citizen-centered language.

- The tool must create transparency by ensuring that users have an accurate understanding of the necessary information.

Information does not equal transparency, instead transparency is created when the information given to the user gives them an accurate understanding of the information. Users need to understand the purpose and possible risks of using the tool. There is not one way to inform the users, therefore, to give different users an accurate understanding of the information provided, the information must be available in multiple versions. Transparency can be created by utilizing different methods such as frameworks and storytelling, which uses text, images, visuals, and universal iconography. Achieving transparency should not require a big effort of the user, instead it should be embedded in the tool's functionality.

- The user must be able to exercise their data rights.

The user must be able to conveniently exercise their data rights, by making it easy to withdraw consent at any time and enable the users to change their mind on what they consent to while using the tool. Decisions about consent should be made in a context where the user has time to consider pros and cons properly and not be put in a situation where the user feels forced to consent to be able to use the tool.

- The tool must be able to explain the purpose of a personalized profile.

When offering a personalized profile to the user, the tool must be able to explain how personalization can assist the user in identifying disinformation. Explaining to the user the short- and long-term benefits and consequences of a personalized profile should make the user capable of deciding whether they want to use a personalized profile considering TITANs aim of fighting disinformation. The tool must map what data is being used for what purpose in the personalization and ask for consent for each specific use of data.

- The tool should offer granularity of consent to the user.

The user should have freedom of choice - consent should not be an all-or-nothing option. Users need to have different options to choose from. Granularity can be achieved by offering consent for time periods, for different data amounts and for gradual disclosure, asking for more data only when the functionality of the tool requires it.

12.3 USER REQUIREMENTS ON DATA AND TRUST

- The tool should only ask the user to provide the data necessary for the service.

The user should be the center of attention for data use. Therefore, the TITAN tool should not ask for data that is merely nice to have, but only data that is necessary to have to offer the service. Many digital services ask for more data than what the service needs - a habit that the tool should not repeat.

- The tool should delete user data as soon as it is not needed anymore.

To be transparent about data use, the tool should by default delete user data when it is no longer needed for the purpose the user was asked to provide it for. If the tool also wants to use user data provided to enable a specific functionality of the tool, e.g., to train the AI, the user should be asked if the user data can be saved for a longer period to train the AI.

- The tool must only use data for the purpose explained to the user.

The tool must consider that many users do not have the capabilities to understand technical explanation about data use and functionalities of the system, but the user must be explained the context and what they can gain from sharing their data. E.g., the tool has a clear purpose of wanting to do a critical thinking assessment of the user, which makes it possible to explain the purpose of the data users need to provide. Data should never be used for other purposes than what it was directly provided for.

- The tools must have values that build trust towards the user.

The tool must define its values and communicate them to the user. The values must not change over time, and they can be made legally binding by the provider of the tool, by e.g., ensuring that the provider will stay non-profit, that the tool will not be sold to third parties, and by using a Trustworthy AI setup. Objective and knowledgeable actors could be used to evaluate and recommend the tool to potential users, to create societal trust in the tool.

- The user must have trust in the data handling and storage.

Users' trust in the tool is core, but trust is challenged in many contexts, and creating trust in a digital tool has various aspects. Users need to trust that when they give a tool their data, the tool delivers something of value to them. There is no model or standard for how to create trust in a digital tool, but various steps can enable trust. The tool can earn trust via communication, strategic design, and the right set of values. The tool could use a bottom-up infrastructure (e.g., a data pod or data trusts), since having a neutral intermediate to handle and store user data could build trust with the user in a different way than a private service.

- The tool should make it optional for the user to make a personalized profile.

A personalized profile should not be mandatory for using the tool. Users will have different ways of using the tool. For some users it will be beneficial for them to follow e.g., their development in their critical thinking assessment, while for other users it will be a barrier to have a profile storing the assessment data on them. It should be possible to do the critical thinking assessment without a profile, but also do critical thinking assessment without the data being stored on the profile.

12.4 USER REQUIREMENTS ON AI LEARNING TOOLS AND CITIZENS NEEDS

- The tool could be a website platform with mobile first design.

Hosting the tool on a website allows it to be used in various contexts, e.g., in a teaching setup on a bigger screen. Young people can be reluctant to download an app, since it takes up space on their phone. With mobile-first design on a website you have a mobile friendly website that can also be used on bigger screens.

- The tool must have micro-lessons with different levels.

The micro-lessons must have different levels to start from since users will have different levels of critical thinking when they begin using the tool. If all users start at the same level, some users will not find the tool beneficial.

- The tool should have micro-lessons with various formats.

The formats of the micro-lessons should encourage reflection using e.g., exercises with informational videos and audio, and interactive quizzes.

- The tool should fulfil its purpose while requiring as little time as possible from the user.

The tool should balance the purpose of the tool and the expectation of the user's time engagement, to not create a barrier for people to use the tool. The time a user spends with the tool depends on how the user finds the tool and their engagement, but some users will spend maximum 10 min with the tool. The learning aspect should not be compromised to limit the time spent with the tool, but time aspects should be considered in the interaction with the chat-bot, the micro-lesson, and surveys.

- The tool must focus on formative assessment when giving feedback to the user.

Because the tool focuses on competencies of critical thinking, it should use formative assessment when giving feedback to the user. Ongoing feedback will allow the user to see their progress over time and give them the opportunity to reflect on what they have learned and what they want to learn next. The feedback should start with a positive description of their choices, followed by suggestions on what they can consider. Avoid a judgment tool where feedback is based on an end goal, grades, numbers, or skill-based feedback.

- The tool could have gamification elements that fit the vision and the purpose of the tool.

Gamification elements could help to get young people to engage with the tool, because it gives incentive to keep learning, give immediate feedback, create competition elements, and motivational nudges. However, the risk of adding gamification elements is that the user forgets the purpose of the tool – developing a critical thinking mindset – instead they return to the tool for the sake of the game.

- The tool should have inclusive push notifications methods.

The tool should apply inclusive push notification methods, by allowing the user to decide if they want notifications or not. Push notifications are ok to use if the user is notified when using the tool, not outside of the tool, unless the user has agreed to get notification outside of the tool e.g., on email. Push notification could be a symbol that pops-up, and if the user presses the symbol a brief message will show, which is not repetitive.

- The tool must consider disabilities among users in the design of the tool.

The tool must consider web accessibility to accommodate users with various disabilities. E.g., deciding on symbols and colors of push notifications that people with cognitive disabilities or visually impaired people can also use.

- The tool could include a dialogue between users.

Group dialogue and collaboration or competition with other users could be an amplifier for people to engage with the tool. A broader social network could be considered for the tool, where the user consults the tool to learn from previous users' actions and reflection on disinformation. Dialogue between users could enable engagement because users learn from real people's experience, they communicate with other users and help each other learn.

- The tool must be able to consider personal differences.

Ensuring that the tool considers personal differences in guidance and training for identifying disinformation signals is vital for effectiveness. People come from diverse backgrounds and have varying levels of prior knowledge and cultural contexts. Tailoring guidance and training to individual needs acknowledges these differences, making the learning experience more relevant and impactful. Additionally, incorporating national differentiation further enhances the tool's adaptability by addressing specific disinformation challenges unique to different regions or countries. This personalized and culturally sensitive approach fosters a more inclusive and effective disinformation combat system that respects individual differences while promoting critical thinking skills universally.

- The tool should interact with the user in the consumption of the news and no later start.

Ensuring that the tool interacts with users at the time they begin consuming news, rather than later when they initiate a conversation, is crucial for timely and proactive engagement. Many users share news privately as soon as they encounter it online. By intervening at this initial stage, the tool can provide real-time feedback, disinformation signal identification, or critical thinking prompts, effectively curbing the spread of potential disinformation.

- Warning/highlight of text to identify disinformation.

Incorporating warnings or highlights within text to identify disinformation is a practical and user-friendly approach. This feature helps users quickly spot potentially misleading or false information within an article or news piece. By drawing attention to potential disinformation, the tool empowers users to make informed judgments about the content they encounter, strengthening their critical thinking skills and enhancing their ability to recognize and combat misleading information effectively.

- A system with high user friendliness.

A system with high user-friendliness is essential for a positive and productive user experience. This feature ensures that the tool is intuitive, easy to navigate, and accessible to users of all backgrounds and abilities. A user-friendly system simplifies the learning process, reduces frustration, and encourages active engagement.

- A system with good design on the interface.

A system with a well-designed interface is crucial for providing an engaging and visually appealing user experience. This feature encompasses intuitive layouts, clear navigation, and aesthetically pleasing elements that enhance user engagement. A thoughtfully designed interface not only makes it easier for users to access and interact with the tool but also reinforces their trust in its reliability and professionalism.

- A system with general fast response (time efficient).

A system with fast response times is essential for user efficiency and satisfaction. This feature ensures that users can access information, complete tasks, and interact with the tool swiftly, without unnecessary delays or waiting times. A responsive system not only saves users valuable time but also enhances their overall experience, making their interactions with the tool smooth and seamless.

- A system with limited bugs.

A system with limited bugs is crucial for providing a reliable and frustration-free user experience. This feature ensures that the tool operates smoothly, without disruptive glitches or errors. A bug-free system promotes user confidence in the tool's functionality and credibility. It reduces user frustration, enhances trust, and encourages continued use.

13 FOCUS GROUPS WITH TITAN PARTNERS

Within the domain of user needs analysis, we find a multifaceted approach that extends beyond the perspectives of the three primary target groups and the insights generated during co-creation sessions. Here, technology experts affiliated with the consortium assume a pivotal role in the meticulous refinement and augmentation of user requirements. Drawing from their extensive technical expertise and wealth of experience, these experts serve as essential bridges, connecting the intricacies of user needs with the technical capabilities of the chatbot. Their invaluable insights form the cornerstone of ensuring the tool's viability, security, scalability, and overall efficacy in the battle against disinformation. It is their contributions that breathe life into innovative, practical solutions capable of navigating the intricate terrain of the digital landscape. Moving forward, we have strategically identified and categorized the user needs analysis into distinct segments to ensure a comprehensive understanding of the requirements. This segmentation allows us to delve deeply into specific aspects critical for the development of the chatbot tool.

Socratic Method

Centering our attention on the Socratic method (see details in Section 4.3.3), our goal is to gain comprehensive insights into how users can most effectively interact with the chatbot to cultivate and enhance their critical thinking abilities. This entails a deep exploration of how users not only engage with but also actively respond to this pedagogical approach. Acquiring a profound understanding of how users derive benefits from the Socratic method is of utmost importance, as it serves as the bedrock for crafting the chatbot's interactions in a manner that aligns seamlessly with the users' needs, learning styles, and educational objectives. This comprehension is the linchpin in our endeavour to ensure that the chatbot becomes a potent tool for the cultivation of critical thinking skills, providing users with an enriching and personalized learning experience.

- The user may need to possess background knowledge about Socrates and his Socratic method. The user will be able to keep up with the procedure followed (6 steps of Socratic dialogue) to reach a conclusion and understand the outcome of the dialogue.

- The user may need to be informed about the type of questions and answers that are needed to converse with a chatbot using the Socratic method.

Since the chatbot can engage in two-way conversation, it would be good for the user to know if the chatbot can ask and answer open-ended questions, close-ended questions or/and multiple-choice questions.

- The user may need to be informed whether the data/output of his/her interaction with the Socratic chatbot will be anonymous or public.

Having agreed on terms about sharing conversation data before starting the conversation, may make the user more willing to have a conversation with the Socratic chatbot.

- The user may need to give some feedback and rate his/her dialogue with the Socratic chatbot. The user will be able to rate his/her conversation with the Socratic chatbot (e.g., from 1-5), write comments and suggest ways (if the user wishes) to improve the chatbot.

- The user may need to possess background knowledge about disinformation and disinformation signals. The user may require the meaning of disinformation or/and disinformation signals prior to chatting with the Socratic chatbot for more accurate communication.

- The chatbot may need to give concrete answers.

The user may need specific answers from the model, without made up answers, and if the chatbot doesn't know a specific answer, it could request more information.

Critical thinking assessment

Delving deeply into the domain of critical thinking (a comprehensive analysis can be found in Section 3) as a distinct category offers us a valuable vantage point from which to gain a nuanced understanding of the precise skills and cognitive processes that users aspire to refine and fortify. This deeper comprehension serves as the cornerstone of our approach, as it empowers us to meticulously tailor the content and methods deployed by the chatbot.

With this knowledge in hand, we can seamlessly align the chatbot's instructional content and pedagogical methods with the distinct objectives and aspirations of our users. It ensures that the chatbot becomes a finely tuned instrument, finely attuned to the individualized goals of users who are on a journey to evolve into more discerning consumers of information. In essence, this process allows us to craft a personalized and impactful learning experience that empowers users to critically evaluate and navigate the vast sea of information with confidence and acumen.

- Initial questionnaire or mini survey

The user should answer a short questionnaire to assess some individual and personal dimensions associated with critical thinking (cognitive and reflective skills).

- Critical thinking assessment: dispositional aspect

The user should respond to a number of statements about their usual way of approaching a problem, answering on a 7-point frequency scale (from Never to Always)

- Critical thinking assessment: objective multiple-choice test

After a brief description of the scenario, the user must respond regarding the accuracy of a statement among few possibilities. The answers are either correct or incorrect.

Disinformation Signals

The meticulous analysis of disinformation signals (see details in Section 4.2) offers us a unique vantage point from which we can discern and classify the myriad forms of misinformation and disinformation that users commonly encounter in the digital landscape. This intricate examination equips us with invaluable insights, acting as a compass that guides the crafting of the chatbot's functionalities.

By distilling these insights into the design process, we empower the chatbot with the capacity to astutely detect and effectively counteract these deceptive signals. This technological prowess is pivotal in our mission to guarantee that users are not just consumers of information but discerning evaluators who receive, trust, and share only accurate and verified information. In essence, it is through this careful analysis that the chatbot becomes a vigilant guardian of information integrity, fortifying the digital landscape against the pervasive threat of disinformation.

- The disinformation signals the chatbot is capable of detecting.

The user may need to be informed about the disinformation signals that the Socratic chatbot is able to detect before engaging in conversation with the chatbot.

- Detecting how many times a particular disinformation signal appears in an article.

The user may need to be informed about the frequency of a disinformation signal in an article.

- Detecting how many disinformation signals appear in an article.

The Socratic chatbot may need to inform the user about the number of disinformation signals that appear in an article to encourage further interaction and critical assessment.

- The user may need to select to learn about a specific number of disinformation signals.

It would be good for the user to select to learn about a specific disinformation signal or more than one disinformation signal.

- Providing information about uncertainty about the signal

There is no perfect method of evaluating disinformation, so the user may want to know that the system is not 100% sure of some assessment.

- Providing explanation why some information may contain signal that allows for classification as (potential) disinformation

Users should know what the automatic evaluation is based on so that they can independently verify whether they should trust such an evaluation, and what its strengths and weaknesses are.

Propagation Impact assessment

When a person shares an article in social media, it will reach people in her/his social network in the service. Some of these recipients will share the article again in their social networks, and the article propagates further to new audiences. The resharing pattern will be repeated and the propagation will continue basically indefinitely.

Propagation Impact Assessment (PIA) component predicts the effect of sharing an article in social media: “If I share this, how efficiently will it reach others?”

PIA is based on using and analyzing the content of the article (text) and the social network information of the user. The article content is obtained via URL or using the text content of the article directly. The information related to social network of the user is fetched using APIs of the social media service. However, this requires that the user must specify her/his user account (account name/id) in the social media service. The text content and social network information is preprocessed and used as an input for a machine learning algorithm. As a result, the algorithm gives an estimate for the propagation impact. The metric for the propagation impact depends on the social media service and the respective dataset used for training the model. For instance, in the case of Mastodon, the options are reblog count (i.e., retweet in Mastodon), reply count and favorites count. On some occasions it may be relative to user’s impact in the past (e.g., reblog count is used as a basis for calculating the relative impact in range 0...100).

In case the user does not want to share her/his social media account id/name, the PIA is computed based only on the text content. Text-only features, such as length of the post, presence of photos or videos and the amount of other shared posts with the same hashtags, are calculated and used as an input for the machine learning model that predicts propagation impact. In this case the model does not take the user-specific information, such as social network information, into account.

- The PIA component should give the results as a user understandable metric.

When the users ask for PIA, the PIA component computes the estimate and gives the result as numerical value. Ensuring that the PIA component provides results in a user-understandable metric is essential for enabling users to comprehend the reach and impact of disinformation.

- The user should be able to feed text content of an article to the PIA component.

This PIA tool relies on the actual textual content of an article to assess its potential impact. In other words, to gauge how widely and deeply the information within an article might spread and influence people, it requires access to the article's written content.

- The user needs to share his/her social media user account name with the TITAN system for a better estimation of the PIA.

Allowing users to share their account name for the selected social media service, along with access to their social graph information and historical post data, enables the system to create a more comprehensive and accurate social network analysis. This feature enhances the propagation impact assessment, as it can leverage the user's existing connections and posting history to gauge the potential reach and impact of disinformation shared within their network.

- The user should be able to feed meta-data related to the article to the PIA component.

Enabling users to feed metadata related to the article into the PIA tool enhances the depth and accuracy of the assessment. Metadata, such as publishing time, provides valuable context that can significantly influence the potential impact of an article. By incorporating this metadata, the assessment component can make more informed judgments about how the information might propagate and impact audiences.

- The user should be able to select the social media service that the account name is linked to from a list of supported social media services.

Allowing users to select the social media service to which their account name is linked from a list of supported platforms ensures flexibility and compatibility. This feature simplifies the process of connecting social media profiles to the tool by providing a user-friendly selection interface. Conceptually, users can choose the specific platform they use, enhancing the tool's adaptability to various social media services; in TITAN implementation, only one social media will be supported.

Micro-lessons

Gaining a profound insight into user preferences regarding micro-lessons (see details in Section 4.3) plays a pivotal role in our mission to craft educational content that is both captivating and highly effective. This deeper understanding serves as a guiding light, illuminating the path toward creating content that resonates with users on a personal level. Through a thorough examination of this category, we can engage in a careful dissection of user preferences, allowing us to discern the specific formats, durations, and presentation styles that best align with their learning inclinations. This dissection process facilitates a highly tailored approach in which we can fine-tune the chatbot's educational offerings to seamlessly match each user's individualized preferences. As a result, the chatbot's lessons become not just informative but also engaging, delivering knowledge in precisely the manner that resonates with and empowers users to enhance their critical thinking skills effectively.

- The user may need to get informed about the estimated amount of time that is required to take a micro-lesson.

Providing users with information about the estimated amount of time required to complete a micro-lesson is essential for effective time management and user engagement. This feature empowers users to make informed decisions about when to engage with a lesson, aligning their learning goals with their available time.

- The user may need to select the amount of time to take a micro-lesson.

Allowing users to select the amount of time they wish to dedicate to a micro-lesson offers flexibility and personalized learning. This feature recognizes that users have varying time constraints and learning preferences. By providing options for time allocation, users can tailor their engagement with micro-lessons to fit their schedules and learning pace.

- The user may need to select to take a micro-lesson concerning specific disinformation signals or concerning disinformation signals in general.

Allowing users to choose between taking a micro-lesson focused on specific disinformation signals or a general disinformation signals lesson enhances their learning experience. This feature caters to users with varying levels of familiarity and expertise in the field of disinformation. Users can opt for specialized lessons to deepen their knowledge in specific areas or choose general lessons for a broader understanding.

- Feedback on how well someone did the micro lesson.

Providing feedback on how well someone did in a micro-lesson is a fundamental element of the learning process. This feature offers users a valuable assessment of their performance, helping them understand their strengths and areas for improvement in critical thinking skills. Constructive feedback can include insights into correct answers, areas that need more attention, and suggestions for further learning.

Sustainability

The sustainability category shifts our focus toward the enduring aspect of user engagement and the chatbot's perpetuating relevance in users' lives. This facet of our analysis is instrumental in ensuring that the chatbot remains a steadfast companion on users' educational journeys over the long haul.

By immersing ourselves in this category, we embark on a comprehensive exploration of users' aspirations for continuous learning and the kind of sustained support they anticipate. This deliberate examination empowers us to forge a tool that not only imparts knowledge but also evolves in tandem with users, seamlessly adapting to their changing needs and preferences over time. In essence, the sustainability category serves as a compass, guiding us to craft a chatbot that is not just a momentary resource but a lifelong partner in fostering critical thinking and information literacy.

- The user may want to receive badges and points while using TITAN

Awarding users with badges and points as they use TITAN adds a gamified element to the learning experience. This feature encourages active participation and motivation by recognizing and rewarding achievements.

- The user may want to progress overtime (i.e., do better than the previous time you interacted with TITAN)

Allowing users to track their progress over time by aiming to perform better than their previous interactions with the system is a powerful motivational tool. This feature encourages users to set personal improvement goals and engage consistently with the tool to surpass their own achievements.

- The user may want to see how the TITAN community progresses (i.e., see statistics of how the community responded, their progress in TITAN, etc.).

Allowing users to track the progress of the community, including statistics on how the community responds and their overall progress within the system, promotes a sense of belonging and competition. This feature provides users with valuable insights into collective achievements, fostering a sense of community and shared goals.

- The user may want to have progress levels and ranks.

Implementing progress levels and ranks for users adds an element of achievement and recognition to their journey with the tool. This feature allows users to visually track their advancement and feel a sense of accomplishment as they climb through different levels and ranks.

14 USER REQUIREMENTS CONSIDERING LEGAL AND ETHICAL ASPECTS OF TRUSTWORTHY AI

In the development of any project, including the creation of a chatbot to combat disinformation, it is crucial to prioritize legal and ethical considerations. Adhering to legal requirements ensures compliance with applicable laws, regulations, and frameworks, protecting both the project stakeholders and the users of the tool. Ethical considerations, on the other hand, are vital for upholding principles such as fairness, transparency, and accountability, ultimately building trust with users and society at large.

Legal requirements serve as a baseline for responsible development and deployment of the chatbot. They encompass various aspects, including data protection and privacy laws, intellectual property rights, and regulations surrounding the use of artificial intelligence (AI) technologies. Compliance with these legal requirements helps safeguard user data, respect privacy rights, and mitigate potential legal risks associated with the project.

In parallel, ethical requirements play a fundamental role in ensuring that the chatbot operates in a manner consistent with societal values and expectations. Ethical considerations encompass a wide range of aspects, including transparency, fairness, non-discrimination, explainability, and user consent. By adhering to ethical principles, TITAN can mitigate biases, promote responsible use of AI, and build trust with users by demonstrating a commitment to their well-being and rights.

User requirements are equally important when considering the legal and ethical aspects of the project. Including user perspectives and feedback in the development process helps ensure that the chatbot aligns with their needs, concerns, and values. **User requirements, should encompass aspects such as transparency in data handling, control over personal information, clear explanations of how the chatbot operates, and mechanisms for user consent and opt-out options.** By incorporating user requirements, the project can address potential legal and ethical challenges while enhancing user trust and satisfaction.

To provide a comprehensive analysis of the legal, ethical, and user requirements, two separate deliverables have been created, specifically D1.3 Template and guidance for legal and ethical impact assessment and D1.4 Report on legal and ethical impact assessment. These deliverables contain detailed information on the legal and ethical frameworks that the project adheres to, as well as a comprehensive analysis of the user requirements that relates to the legal and ethical aspects. Interested stakeholders can refer to D1.3 and D1.4, but also in Section 5 of Part 1, for a thorough understanding of how the project meets legal obligations, upholds ethical principles, and integrates user requirements to develop a robust, responsible, and user-centric chatbot tool.

By prioritizing legal and ethical requirements and incorporating user perspectives, the project aims to ensure the development of a chatbot that not only effectively combats disinformation but also operates in a responsible, ethical, and user-friendly manner.

15 MOCK-UPS CREATION AND ITERATIONS

The creation of mock-ups serves as a bridge between user requirements and the final implementation of the chatbot system. Mock-ups provide a visual representation of the interface, helping to envision how the system would appear and function in practice. By considering the user requirements as a foundation, these mock-ups facilitate the exploration of different design possibilities and the refinement of features to ensure the system meets user expectations effectively.

During the collection of the user requirements, several iterations were conducted to enhance the mock-ups. Each iteration involved gathering partners' feedback and incorporating it into subsequent iterations, leading to an iterative design approach. The feedback collected during the mock-up stages helped identify areas for improvement, ensuring that the final application aligns with the needs, preferences, and goals of the users.

Collaboration with project partners was also a key aspect of the mock-up creation process. By sharing the mock-ups with partners, including disinformation experts, digital literacy advocates, and potential end-users, valuable insights and perspectives were gathered. This collaborative effort allowed for a comprehensive evaluation of the mock-ups, ensuring that the application design and functionality addressed the diverse requirements and expectations of the stakeholders.

Through the iterative process and partner collaboration, the mock-ups evolved, providing a visual representation of how the chatbot system would look and function. The incorporation of user requirements, along with the insights from partners, shaped an initial design that will be further changed if needed after the living labs and the users' feedback. In Fig. 15.1 below the reader can see the landing page of the TITAN platform as it is envisioned from the outcomes of the first year of the project. Similar pages have been created for the rest of the functionalities that appear in the landing page.

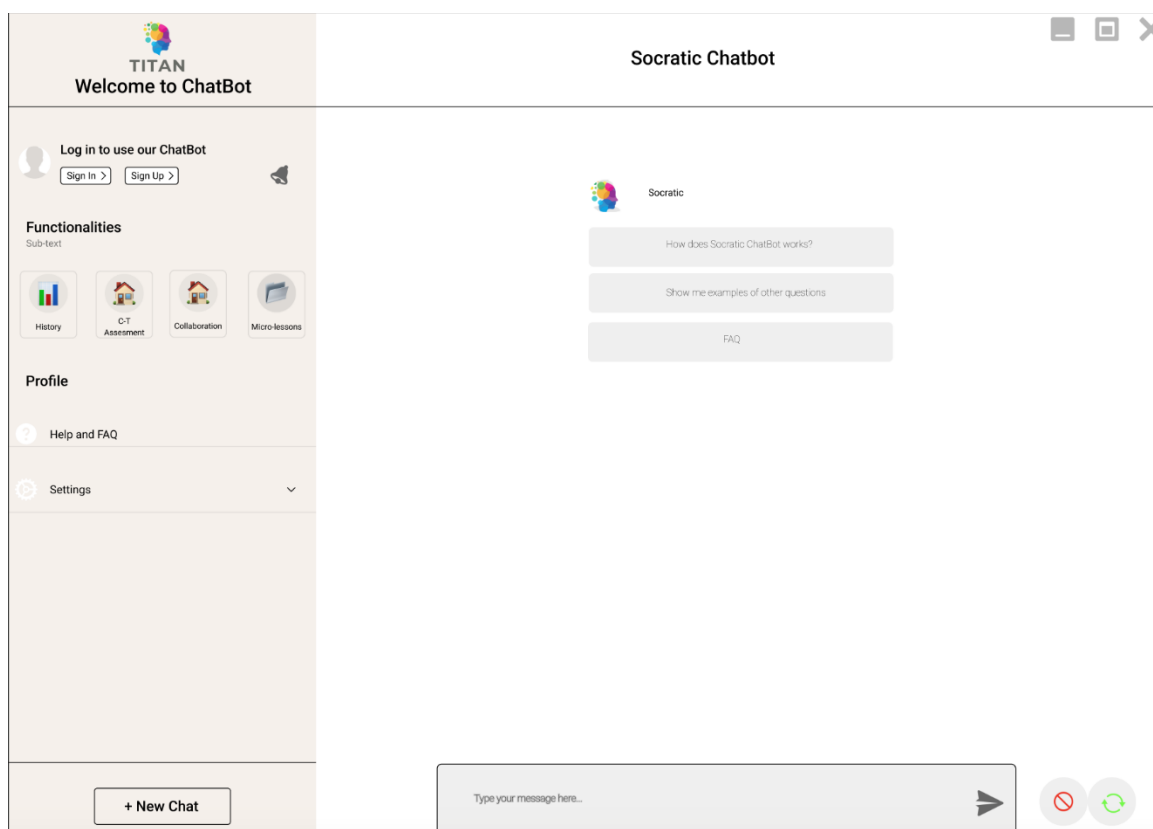


Figure 15.1. TITAN's chatbot landing page

16 CONCLUSION AND NEXT STEPS PART 2: USER NEEDS ANALYSIS

In conclusion, Part 2 has focused on the user requirements analysis for the development of an intelligent TITAN system aimed at combating disinformation. By exploring various sources such as the literature on user coping mechanisms, gathering user needs from user stories, conducting co-creation workshops, incorporating expert insights, and considering legal and ethical requirements, we have gained valuable insights into the specific user requirements necessary for an effective TITAN coaching environment.

The user-centered design approach has been at the core of this research, recognizing the importance of placing users and their tasks at the forefront of the development process. User experience requirements, encompassing factors such as transparency, efficiency of use, intuitiveness, low perceived workload, and accessibility, have been identified as key elements that should shape the design and functionality of the TITAN system.

Moving forward, the next steps in the development process involve translating the identified user requirements into concrete design and implementation decisions. The iterative design process, incorporating user feedback, will be crucial in refining the TITAN system's features, interface, and overall user experience.

Furthermore, close collaboration with stakeholders, including disinformation experts, digital literacy advocates, and potential end-users, will be essential in ensuring that the chatbot effectively addresses their needs and aligns with their expectations. This collaborative approach will foster continuous improvement and help establish the TITAN coaching system as a valuable tool in the fight against disinformation.

In conclusion, by adhering to user requirements, leveraging user-centered design principles, and iteratively refining the TITAN system based on user feedback, we can create a powerful tool that guides users in combating disinformation, empowering them to navigate the complex landscape of information with confidence and critical thinking.

The next steps in the development of the TITAN system involve several key actions, such as design iterations, implementation, user testing and evaluation, stakeholder collaboration, ethical considerations and deployment and evaluation.

More specifically, we will utilize the user requirements gathered through user stories, co-creation workshops, and expert insights to iteratively refine the design of the chatbot. Incorporate user feedback and conduct usability testing to ensure the system meets users' needs effectively. Furthermore, the design decisions must be translated into actionable development tasks and utilize appropriate technologies, such as natural language processing and machine learning, to implement the desired functionalities and features of the chatbot.

Comprehensive user testing will be conducted with a diverse range of potential users to gather feedback on usability, functionality, and overall user experience to identify areas for improvement and optimization.

Further engagement in ongoing collaboration with stakeholders follows, including disinformation experts, digital literacy advocates, and potential end-users. Solicit their input and insights to ensure the TITAN system aligns with their needs, goals, and expectations.

Continuously assess and address ethical considerations related to user privacy, data security, and potential biases in the system's algorithms. Strive for transparency and accountability in system operations.

Finally, the chatbot must be deployed in real-world settings and evaluate its effectiveness in guiding users to combat disinformation. Monitor its performance, gather user feedback, and assess its impact on users' ability to identify and counteract false information.

By following these steps, we can refine and optimize the TITAN system, continuously improving its capabilities to guide users in the fight against disinformation, ultimately fostering a more informed and resilient society.

TITAN

Grant Agreement #767227

17 ANNEX 1: FALLACIES AND FACT-CHECKING DATASETS

17.1 FALLACIES DATASET SUMMARY

Task	Labels	Dataset	Extracted features	Approach	Link	References
Fallacy Type Recognition (6 classes)	<ol style="list-style-type: none"> 1. ad hominem 2. appeal to emotion 3. red herring 4. hasty generalization 5. irrelevant authority 6. non-fallacious 	ARGOTARIO (general domain): 1160 examples	German 64-dimensional embeddings (Habernal et al., 2018 for German)	LSTM, SVM, CNN	https://github.com/UKPLab/argotario	Habernal et al., 2017
Fallacy Type Recognition (5 classes)	<ol style="list-style-type: none"> 1. Begging the question 2. Hasty Generalization 3. Questionable cause 4. Red Herring 5. None 	RIPOSTE: 18887 counter-arguments	SpaCy, Glove	LSTM, se2seq	https://www.kaggle.com/datasets/xabierdezuazo/riposte-a-large-corpus-of-counter-arguments?resource=download	Reisert et al., 2019
Propaganda Techniques Detection (18 classes)	<ol style="list-style-type: none"> 1. Loaded Language 2. Name Calling or Labelling 3. Exaggeration or Minimization 4. Doubt 5. Appeal to Fear/Prejudice 6. Flag-Waving 7. Causal Oversimplification 8. Slogans 9. Appeal to Authority 10. Black-and-White Fallacy 	PROPAGANDA (politics domain): 7485 news	BERT tokenizer (WordPiece)	BERT and variations	https://propaganda.qcri.org/nlp4if-shared-task/data/datasets-v2.tgz https://propaganda.qcri.org/fine-grained-propaganda-emnlp.html	Da San Martino et al., 2019b

	<ol style="list-style-type: none"> 11. Thought-Terminating Cliche 12. Whataboutism 13. Reductio ad Hitlerum 14. Red Herring 15. Strawman 16. Repetition 17. Bandwagon 18. Obfuscation, intentional vagueness 					
<p>Propaganda detection (binary)</p> <p>Propaganda Techniques Detection (14 classes)</p>	<ol style="list-style-type: none"> 1. Appeal to Authority 2. Doubt 3. Repetition 4. Appeal to fear-prejudice 5. Slogans 6. Black and White Fallacy 7. Loaded Language 8. Flag-Waving 9. Name Calling, Labeling 10. Whataboutism, Straw Men, Red Herring 11. Causal Oversimplification 12. Exaggeration, Minimisation 13. Bandwagon, Reducti ad hitlerum 14. Thought-terminating Cliches 	SemEval2020: 536 news articles (8,981 propaganda text snippets)	TFIDF, transformers	RoBERTa, CRF and transfer learning (https://github.com/ascern/semEval2020_task1)	https://zenodo.org/record/3952415#.ZEJ8qnZBy3A https://huggingface.co/datasets/sem_eval_2020_task_11	Da San Martino et al., 2020
Fallacy Type Recognition (9 classes)	<ol style="list-style-type: none"> 1. Appeal to authority 2. Appeal to majority 3. Appeal to nature 	Informal fallacies from	BERT tokenizer (WordPiece)	BERT, MGN	https://github.com/sahaisaumya/informal_fallacies	Sahai et al., 2021

	<ol style="list-style-type: none"> 4. Appeal to tradition 5. Appeal to worse problems 6. Black-or-white 7. Hasty generalization 8. Slippery slope 9. Non-fallacious 	Reddit: 3358 comments				
Logical Fallacy Detection (binary)	<ol style="list-style-type: none"> 1. Valid argument 2. Invalid argument 	BigBench (general domain): 2800 examples	SentencePiece, byte-pair encoding (BPE) (Hoffmann et al., 2022, Rae et al., 2022)	Transformer-based LM Gopher, Chinchilla, GPT-3, Jurassic-1, Megatron-Turing NLG, LaMDA (Hoffmann et al., 2022, Rae et al., 2022)	https://github.com/google/BIG-bench/tree/main/bigbench/benchmark tasks/logical fallacy detection https://huggingface.co/datasets/tasksource/bigbench/viewer/logical fallacy detection/train	Srivastava et al., 2022
Fallacy Type Recognition (14 classes)	<ol style="list-style-type: none"> 1. Faulty generalization 2. False causality 3. Circular reasoning 4. Ad populum 5. Ad hominem 6. Fallacy of logic 7. Appeal to emotion 8. False dilemma 9. Equivocation 10. Fallacy of extension 11. Fallacy of relevance 12. Fallacy of credibility 13. Miscellaneous 14. Intentional 	LOGIC (online educational material): 2449 examples LOGICCLIMATE: 1109 examples	Contextualized embeddings / NLTK	BERT and variations, Electra and variations, GPT3	https://github.com/causalNLP/logical-fallacy	Jin et al., 2022
Coarse-grained Fallacy Classification (4 classes)	<ol style="list-style-type: none"> 1. Fallacy of Relevance 2. Fallacy of defective Induction 3. Fallacy of presumption 	LOGIC (online educational material): 2449 examples	Word2Vec / transformer embeddings	BERT/ DeBERTa/ DistilBERT/ RoBERTa/ Electra	https://github.com/tmakesense/logical-fallacy/tree/main/Dataset-fixed	Sourati et al., 2022

	4. Fallacy of ambiguity	LOGICCLIMATE: 1109 examples			https://github.com/usc-isi-i2/logical-fallacy-identification	
Fallacy Type Recognition (11 classes)	<ol style="list-style-type: none"> 1. Cherry Picking 2. Evading the Burden of Proof 3. False Analogy 4. False Authority 5. False Cause 6. Hasty Generalization 7. No Fallacy 8. Post Hoc 9. Red Herring 10. Strawman 11. Vagueness 	newCLIMATE: 477 news	Word2Vec / transformer embeddings	T5, UnifiedQA, GPT-3, BERT	https://github.com/Tariq60/fallacy-detection/tree/master/data	Alhindi et al., 2022
Fallacy Type Recognition (11 classes)	<ol style="list-style-type: none"> 1. Cherry Picking 2. Evading the Burden of Proof 3. False Analogy 4. False Authority 5. False Cause 6. Hasty Generalization 7. No Fallacy 8. Post Hoc 9. Red Herring 10. Strawman 11. Vagueness 	COVID-19: 621 news			https://github.com/Tariq60/fallacy-detection/tree/master/data	Musi et al., 2022
Fallacy Type Recognition (6 classes)	<ol style="list-style-type: none"> 1. ad hominem 2. appeal to emotion 3. red herring 4. hasty generalization 5. irrelevant authority 6. non-fallacious 	Covid-19 news: 226 examples	BPE tokenizer	PET classifier on top of RoBERTa	The corpus will be made available via data-sharing agreement pending publication	Bonial et al., 2022

Fallacy Type Recognition (22 classes)	<ol style="list-style-type: none"> 1. Appeal to Authority 2. Appeal to Belief 3. Argument against theMan 4. Argument from force 5. Argument to pity 6. Begging the question 7. Fallacy of false cause 8. Fallacy of irrelevant purpose 9. Fallacy of Is to Ought 10. Fallacy of previous this 11. Fallacy of red herring 12. Fallacy of slippery slope 13. Fallacy of special pleading 14. Fallacy of the beard 15. False analogy 16. Irrelevant conclusion 17. Joint effect 18. Poisoning the well 19. Pragmatic Fallacy 20. Prejudicial language 21. Slothful induction 22. Wrong direction 	Logical.ly (scrapped from argueman): 916 examples			https://github.com/SeanFlanner/Logical.ly/tree/master/src/data	
Logical Fallacy Label (Yes/No)		LOGIFALLA			PENDING https://ist.psu.edu/research/projects/online-conversations	Huang, Ting Hao (Kenneth)
Fallacy Type (if Yes)						

References

- [1] T. Alhindi, T. Chakrabarty, E. Musi, and S. Muresan, “Multitask Instruction-based Prompting for Fallacy Recognition,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Apr. 2022, pp. 8172–8187. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.560> .
- [2] C. Bonial *et al.*, “The Search for Agreement on Logical Fallacy Annotation of an Infodemic,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Apr. 2022, pp. 4430–4438. [Online]. Available: <https://aclanthology.org/2022.lrec-1.471> .
- [3] I. Habernal, R. Hannemann, C. Pollak, C. Klamm, P. Pauli, and I. Gurevych, “Argotario: Computational Argumentation Meets Serious Games,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Apr. 2017, pp. 7–12. doi: 10.18653/v1/D17-2002.
- [4] Z. Jin *et al.*, “Logical Fallacy Detection,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, Apr. 2022, pp. 7180–7198. [Online]. Available: <https://aclanthology.org/2022.findings-emnlp.532> .
- [5] G. D. S. Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov, “SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles,” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Apr. 2020, pp. 1377–1414. doi: 10.18653/v1/2020.semeval-1.186.
- [6] G. D. S. Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, “Fine-Grained Analysis of Propaganda in News Article,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Apr. 2019, pp. 5636–5646. doi: 10.18653/v1/D19-1565.
- [7] E. Musi and C. Reed, “From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media,” *Discourse & Society*, vol. 33, no. 3, pp. 349–370, 2022, doi: 10.1177/09579265221076609.
- [8] P. Reiser, B. Heinzerling, N. Inoue, S. Kiyono, and K. Inui, “Riposte! A Large Corpus of Counter-Arguments,” *CoRR*, vol. abs/1910.03246, 2019, [Online]. Available: <http://arxiv.org/abs/1910.03246> .
- [9] S. Sahai, O. Balalau, and R. Horincar, “Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Apr. 2021, pp. 644–657. doi: 10.18653/v1/2021.acl-long.53.
- [10] Z. Sourati *et al.*, “Robust and explainable identification of logical fallacies in natural language arguments,” *Knowl Based Syst*, vol. 266, p. 110418, 2023, doi: <https://doi.org/10.1016/j.knosys.2023.110418> .
- [11] A. Srivastava *et al.*, “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.” Apr. 2022. doi: 10.48550/arXiv.2206.04615.

17.2 FACT-CHECKING DATASET SUMMARY

Task	Dataset	Link	Extracted features	Approach	References
Claim classification (multiclass)	<p>FEVER (Fact Extraction and VERification) consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. For the first two classes, the annotators also recorded the sentence(s) forming the necessary evidence for their judgment.</p> <p><i>Supported</i>: 93367 claims <i>Refuted</i>: 43107 claims <i>NotEnoughInfo</i>: 48971 claims</p>	https://fever.ai/dataset/fever.html	<p><i>id</i>: The ID of the claim <i>label</i>: The annotated label for the claim. Can be one of <i>SUPPORTS/REFUTES/NOT ENOUGH INFO</i>. <i>claim</i>: The text of the claim. <i>evidence</i>: A list of evidence sets (lists of [Annotation ID, Evidence ID, Wikipedia URL, sentence ID] tuples) or a [Annotation ID, Evidence ID, null, null] tuple if the label is NOT ENOUGH INFO.</p> <p>Extracted features: Term frequencies TF-IDF cosine similarity between claim and evidence</p>	MLP, Decomposable Attention (DA)	James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal, 2018
Claim classification (multiclass)	<p>The FEVER 2.0 Dataset consists of 1174 claims created by the submissions of participants in the <i>Breaker</i> phase of the 2019 shared task. Participants (<i>Breakers</i>) were tasked with generating adversarial examples that induce classification errors for the existing systems. Breakers submitted a dataset of up to 1000 instances with equal number of instances for each of the three classes (<i>Supported, Refuted NotEnoughInfo</i>). Only novel claims (i.e. not contained in the original FEVER dataset) were considered as valid entries to the shared</p>	https://fever.ai/dataset/adversarial.html	<p>Same as FEVER: Term frequencies TF-IDF cosine similarity between claim and evidence</p>	MLP, Decomposable Attention	James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal, 2019

	task. The submissions were then manually evaluated for <i>Correctness</i> (grammatical, appropriately labeled and meet the FEVER annotation guidelines requirements).				
Claim classification (multiclass)	FEVEROUS (Fact Extraction and VERification Over Unstructured and Structured information) is a fact verification dataset which consists of 87,026 verified claims. Each claim is annotated with evidence in the form of sentences and/or cells from tables in Wikipedia, as well as a label indicating whether this evidence supports, refutes, or does not provide enough information to reach a verdict. The dataset also contains annotation metadata such as annotator actions (query keywords, clicks on page, time signatures), and the type of challenge each claim poses.	https://fever.ai/dataset/feverous.html	TF-IDF, sentences	RoBERTa encoder with a linear layer on top	Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos et al., 2021
Fake News Detection (multiclass)	LIAR dataset: 12.8K manually labeled short statements from POLITIFACT.COM <i>Pants-fire</i> : 1,050 statements <i>False</i> : ~2350 statements <i>Barely-true</i> : ~2350 statements <i>Half-true</i> : ~2350 statements <i>Mostly-true</i> : ~2350 statements <i>True</i> : ~2350 statements	https://paperswithcode.com/dataset/liar	Matrix of embedding vectors	LR, SVM, Bi-LSTM, CNN	William Yang Wang, 2017
Fake News Detection (multiclass)	LIAR-PLUS dataset: extension of the LIAR dataset by automatically extracting the justification from the fact-checking article used by humans to label a given claim (metadata)	https://paperswithcode.com/paper/where-is-your-evidence-improving-fact	tf-idf weighted unigram features <i>Glove</i> word embeddings sentiment strength using <i>SentiStrength</i>	LR , SVM, Bi-LSTM, P-BiLSTM	Tariq Alhindi, Savvas Petridis, Smaranda Muresan, 2018

			emotion using the NRC Emotion Lexicon (EmoLex), and the Linguistic Inquiry and Word Count(LIWC) number of claims each speaker makes for every truth-label (history)		
Fake news classification (binary)	Dataset based on KaggleFN <i>Fake</i> : ~1320 claims from various articles <i>Real</i> : ~796 claims from various articles	https://www.kaggle.com/datasets/ruchi798/source-based-news-classification?resource=download	TF/IDF values <i>Word2Vec</i> embeddings <i>GloVe</i> embeddings	SVM, Naive Bayes, RF, LR, AdaBoost, DT, NN	Avinash Bharadwaj et al., 2020
Fake News Detection (multiclass)	data from 2013 to 2020 using BeautifulSoup library of Python from Politifact.com TODO: write labels	https://www.kaggle.com/datasets/techykajal/fake-real-news	-	-	-
Scientific claim classification (multiclass)	SciFact dataset: 1409 scientific claims accompanied by expert annotations and papers that support or contradict each claim <i>SUPPORT</i> : 556 claims <i>NOINFO</i> : 516 claims <i>CONTRADICTS</i> : 337 claims	https://www.kaggle.com/datasets/thedevastator/unlock-insight-into-scientific-claims-with-scifa?select=claims_train.csv	“claim-only” and “abstract-only” k sentences from the cited abstract with highest TF-IDF similarity to the claim Rationale sentence selection RoBERTa-large sentence encoder	NN + softmax (BERT-based model)	David Wadden et al., 2020
News classification (binary)	ISOT Fake News dataset <i>Real</i> : 21417 world and politics news from ‘Reuters’ <i>Fake</i> : 23481 government, US, left, politics and general news from ‘Wikipedia’ and ‘Politifact’	https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets	Term frequencies TF-IDF	SGD, SVM, LSVM, KNN, DT	

COVID claims (kind-of multiclass)	<p>FakeCovid dataset: 7623 claims from 92 fact-checking websites located in 105 countries (references on ‘Poynter’ and ‘Snopes’)</p> <p>False: 82%</p> <p>Unclear (e.g misleading, mostly-true): 18%</p>	<p>https://www.kaggle.com/datasets/thedevastator/fake-covid-fact-checked-news-dataset</p>	-	BERT-based model	
COVID claim classification (binary)	<p>ACOVMD COVID Infodemic dataset: 500 tweet samples from 2021</p> <p><i>True</i>: 250 Tweets from Centers for Disease Control and Prevention, World Health Organization, Indian Council of Medical Research, and Official Government of India</p> <p><i>False</i>: 250 claims from tweets, Instagram posts, Facebook posts, web-based news information, fact-checking websites namely, ‘Politifact’ and ‘Poynter’</p>	<p>https://www.kaggle.com/datasets/selvabirunda/acovmd-covid-infodemic</p>	-	-	-
Fake news classification (multiclass)	<p>FNC-1 dataset: 75385 articles (49972 training data)</p> <p>Training data: Unrelated: ~73% Discuss: ~18% Agree: ~7%</p>	<p>https://paperswithcode.com/paper/a-simple-but-tough-to-beat-baseline-for-the</p>	<p>TF vector of article headline TF vector of article body Cosine similarity between the L2-normalised TF-IDF vectors of the headline and body</p>	MLP	Benjamin Riedel et al., 2018

	Disagree: ~2%				
Claim classification (multiclass, multilingual)	X-FACT dataset: 31189 general domain non-English claims from 25 languages. Fact-checked by International Fact-Checking Network (IFCN) and Duke Reporter's Lab	https://github.com/utahnlip/x-fact	Claim-only Claim+Metadata Attention-based Evidence Aggregator (Attn-EA)	mBERT (multilingual BERT)	Ashim Gupta, Vivek Srikumar, 2021
Claim classification (multiclass)	VitaminC dataset: more than 450,000 claim-evidence pairs for fact verification and factual consistent generation. Based on over 100,000 revisions to popular Wikipedia pages, and additional "synthetic" revisions. Supports: 185714 Refutes: 131958 NotEnoughInfo: 52981	https://huggingface.co/datasets/tals/vitaminc	Average fastText word embedding	Bag of words+MLP ALBERT	Tal Schuster, Adam Fisch, Regina Barzilay, 2021
Claim checking (binary)	Snopes dataset: Fact-checking (FC) articles which contains ~16000 pairs (multimodal tweet and a FC-article) from snopes.com <i>Real</i> : 11202 positive tweets <i>Fake</i> : ~5000 fake tweets	https://github.com/nguyenvo09/EMNLP2020	Glove embeddings ELMo embeddings Glove+ELMo	CNN Ranking models: DVSH, TranSearch, ESIM, NSMN, Match-Pyramid, KNRM, ConvKNRM, CoPACRR, DUET	Nguyen Vo, Kyumin Lee, 2020



<p>Claim checking (binary)</p>	<p>Politifact dataset: Fact-checking (FC) articles which contains ~3000 pairs (multimodal tweet and a FC-article) from politifact.com</p> <p><i>Real</i>: 2037 positive tweets <i>Fake</i>: ~1000 fake tweets</p>	<p>https://github.com/nguyenvo09/EMNLP2020</p>	<p>Glove embeddings ELMo embeddings Glove+ELMo</p>	<p>CNN</p> <p>Ranking models: DVSH, TranSearch, ESIM, NSMN, Match-Pyramid, KNRM, ConvKNRM, CoPACRR, DUET</p>	<p>Nguyen Vo, Kyumin Lee, 2020</p>
<p>Claim classification (multiclass)</p>	<p>CLIMATE-FEVER dataset: 1,535 real-world claims regarding climate-change. Each claim is accompanied by five manually annotated evidence sentences retrieved from Wikipedia that support, refute or do not give enough information to validate the claim. The total dataset thus contains 7,675 claim-evidence pairs. Furthermore, the dataset features challenging claims that relate multiple facets and disputed cases of claims where both supporting and refuting evidence are present.</p> <p><i>SUPPORTS</i>: 655 claims <i>REFUTES</i>: 253 claims <i>DISPUTED</i>: 153 claims <i>NOTENOUGHINFO</i>: 474 claims</p>	<p>https://www.sustainablefinance.uzh.ch/en/research/climate-fever.html</p>	<p>Sentence embeddings</p>	<p>ALBERT (large-v2)</p>	<p>Thomas Diggelmann et al., 2020</p>
<p>Covid tweet claim</p>	<p>CoVERT dataset: tweets with a focus on the domain of biomedicine and COVID-19-related (mis)information. The corpus</p>	<p>https://www.ims.uni-stuttgart.de/forschung/re</p>	<p>Textual entailment features by AllenNLP model</p>	<p>BERT, BioBERT, MLP</p>	<p>Isabelle Mohr et al., 2022</p>

checking (multiclass)	consists of 300 tweets, each annotated with medical named entities and relations. Employs a novel crowdsourcing methodology to annotate all tweets with fact-checking labels and supporting evidence, which crowdworkers search for online. This methodology results in moderate inter-annotator agreement. <i>SUPPORTS</i> : 198 claims <i>REFUTES</i> : 66 claims <i>NotEnoughInfo</i> : 36 claims	ssourcen/korpora/bioclim/			
Claim classification (multiclass)	Moche dataset: consists of 21,184 claims, where each claim is assigned a truthfulness label and ruling statement, with 58,523 pieces of evidence in the form of text and images (fact-checked from 'Snopes' and 'Politifact'). It supports the end-to-end multimodal fact-checking and explanation generation, where the input is a claim and a large collection of web sources, including articles, images, videos, and tweets, and the goal is to assess the truthfulness of the claim by retrieving relevant evidence and predicting a truthfulness label (i.e., support, refute and not enough information), and generate a rationalization statement to explain the reasoning and ruling process.	https://zenodo.org/record/6657096#.ZD60YI5Byhc	Claim and evidence word representations	CLIP model for contextual representations, BERT-based, BART, SBERT, NN (Stance detection layers, sentiment detection layers)	Barry Menglong Yao et al., 2022

	<p><i>SUPPORTS</i>: 7803 claims <i>REFUTES</i>: 8004 claims <i>NEI</i>: 5377 claims</p>				
Tweet claim classification (multiclass)	<p>Stanceosaurus dataset: 28,033 tweets in English, Hindi, and Arabic annotated with stance towards 251 misinformation claims. The claims in Stanceosaurus originate from 15 fact-checking sources that cover diverse geographical regions and cultures. Unlike existing stance datasets, it introduces a more fine-grained 5-class labeling strategy with additional subcategories to distinguish implicit stance.</p> <p>English subset</p> <p>-----</p> <p>5-class mode: Irrelevant: 7885 tweets Supporting: 2549 tweets <i>Refuting</i>: 1090 tweets Discussing: 8884 tweets</p> <p>3-subclass mode: Supporting: 4302 tweets <i>Refuting</i>: 2735 tweets <i>Other</i>: 1847 tweets</p>	<p>https://github.com/JonathanQZheng/Stanceosaurus/</p>	Claim-tweet tokens	BERT, BERTweet, mBERT, XLM-RoBERTa	Jonathan Zheng et al., 2022
Factual claim checking	<p>MultiFC dataset: naturally occurring factual claims for the purpose of automatic claim verification. It is collected from 26 fact checking websites in English, paired with textual sources</p>	<p>http://www.copenlu.com/publication/2019_emnlp_augenstein/</p>			Isabelle Augenstein et al., 2019

	and rich metadata, and labelled for veracity by human expert journalists (NEEDS LOGIN TO ACCESS DATASET)				
COVID claim classification (binary)	COVID-Fact dataset: a FEVER-like dataset of claims concerning the COVID-19 pandemic. The dataset contains claims, evidence for the claims, and contradictory claims refuted by the evidence. <i>Supported: 1296 claims</i> <i>Refuted: 2790 claims</i>	https://github.com/asaakyan/covidfact	Sentence embeddings	SBERT, RoBERTa	Arkadiy Saakyan et al., 2021
Scientific news classification	Spiced dataset: paraphrase dataset of scientific findings annotated for degree of information change. Spiced contains 6,000 scientific finding pairs extracted from news stories, social media discussions, and full texts of original papers. Information Matching Score (IMS) classification (matching score: 1-5)	https://huggingface.co/datasets/copenlu/spiced	Sentence embeddings	RoBERTa, SBERT with MiniLM, MPNet-FT, SciBERT, CiteBERT	Dustin Wright et al., 2022

References

- [1] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: a Large-scale Dataset for Fact Extraction and VERification,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, Apr. 2018, pp. 809–819. doi: 10.18653/v1/N18-1074.
- [2] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, “The FEVER2.0 Shared Task,” in Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Apr. 2019, pp. 1–6. doi: 10.18653/v1/D19-6601.

- [3] R. Aly et al., “FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information.” 2021.
- [4] T. Alhindi, S. Petridis, and S. Muresan, “Where is Your Evidence: Improving Fact-checking by Justification Modeling,” in Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Apr. 2018, pp. 85–90. doi: 10.18653/v1/W18-5513.
- [5] W. Y. Wang, ““Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Apr. 2017, pp. 422–426. doi: 10.18653/v1/P17-2067.
- [6] A. Bharadwaj, B. Ashar, P. Barbhaya, R. Bhatia, and Z. S. Shaikh, “Source Based Fake News Classification using Machine Learning,” 2020.
- [9] D. Wadden et al., “Fact or Fiction: Verifying Scientific Claims,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Apr. 2020, pp. 7534–7550. doi: 10.18653/v1/2020.emnlp-main.609.
- [10] I. and S. S. A. H. and Traore, “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques,” in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, I. and A. A. T. I. and Woungang, Ed., Springer International Publishing, 2017, pp. 127–138.
- [11] G. K. Shahi and D. Nandini, FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19. ICWSM, 2020. doi: 10.36190/2020.14.
- [12] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, “A simple but tough-to-beat baseline for the Fake News Challenge stance detection task.” 2018.
- [13] A. Gupta and V. Srikumar, “X-Fact: A New Benchmark Dataset for Multilingual Fact Checking,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Apr. 2021, pp. 675–682. doi: 10.18653/v1/2021.acl-short.86.
- [14] T. Schuster, A. Fisch, and R. Barzilay, “Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Apr. 2021, pp. 624–643. doi: 10.18653/v1/2021.naacl-main.52.
- [15] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, “CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims.” 2021.
- [16] N. Vo and K. Lee, “Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Apr. 2020, pp. 7717–7731. doi: 10.18653/v1/2020.emnlp-main.621.
- [17] I. Mohr, A. Wüthrich, and R. Klinger, “CoVERT: A Corpus of Fact-checked Biomedical COVID-19 Tweets.” 2022.
- [18] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, and L. Huang, “End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models.” 2022.
- [19] J. Zheng, A. Baheti, T. Naous, W. Xu, and A. Ritter, “Stanceosaurus: Classifying Stance Towards Multicultural Misinformation,” in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Apr. 2022, pp. 2132–2151. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.138>
- [20] I. Augenstein et al., “MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Apr. 2019, pp. 4685–4697. doi: 10.18653/v1/D19-1475.
- [21] A. Saakyan, T. Chakrabarty, and S. Muresan, “COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Apr. 2021, pp. 2116–2129. doi: 10.18653/v1/2021.acl-long.165.
- [22] D. Wright, J. Pei, D. Jurgens, and I. Augenstein, “Modeling Information Change in Science Communication with Semantically Matched Paraphrases.” 2022.

Table 17.3. HATE SPEECH DATASETS SUMMARY

Task	Labels	Dataset Type & Size	Language	Dataset Type	Dataset Links	Extracted features	Approach	References
Clickbait classification (binary)	Non-clickbait, Clickbait	Non-clickbait: 18.513 Wikinews articles collected by NewsReader Clickbait: 7.623 articles from BuzzFeed, Upworthy, ViralNova, Scoopwhoop, ViralStories	English	Web articles	https://github.com/bhargaviparaniapa/clickbait https://www.kaggle.com/datasets/amanandrai/clickbait-dataset?resource=download	Sentence structure Word patterns Clickbait language N-gram, POS N-grams	SVM, Decision Tree, RF	Chakraborty et al., 2016
Hate speech classification (3 classes)	Hate Speech, Offensive, Neither	24.783 unique tweets	English	Tweets	https://github.com/t-davidson/hate-speech-and-offensive-language	TF-IDF; Features for the number of characters, words, and syllables	Logistic regression, Naive Bayes, Decision trees, Random Forest, SVM	Davidson, et al., 2017
Hate speech classification (4 classes)	Hate Abusive Spam None	80.000 entries 0.18% abusive	English	Tweets	https://dataverse.mpi-sws.org/dataset.xhtml?persistentId=doi:10.5072/FK2/ZDTEMN	TF-IDF, Glove, Transformers	BERT, ELECTRA, CNN, MLP	Founta et al. 2018
Hate speech classification (binary)	Not Racist/Not Sexist (0),	Twitter Sentiment Analysis (TSA)	English	Tweets	https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech	TF-IDF, Glove,	BERT, ELECTRA, CNN, MLP	

	Racist/Sexist (1)	31.961 entries by Analytics Vidhya 0.07% abusive				Transformers		
Offensive language classification (binary, 3 classes)	Sub-task A: Not Offensive (NOT) Offensive (OFF) Sub-task B: Targeted Insult (TIN) Untargeted (UNT) Sub-task C: Individual (IND) Group (GRP) Other (OTH)	Offensive Language Identification Dataset (OLID) Official dataset of SemEval 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) 13.240 unique entries on the train & validation set 4.747 unique entries on the test set A: Offensive Language Detection B: Categorization of Offensive Language C: Offensive Language Target Identification	English	Tweets	https://scholar.harvard.edu/malmasi/olid https://competitions.codalab.org/competitions/20011	Unigrams, FastText embeddings	SVM, BiLSTM, CNN, RNN	Zampieri et al., 2019

Toxic Language Detection Text Classification (binary)	TOXIC, NOT	Toxic Tweets Dataset from Kaggle 53.619 unique entries	English	Tweets	https://www.kaggle.com/datasets/ashwinier176/toxic-tweets-dataset			
Offensive language classification (binary, 3 classes)	Sub-task A: Not Offensive (NOT) Offensive (OFF) Sub-task B: Targeted Insult (TIN) Untargeted (UNT) Sub-task C: Individual (IND) Group (GRP) Other (OTH)	SemEval-2019 Task 6 (OffensEval) 13.181 unique entries	English	Tweets	https://github.com/cardiffnlp/tweeteval		RNN, LSTM, BERT, BiLSTM, BiGRU, Ensemble	Zampieri et al., 2019
Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (binary)	Hateful (HATE), Non-hateful (NOT)	SemEval-2019 Task 5 (HatEval) 997.2 unique entries on the train & validation set 2.924 unique entries on the test set	English	Tweets	https://github.com/cardiffnlp/tweeteval	TF-IDF, Glove embeddings, different types of n-grams	MFC, SVM, CNN, LSTM, BiGRU, BERT, Logistic Regression, Multinomial Naive Bayes	Basile et al., 2019
Hate Speech Detection (binary, 7 sub-)	Hate, Not Hate	HateCheck: Functional Tests for Hate Speech Detection Models	English	Posts	https://github.com/paul-rottger/hatecheck-data		BERT, Google Jigsaw's	Rottger et al., 2021

classes within Hate)	7 Targets Within Hate: (Women, Trans people, Black people, Gay people, Disabled people, Muslims, Immigrants)	A checklist of functional tests to evaluate hate speech detection models. 3.728 synthetically-generated texts 68% abusive					Perspective (P), Two Hat's SiftNinja (SN)	
Toxic Spans Detection (binary)	Toxic, non-toxic	Semeval-2021 Task 5: Toxic Spans Detection 10.629 entries 0.56% abusive	English	Posts of civil comments	https://github.com/ipavlopoulos/toxic-spans	FLAIR, FastText embeddings	Transformers (RoBERTa, BERT), LSTM	Pavlopoulos et al., 2021
Offensive Speech and Stance Detection (binary)	Hate / Offensive, Neither	Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection Data collected by supporters of Trump or Biden 3.000 entries 12% abusive	English	Tweets	https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/stance-hof/			Grimminger et al., 2020
Hate Speech Classification (binary)	Hate, noHate	Hate Speech Dataset from Stormfront, a	English	Forum posts	https://github.com/Vicomtech/hate-speech-dataset	Unigrams, bi-grams	SVM, CNN, RNN with LSTM	Gibert et al., 2018

		White Supremacy Forum 10.944 entries 0.11% abusive Sentences by taking into account the context						
Hate Speech Classification (binary)	Hateful, Not	A Benchmark Dataset for Learning to Intervene in Online Hate Speech 33.776 entries from Gab, 0.43% abusive 22.324 entries from Reddit, 0.24% abusive	English	Gab posts Reddit posts	https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech	TF-IDF, 2-grams, Randomly initialized word embeddings, Word2Vec embeddings on Google News	Logistic Regression, SVM, CNN. RNN with BiGRU	Qian et al., 2019
Hate Speech Classification (binary, 8 classes)	Hate, Not Multi-label: violence (1) or not violence (0), directed (1) or general (0), 6 labels about the category of hate	ETHOS: an Online Hate Speech Detection Dataset 998 entries, 565 not hate speech, 433 hate speech 433 entries with hate speech content 0.43% abusive	English	YouTube and Reddit comments	https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset		Multinomial, BernoulliNB, Logistic Regression, SVM, Random Forests, Gradient Boosting,	Mollas et al., 2022

	speech: gender, race, national_ori gin, disability, religion and sexual_orienta tion						CNN+Atten tion+FT+GV , LSTM+FT+G V, FF+LSTM+C NN+FT+GV, BiLSTM+FT +GV, BERT, DistilBERT	
Hate Speech Classification (binary, multinomial)	Vulgarity and/or Offensive language (VO), assaults on Human Dignity (HD), Calls for violence (CV), Sub-task on message delivery: <i>explicit/implici t</i> , Multinomial classification: Identity-based hate (race/ethnicity	The Gab Hate Corpus: A collection of 27,665 posts annotated for hate speech 0.09 Hate, 0.06 Offensive/Vulgar	Englis h	Gab posts	https://osf.io/edua3/	Frequency of Linguistic Inquiry and Word Count (LIWC), TF-IDF 73- length vectors	Linear SVMs, BERT	Kennedy et al., 2018

	, nationality/regionalism/xenophobia, gender, religion/belief system, sexual orientation, ideology, political identification/party, mental/physical health)							
Group-directed, Person-directed Hate speech classification (binary, 21 classes)	Hateful, Neutral Multinomial classification (21 categories divided into: 1. hateful language (Accusations, Humiliation, Swearing, Promoting Violence) , 2.hate targets (Financial Power, Political Issues, Racism & Xenophobia, Religion,	Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media 5. 143 entries 82% abusive	English	YouTube and Facebook comments	https://www.dropbox.com/s/21wtzy9arc5skr8/ICWSM18%20-%20SALMINEN%20ET%20AL.xlsx?dl=0	1-3 N-grams, TF-IDF, hateful/no n-hateful words, word2vec, doc2vec embeddings	Logistic Regression, Decision Tree, Random Forest, Adaboost, SVM	Salminen et al., 2018

	Specific Nation(s), Media, Armed Forces, Behavior) 3.hate sub-targets (Corporation, Wealthy, Terrorism, Politics, Ideology, Anti-white, Anti-black, Xenophobia, Anti-Islam, Anti-Semitist, Towards media company, Other, Police, Military, Humanity, Other)							
Offensive Language Detection (binary)	Not Offensive (NOT), Offensive (OFF)	Greek Tweet Dataset (OGTD) 4.749 entries 0.29% abusive	Greek	Tweets	https://opendatalab.com/OGTD	TF/IDF unigrams, TF/IDF bigrams, POS tags, dependency relation tags,	Linear SVM, RBF SVM, Stochastic Gradient Descent, Multinomial Naïve Bayes, Bernoulli Naïve Bayes,	Pitenis et al., 2020

							Stacked LSTM with Attention, LSTM and GRU with Attention, 2D Convolution with Pooling, GRU with Capsule, LSTM with Capsule and Attention, Multi-lingual BERT	
Offensive language classification (binary, 3 classes)	Sub-task A: Not Offensive (NOT) Offensive (OFF) Sub-task B: Targeted Insult (TIN) Untargeted (UNT) Sub-task C: Individual (IND) Group (GRP) Other (OTH)	Offensive Language and Hate Speech Detection for Danish 3.600 entries 0.12% abusive	Danish	Facebook - Ekstra Bladet, Reddit, r/Denmark, r/DANMAG		Word2vec, comment2vec, randomly initialized embeddings	Logistic Regression, Learned-BiLSTM, FastBiLSTM, AUX-Fast-BiLSTM	Sigurbergsson et al., 2019



Abusive Speech detection (binary)	True (Abusive), False (Not abusive)	Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Multilingual, Abusive Words, Political 2,000 entries 32.7% abusive	Surzhuk (Russian & Ukrainian)	YouTube comments	https://github.com/bohdan1/AbusiveLanguageDataset	5 most top-used abusive words 600 abusive and obscene words (bad_words.txt)	SD – seed dictionary RD – relative distance LOR – log odds ratio STM – stemmed dictionary MSD – micro seed dictionary Words – number of new abusive terms added to the seed dictionary	Andrusyak et al., 2018
Flagged content detection (binary)	Deleted, Not	Latvian newspaper user comment dataset 1.421.883 comments from 2019	Latvian & Russian	Newspaper comments	https://www.clarin.si/repository/xmlui/handle/11356/1407	BERT embeddings	EMBEDDIA Media Assistant with the TEXTA Toolkit, multilingual BERT	Pollak et al., 2021



Multi-topic harassment detection	Racism, Sexism, Appearance-related, Intellectual, Political	A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research 24.189 entries 0.13% abusive	English	Tweets	https://github.com/Mrezvan94/Harassment-Corpus			Rezvan et al., 2018
Abusive Keywords	PS:negative stereotypes ethnic slurs RCI:locations and demonyms PA:professions and occupations DDF:physical disabilities and diversity DDP:cognitive disabilities and diversity DMC:moral and behavioral defects IS:words related to social and economic disadvantage OR:plants AN:animals ASM:male genitalia	HurtLex is a lexicon of offensive, aggressive, and hateful words in over 50 languages. The words are divided into 17 categories, plus a macro-category indicating whether there is a stereotype involved.	Over 50 languages	Abusive words & phrases	https://github.com/valeriobasile/hurtlex	Lemmas, POS tags, category, level		Bassignana et al., 2018

	ASF:female genitalia PR:words related to prostitution OM:words related to homosexuality QAS:with potential negative connotations CDS:derogator y words RE:felonies and words related to crime and immoral behavior SVP:words related to the seven deadly sins of the Christian tradition							
--	---	--	--	--	--	--	--	--

References

- [1] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop Clickbait: Detecting and preventing clickbaits in online news media," in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, Institute of Electrical and Electronics Engineers Inc., Nov. 2016, pp. 9–16. doi: 10.1109/ASONAM.2016.7752207.
- [2] T. Davidson, D. Warmley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," 2017. [Online]. Available: www.aaai.org.
- [3] A.-M. Founta *et al.*, "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior," Feb. 2018, [Online]. Available: <http://arxiv.org/abs/1802.00393>.

- [4] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval).” [Online]. Available: <http://competitions.codalab.org/>
- [5] V. Basile *et al.*, “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter.” [Online]. Available: <http://evalita.org>.
- [6] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. B. Pierrehumbert, “HateCheck: Functional Tests for Hate Speech Detection Models,” Dec. 2020, doi: 10.18653/v1/2021.acl-long.4.
- [7] J. Pavlopoulos, J. Sorensen, L. Laugier, and I. Androutsopoulos, “SemEval-2021 Task 5: Toxic Spans Detection.” [Online]. Available: <https://aclanthology.org/2021.semeval-1.6/>.
- [8] L. Grimminger and R. Klinger, “Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection,” 2021. [Online]. Available: <https://www.ims.uni-stuttgart.de/>.
- [9] O. De Gibert, N. Perez, A. García-Pablos, and M. Cuadros, “Hate Speech Dataset from a White Supremacy Forum,” 2018. [Online]. Available: <https://aclanthology.org/W18-5102/>.
- [10] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, “A Benchmark Dataset for Learning to Intervene in Online Hate Speech,” Sep. 2019, [Online]. Available: <http://arxiv.org/abs/1909.04251>.
- [11] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, “ETHOS: an Online Hate Speech Detection Dataset,” Jun. 2020, doi: 10.1007/s40747-021-00608-2.
- [12] B. Kennedy *et al.*, “Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale,” *Lang Resour Eval*, vol. 56, no. 1, pp. 79–108, Mar. 2022, doi: 10.1007/s10579-021-09569-x.
- [13] J. Salminen *et al.*, “Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media,” 2018. [Online]. Available: www.aaai.org.
- [14] Z. Pitenis, M. Zampieri, and T. Ranasinghe, “Offensive Language Identification in Greek,” Mar. 2020, [Online]. Available: <http://arxiv.org/abs/2003.07459>
- [15] G. I. Sigurbergsson and L. Derczynski, “Offensive Language and Hate Speech Detection for Danish.” [Online]. Available: <https://www.reddit.com/r/Denmark/comments/>.
- [16] B. Andrusyak, M. Rimel, and R. Kern, “Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Languages,” 2018.
- [17] S. Pollak Jožef Stefan Institute *et al.*, “Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions,” 2021. [Online]. Available: <https://docs.texta.ee/>.
- [18] M. Rezvan, K. Thirunarayan, S. Shekarpour, V. L. Shalin, L. Balasuriya, and A. Sheth, “A quality type-aware annotated corpus and lexicon for harassment research,” in *WebSci 2018 - Proceedings of the 10th ACM Conference on Web Science*, Association for Computing Machinery, Inc, May 2018, pp. 33–36. doi: 10.1145/3201064.3201103.
- [19] E. Bassignana, V. Basile, and V. Patti, “Hurtlex: A Multilingual Lexicon of Words to Hurt.” [Online]. Available: <http://hatespeech.di.unito.it/>.

18 ANNEX 2: CONCLUSION LIST OF USER REQUIREMENTS

This section presents the culmination of TITAN’s extensive process, incorporating all relevant considerations mentioned throughout Part 2 and actively involving users in the development of the chatbot. After considering the insights from literature reviews, co-creation workshops, stakeholder involvement, legal and ethical analysis²³, user perspectives and technical constraints, we have compiled a complete²⁴ list of user requirements (see Table 18.1 below). These requirements serve as a conclusion, guiding the further refinement and development of the chatbot to effectively combat disinformation while addressing the needs and expectations of users.

#	ID	Requirement	Rationale	Category
1	BB-1	The chatbot will need to be explicit about its premise i.e., an informed citizenship.	The chatbot should provide information about the concept of an informed populace, highlighting the importance of logical thinking, information-seeking, and political involvement among citizens in a democracy.	Non-Functional
2	BB-2	The chatbot will need to be explicit about its premise i.e., the significance of critical thinking in media literacy education.	The chatbot should explain how questioning sources, considering credibility, and understanding own biases are essential aspects of evaluating information.	Non-Functional
3	BB -3	The user may need to get a trust indicator for sources.	The chatbot should give an indication of trustworthiness of sources e.g., by stating the trust scores of brands in the Digital News Report.	Non-Functional
4	BB -4	The chatbot will need to invite users to define their level of trust in provided sources.	When providing an indication of trustworthiness, the chatbot should invite the user to make their own assessment of trust in the source.	Non-Functional
5	BB -5	The user may need to be challenged regarding the ways in which they assign trust to sources	The chatbot should invite users to reflect on the assumptions underlying how they assign trust to sources, specifically helping them reflect on how emotional attachment, habits, and cognitive biases can overrule reliability and accuracy.	Functional
6	BB -6	The user may need to be made aware of the importance of emotions in how they deal with information.	The chatbot should highlight when the user is responding to emotions rather than ideas or facts. It should then prompt the user to reflect on how these non-rational factors may influence their trust in news and	Functional

²³ As a reminder to the reader a comprehensive analysis of the legal and ethical aspects and the corresponding user needs can be found in D1.3 and thus are omitted from the conclusion list of user requirements.

²⁴ A more comprehensive version of the user requirements can be found in [this excel](#).

			assessment of information. It should provide users with tools to distinguish between emotional appeals and evidence-based arguments and then with strategies to make more informed, balanced, and objective judgments.	
7	BB -7	When presented with facts that are potentially discomfoting or disturbing, the user may need to get a warning or accompanying text.	The chatbot will need to consider the impact of disturbing or discomfoting factual information for the user e.g., confrontational facts about war, famine, global warming, injustice, etc. – ex ante.	Functional
8	BB -8	When presented with facts that are potentially discomfoting or disturbing, the user may need to be invited to share their concerns or fears.	The chatbot will need to offer guidance through dialogue that can help the user cope with discomfoting or disturbing facts – post hoc.	Functional
9	BB -9	The user may be experiencing how critical thinking about news sources leads to increased levels of distrust in those sources.	The chatbot should help users understand the balance between trust and scepticism in their engagement with information sources.	Functional
10	BB-10	The chatbot may need to inform the user about the potential weaponization of critical thinking.	When identifying an overtly critical or biased form of reasoning by the user, the chatbot should explain how critical thinking can be exploited by anti-democratic forces, by giving examples about the manipulation of information or the promotion of conspiracy theories. It should raise awareness of the dangers of falling into ideological echo chambers or being influenced by propaganda disguised as critical thinking.	Non-Functional
11	BB-11	The chatbot should explore strategies or approaches to encourage critical thinking while maintaining trust in reliable news sources.	The chatbot should provide guidance on how citizens can navigate the complex interplay between trust and critical thinking in their news consumption habits. It should attend the user on how personal beliefs, political affiliations, religious beliefs, or following their gut feeling affects trust in media and scientific sources. It should prompt users to reflect on how this dynamic can inadvertently steer individuals towards untruthful sources aligned with their existing beliefs.	Functional

12	BB-12	The user may need to be attended to the role of biases and worldviews in critical thinking.	The chatbot will need to address how individuals' personal biases and worldviews influence their perception of what is right and wrong, true and false. It should highlight the possibility that critical thinking can be influenced more by personal ways of knowing rather than specific sources of information. When pointing out the biases in the user's reasoning, the chatbot should do so in a constructive and considerate manner. It can provide examples, such as alternative media, populist politicians, and fearmongers, who have mastered ways to gain people's trust despite disseminating inaccurate information.	Functional
13	BB-13	The chatbot should facilitate users' reflection on their own dispositions and biases.	The chatbot should prompt users to consider how their beliefs and preconceived notions may impact their evaluation of information. It should encourage users to be aware of their own cognitive biases and strive for balanced and objective assessments.	Functional
14	BB-14	When critical thinking is leading the user to dismiss factual information, the user may need to be encouraged to build its argument on factual sources.	The chatbot will need to be able to converse with a user that does not consider factual information that is offered to them. The conversation will need to invite the user to reflect on why they dismiss the facts – continue dialogue.	Functional
15	BB-15	When critical thinking continues to lead the user to dismiss factual information or to distrust the chatbot, the chatbot should explain that critical thinking ends where distrust begins.	The chatbot will need to confront the user with the conundrum of trust and critical thinking, explaining that a level of scepticism is encouraged but too much hinders Socratic/democratic dialogue.	Functional
16	BB-16	When critical thinking is consistently leading the user to dismiss factual information, the chatbot must stop the conversation.	When a user repeatedly dismisses factual information, the chatbot will need to stop the discussion and (re)state its underlying premise that factual information forms the basis of democratic dialogue.	Functional
17	UC1-1	As a user I want to be informed about the purpose of the tool	The users need a tutorial that says what they can expect.	Non-Functional

18	UC1-2	Users need info in Is informed about the duration of one session (conversation)	Expectation management is important, users need info beforehand so that they can decide when to use it (en route, in a workshop...)	Functional
19	UC1-3	Upload new articles (URL) to TITAN-tool	If we want retention, the content on the platform should be updated	Functional
20	UC1-4	Highlight text as part of micro-lesson	Interactivity is key for learning, a feature that makes it possible to work with the article, makes it more tangible	Functional
21	UC1-5	Can start a session (conversation), pause it and continue later	Users need some agency, flexibility when using the tool. The tool should be able to remember previous sessions.	Functional
22	UC1-6	Users can consume the micro-lessons at any time, and not only during a conversation	Micro-lessons are needed when the conversation is ongoing, but these interesting lessons should also be available when users want to learn about something, at any given time	Functional
23	UC1-7	Users can filter articles based on disinformation signals, so they can exercise only on specific signals	Users can randomly choose articles from the platform or upload an article, but at some point, they want to choose what to want to learn	Functional
24	UC1-8	Users can flag possible mistakes, made by the system	Users can indicate/flag when they have doubts about the detected disinformation signals	Functional
25	UC1-9	Users expect recommendations of other relevant articles they can analyze (to start a dialogue with)	Users do not always know what to do next, so we need to give relevant next articles (e.g., repeat previous disinformation signal or new information signal)	Functional
26	UC1-10	Users need rewards/achievements when finishing a conversation	We need to keep the learner motivated, create positive reinforcements	Functional
27	UC1-11	Users want to share something they have learned.	When a user has learned something, he wants to share this new knowledge with his peers. A share-button could be an easy feature	Functional
28	UC1-12	Users need encouragements	We should also incorporate positive encouragements during the conversations, to make it lighter	Functional
29	UC1-13	After a conversation the user receives concrete tips & tricks to factcheck information by themselves	Providing users with concrete tips and tricks to fact-check information by themselves after a conversation is instrumental in empowering them with practical skills and promoting independent critical thinking. These tips and tricks can include reliable fact-checking sources, critical evaluation techniques, or strategies	Functional

			for verifying information. By equipping users with these tools, the chatbot encourages self-reliance and enables users to become more discerning consumers of information in their daily lives. This feature enhances the long-term impact of the conversation, enabling users to apply fact-checking principles beyond the chatbot interaction and fostering a more informed and vigilant approach to combating disinformation.	
30	UC1-14	As a journalist I want to manually add articles to the sub-site via the CMS	In the use case the facilitators want to curate articles beforehand, so the students can exercise on well-chosen articles	Functional
31	UC1-15	The selection of articles needs to be accessible by the students	During the workshop the students can go to a website where they can explore different articles and choose which they want to tackle	Functional
32	UC1-16	As a journalist I need to have the ability to label an article as fake (wrong) or real (wright) in the CMS	In our workshop the students are detectives and need to figure out which articles are fake or not. The Titan-system is not qualified to do this, but we want the ability to do this manually	Functional
33	UC1-17	As a journalist I can change manually the detected signals of disinformation & arguments, so that I can correct mistakes the system made	In the editor the facilitator can correct the disinformation signals, so the right conversation and micro-lessons can be assigned	Functional
34	UC1-18	As a journalist I can moderate the incoming articles (delete, on hold, approve...), so that content is adapted to target audience	The students can upload articles to the platform that is also open to the other students. As facilitator we want to have the ability to pre-moderate the incoming articles.	Functional
35	UC1-19	As a journalist I want to see what students share to the platform (trends of topics), so that I get insights in the lifeworld of youngsters	As facilitator we want to have an overview of the participants. What is the average score (CTA and other data points), but also what are trends (topics) of the uploaded articles?	Functional
36	UC2-1	As a user, I want the tool to adapt and personalize my learning path based on my progress and preferences.	Provide recommendations for content based on my engagement and performance	Functional
37	UC2-2	As a user, I want the ability to collaborate (in real time) with other NGO experts on analyzing articles and signals.	Enable group discussions and shared analysis.	Functional

38	UC2-3	As a user, I want to customize notification settings to receive updates on new content, discussions, and achievements.	Allow me to choose the frequency and type of notifications	Functional
39	UC2-4	As a user, I want clear information about how my data is used and shared within the tool.	Provide transparent data handling practices and options to control data sharing	Non-Functional
40	UC2-5	As a user, I want the learning experience to include gamified elements, such as challenges, rewards, and leaderboards.	Enhance engagement and motivation through gamification.	Functional
41	UC2-6	As a user, I want the option to download content for offline access and learning.	Allow me to continue learning even without an internet connection.	Functional
42	UC2-7	As a user, I want the ability to easily share interesting content or achievements on social media platforms.	Facilitate sharing of learning experiences with a wider audience.	Functional
43	UC2-8	As a user, I want the ability to create and share content, such as articles or analyses, with the community.	Foster user-generated content to enrich the learning experience	Functional
44	UC2-9	As a user, I want the ability to provide feedback on the tool's functionality and content.	Enable users to contribute to the tool's ongoing improvement	Functional
45	UC3-1	Dialogue presentation should follow a mixed text-image approach.	Visual scenes are required to facilitate contents' and dialogues' accessibility	Functional
46	UC3-2	Dialogues should integrate standard pictograms to text	Pictograms are required to support illiterate migrants understanding	Functional
47	UC3-3	Dialogues should be presented in multiple languages	Migrants' and refugees' engagement should be assured by a multi-language approach (including host countries languages and as many mother languages as possible).	Functional
48	UC3-4	Purpose of the app should be well introduced in a video	App scope communication is required to promote acceptance of the tool.	Non-Functional
49	UC3-5	The key features/ services of the app should be demonstrated through a video or animated tutorial	App key features understanding is required to promote familiarisation with the tool.	Non-Functional

50	UC3-6	Dialogue interactions should be around 1-2 minutes each	Attention span in social media users and specifically in young generation is short; to be attractive and engaging we cannot imagine having long conversations (despite them being or not Socratic). Even if the dialogue itself could be structured on a wide conversation tree, interactions should be managed as a set of “episodes”	Functional
51	UC3-7	Interaction modalities should be flexible	For example, in a dialogue users should have the opportunity to pick from a set of options (multiple choice) leading to different dialogue branches according to the answer of each user. Ideally: in some case managing open answers from the users.	Functional
52	UC3-8	Dialogues should provide a results/ conclusion to the user	Users, not only migrants, should capture the immediate pragmatic value of the tool: not only because they are in an educational setting, but also because they might get a reward/ tangible benefit (e.g., a badge with possibly gamification logics).	Functional
53	UC3-9	The app should motivate the users to get engaged in several dialogues in order to assure critical thinking skills development	Critical thinking skills development is assured by the involvement of the users in a series of interactions with the Titan tool through time. Users’ motivation is one of the fundamental goals the app needs to accomplish. The design of a gamification strategy might be considered.	Functional
54	UC3-10	App interface and interaction design should assure to meet usability, accessibility and universal design principles	UC3 will involve users from a variety of EU and non-EU countries, with a wide spectrum of linguistic, cultural and technological literacy level. Designing interactions with as much broad literacy issue as possible is also required to fight disinformation in vulnerable populations.	Functional
55	CC-1	The tool’s focal point must be the user’s interest.	The user’s interest must be the focal point of the tool. Therefore, it is not sufficient to avoid doing evil, it must be an ambition to do good.	Non-Functional
56	CC-2	The user should only receive information that is necessary for decision making.	The information given to the user, for them to give consent, should be limited to avoid information overload. The user should know the information that makes a difference	Functional

			to their decision to give consent or not.	
57	CC-3	The tool must create transparency by ensuring that users have an accurate understanding of the necessary information.	Information does not equal transparency, instead transparency is created when the information given to the user gives them an accurate understanding of the information.	Non-Functional
58	CC-4	The user must be able to exercise their data rights.	The user must be able to conveniently exercise their data rights, by making it easy to withdraw consent at any time and enable the users to change their mind on what they consent to while using the tool.	Functional
59	CC-5	The tool must be able to explain the purpose of a personalized profile.	When offering a personalized profile to the user, the tool must be able to explain how personalization can assist the user in identifying disinformation.	Non-Functional
60	CC-6	The tool should offer granularity of consent to the user.	The user should have freedom of choice - consent should not be an all-or-nothing option. Users need to have different options to choose from.	Functional
61	CC-7	The tool should only ask the user to provide the data necessary for the service.	The user should be the center of attention for data use. Therefore, the TITAN tool should not ask for data that is merely nice to have, but only data that is necessary to have to offer the service.	Non-Functional
62	CC-8	The tool should delete user data as soon as it is not needed anymore.	To be transparent about data use, the tool should by default delete user data when it is no longer needed for the purpose the user was asked to provide it for.	Non-Functional
63	CC-9	The tool must only use data for the purpose explained to the user.	The tool must consider that many users do not have the capabilities to understand technical explanation about data use and functionalities of the system, but the user must be explained the context and what they can gain from sharing their data.	Non-Functional
64	CC-10	The tools must have values that build trust towards the user.	The tool must define its values and communicate them to the user. The values must not change over time.	Non-Functional
65	CC-11	The user must have trust in the data handling and storage.	Users need to trust that when they give a tool their data, the tool delivers something of value to them. There is no model or standard for how to create trust in a digital tool, but various steps can enable trust.	Functional

66	CC-12	The tool should make it optional for the user to make a personalized profile.	A personalized profile should not be mandatory for using the tool. Users will have different ways of using the tool.	Functional
67	CC-13	The tool could be a website platform with mobile first design.	With mobile-first design on a website you have a mobile friendly website that can also be used on bigger screens.	Functional
68	CC-14	The tool must have micro-lessons with different levels.	The micro-lessons must have different levels to start from since users will have different levels of critical thinking when they begin using the tool. If all users start at the same level, some users will not find the tool beneficial.	Functional
69	CC-15	The tool should have micro-lessons with various formats.	The formats of the micro-lessons should encourage reflection using e.g., exercises with informational videos and audio, and interactive quizzes.	Functional
70	CC-16	The tool should fulfil its purpose while requiring as little time as possible from the user.	The tool should balance the purpose of the tool and the expectation to the user's time engagement, to not create a barrier for people to use the tool.	Functional
71	CC-17	The tool must focus on formative assessment when giving feedback to the user.	Because the tool focuses on competencies of critical thinking, it should use formative assessment when giving feedback to the user.	Functional
72	CC-18	The tool could have gamification elements that fit the vision and the purpose of the tool.	Gamification elements could especially help to get young people to engage with the tool, because it gives incentive to keep learning, give immediate feedback, create competition elements, and motivational nudges.	Functional
73	CC-19	The tool should have inclusive push notifications methods.	The tool should apply inclusive push notification methods, by allowing the user to decide if they want notifications or not.	Functional
74	CC-20	The tool must consider disabilities among users in the design of the tool.	The tool must consider web accessibility to accommodate users with various disabilities.	Functional
75	CC-21	The tool could include a dialogue between users.	Group dialogue and collaboration or competition with other users could be an amplifier for people to engage with the tool.	Functional
76	CC-22	The tool must be able to consider personal differences.	Ensuring that the tool considers personal differences in guidance and training for identifying disinformation signals is vital for effectiveness. People come from	Non-Functional

			<p>diverse backgrounds and have varying levels of prior knowledge and cultural contexts. Tailoring guidance and training to individual needs acknowledges these differences, making the learning experience more relevant and impactful. Additionally, incorporating national differentiation further enhances the tool's adaptability by addressing specific disinformation challenges unique to different regions or countries. This personalized and culturally sensitive approach fosters a more inclusive and effective disinformation combat system that respects individual differences while promoting critical thinking skills universally.</p>	
77	CC-23	The tool should interact with the user in the consumption of the news and no later start.	<p>Ensuring that the tool interacts with users at the time they begin consuming news, rather than later when they initiate a conversation, is crucial for timely and proactive engagement. Many users share news privately as soon as they encounter it online. By intervening at this initial stage, the tool can provide real-time feedback, disinformation signal identification, or critical thinking prompts, effectively curbing the spread of potential disinformation.</p>	Functional
78	CC-24	Warning/highlight of text to identify disinformation	<p>Incorporating warnings or highlights within text to identify disinformation is a practical and user-friendly approach. This feature helps users quickly spot potentially misleading or false information within an article or news piece. By drawing attention to potential disinformation, the tool empowers users to make informed judgments about the content they encounter, strengthening their critical thinking skills and enhancing their ability to recognize and combat misleading information effectively.</p>	Functional
79	CC-25	A system with high user friendliness	<p>A system with high user-friendliness is essential for a positive and productive user experience. This feature ensures that the tool is</p>	Functional

			intuitive, easy to navigate, and accessible to users of all backgrounds and abilities. A user-friendly system simplifies the learning process, reduces frustration, and encourages active engagement.	
80	CC-26	A system with good design on the interface	A system with a well-designed interface is crucial for providing an engaging and visually appealing user experience. This feature encompasses intuitive layouts, clear navigation, and aesthetically pleasing elements that enhance user engagement. A thoughtfully designed interface not only makes it easier for users to access and interact with the tool but also reinforces their trust in its reliability and professionalism.	Non-Functional
81	CC-27	A system with general fast response (time efficient)	A system with fast response times is essential for user efficiency and satisfaction. This feature ensures that users can access information, complete tasks, and interact with the tool swiftly, without unnecessary delays or waiting times. A responsive system not only saves users valuable time but also enhances their overall experience, making their interactions with the tool smooth and seamless.	Functional
82	CC-28	A system with limited bugs	A system with limited bugs is crucial for providing a reliable and frustration-free user experience. This feature ensures that the tool operates smoothly, without disruptive glitches or errors. A bug-free system promotes user confidence in the tool's functionality and credibility. It reduces user frustration, enhances trust, and encourages continued use.	Functional
83	SM-1	The user may need to possess background knowledge about Socrates and his Socratic method.	The user will be able to keep up with the procedure followed (6 steps of Socratic dialogue) to reach a conclusion and understand the outcome of the dialogue.	Non-Functional

84	SM-2	The user may need to be informed about the type of questions and answers that are needed to converse with a chatbot using the Socratic method.	Since the chatbot can engage in two-way conversation, it would be good for the user to know if the chatbot can ask and answer open-ended questions, close-ended questions or/and multiple-choice questions.	Non-Functional
85	SM-3	The user may need to be informed whether the data/output of his/her interaction with the Socratic chatbot will be anonymous or public.	Having agreed on terms about sharing conversation data before starting the conversation, may make the user more willing to have a conversation with the Socratic chatbot.	Non-Functional
86	SM-4	The user may need to possess background knowledge about disinformation and disinformation signals.	The user may require the meaning of disinformation or/and disinformation signals prior to chatting with the Socratic chatbot for more accurate communication.	Non-Functional
87	SM-5	The chatbot may need to give concrete answers.	The user may need specific answers from the model, without made up answers, and if the chatbot doesn't know a specific answer, it could request more information.	Functional
88	CT-1	Initial questionnaire or mini-survey	The user should answer a short questionnaire to assess some individual and personal dimensions associated with critical thinking (cognitive and reflective skills)	Functional
89	CT-2	Critical thinking assessment: dispositional aspect	The user should respond to a number of statements about their usual way of approaching a problem, answering on a 7-point frequency scale (from Never to Always)	Functional
90	CT-3	Critical thinking assessment: objective multiple-choice test	After a brief description of the scenario, the user must respond regarding the accuracy of a statement among few possibilities. The answers are either correct or incorrect.	Functional
91	DS-1	The disinformation signals the chatbot is capable of detecting.	The user may need to be informed about the disinformation signals that the Socratic chatbot is able to detect before engaging in conversation with the chatbot.	Non-Functional
92	DS-2	Detecting how many times a particular disinformation signal appears in an article.	The user may need to be informed about the frequency of a disinformation signal in an article.	Functional
93	DS-3	Detecting how many disinformation signals appear in an article.	The Socratic chatbot may need to inform the user about the number of the disinformation signals that appear in an article to encourage further interaction and critical assessment.	Functional

94	DS-4	The user may need to select to learn about a specific number of disinformation signals.	It would be good for the user to select to learn about a specific disinformation signal or more than one disinformation signal.	Functional
95	DS-5	Providing information about uncertainty about the signal	There is no perfect method of evaluating disinformation, so the user may want to know that the system is not 100% sure of some assessment.	Functional
96	DS-6	Providing explanation why some information may contain signal that allows for classification as (potential) disinformation	Users should know what the automatic evaluation is based on so that they can independently verify whether they should trust such an evaluation, and what its strengths and weaknesses are.	Non-Functional
97	PIA-1	The PIA component should give the results as a user understandable metric.	When the users ask for PIA, the PIA component computes the estimate and gives the result out as numerical value. Ensuring that the PIA component provides results in a user-understandable metric is essential for enabling users to comprehend the reach and impact of disinformation.	Functional
98	PIA-2	The user should be able to feed text content of an article to the PIA component.	This PIA tool relies on the actual textual content of an article to assess its potential impact. In other words, to gauge how widely and deeply the information within an article might spread and influence people, it requires access to the article's written content.	Functional
99	PIA-3	The user needs to share his/her social media user account name with the TITAN system for a better estimation of the PIA.	Allowing users to share their account name for the selected social media service, along with access to their social graph information and historical post data, enables the system to create a more comprehensive and accurate social network analysis. This feature enhances the propagation impact assessment, as it can leverage the user's existing connections and posting history to gauge the potential reach and impact of disinformation shared within their network.	Functional
100	PIA-4	The user should be able to feed meta-data related to the article to the PIA component.	Enabling users to feed metadata related to the article into the PIA tool enhances the depth and accuracy of the assessment. Metadata, such as publishing time, provides valuable	Functional

			context that can significantly influence the potential impact of an article. By incorporating this metadata, the assessment component can make more informed judgments about how the information might propagate and impact audiences.	
101	PIA-5	The user should be able to select the social media service that the account name is linked to from a list of supported social media services.	Allowing users to select the social media service to which their account name is linked from a list of supported platforms ensures flexibility and compatibility. This feature simplifies the process of connecting social media profiles to the tool by providing a user-friendly selection interface. Users can choose the specific platform they use, enhancing the tool's adaptability to various social media services.	Functional
102	MI-1	The user may need to get informed about the estimated amount of time that is required to take a micro-lesson.	Providing users with information about the estimated amount of time required to complete a micro-lesson is essential for effective time management and user engagement. This feature empowers users to make informed decisions about when to engage with a lesson, aligning their learning goals with their available time.	Non-functional
103	MI-2	The user may need to select the amount of time to take a micro-lesson.	Allowing users to select the amount of time they wish to dedicate to a micro-lesson offers flexibility and personalized learning. This feature recognizes that users have varying time constraints and learning preferences. By providing options for time allocation, users can tailor their engagement with micro-lessons to fit their schedules and learning pace.	Non-Functional
104	MI-3	The user may need to select to take a micro-lesson concerning specific disinformation signals or concerning disinformation signals in general.	Allowing users to choose between taking a micro-lesson focused on specific disinformation signals or a general disinformation signals lesson enhances their learning experience. This feature caters to users with varying levels of familiarity and expertise in the field of disinformation. Users can opt for specialized lessons to deepen their knowledge in specific areas or	Functional

			choose general lessons for a broader understanding.	
105	MI-4	Feedback on how well someone did the micro lesson.	Providing feedback on how well someone did in a micro-lesson is a fundamental element of the learning process. This feature offers users a valuable assessment of their performance, helping them understand their strengths and areas for improvement in critical thinking skills. Constructive feedback can include insights into correct answers, areas that need more attention, and suggestions for further learning.	Functional
106	S-1	The user may want to receive badges and points while using TITAN	Awarding users with badges and points as they use TITAN adds a gamified element to the learning experience. This feature encourages active participation and motivation by recognizing and rewarding achievements.	Functional
107	S-2	The user may want to progress overtime (i.e., do better than the previous time you interacted with TITAN)	Allowing users to track their progress over time by aiming to perform better than their previous interactions with the system is a powerful motivational tool. This feature encourages users to set personal improvement goals and engage consistently with the tool to surpass their own achievements.	Functional
108	S-3	The user may want to see how the TITAN community progresses (i.e., see statistics of how the community responded, their progress in TITAN, etc)	Allowing users to track the progress of the community, including statistics on how the community responds and their overall progress within the system, promotes a sense of belonging and competition. This feature provides users with valuable insights into collective achievements, fostering a sense of community and shared goals.	Functional
109	S-4	The user may want to have progress levels and ranks	Implementing progress levels and ranks for users adds an element of achievement and recognition to their journey with the tool. This feature allows users to visually track their advancement and feel a sense of accomplishment as they climb through different levels and ranks.	Functional

Table 18.1. List of user requirements

The user requirements presented here encompass a wide range of aspects, including transparency, privacy, data protection, fairness, explainability, and user control. By incorporating these requirements, our project aims to ensure the chatbot not only delivers accurate and reliable information but also operates in a manner that respects user rights, builds trust, and empowers users in their fight against disinformation.

The user requirements list represents a crucial outcome of our project, shaped by the active participation and insights of various stakeholders, including higher education students, NGOs, migrants, and refugees. It reflects the collective efforts to create a tool that aligns with legal and ethical principles, meets societal expectations, and addresses the diverse needs of users in combating disinformation.

In conclusion, the user requirements presented in this section mark the culmination of our project's inclusive and user-centered approach. Through workshops, stakeholder involvement, legal and ethical analysis, and engagement with users, we have developed a comprehensive set of requirements that embody the values of transparency, privacy, fairness, and user empowerment.

These user requirements serve as a roadmap for the further refinement and development of the chatbot. By addressing these requirements, we aim to build a tool that not only effectively combats disinformation but also prioritizes user satisfaction, respects legal obligations, and upholds ethical standards.

Moving forward, these user requirements will guide the project team in iteratively enhancing the chatbot's functionality, user experience, and adherence to legal and ethical frameworks. By continuously incorporating user feedback and staying attuned to evolving user needs, we ensure that the chatbot remains a reliable and trusted tool in the fight against disinformation.

The user requirements presented in this section represent a significant milestone in our project's journey. They embody the collective efforts of the project team, stakeholders, and users in creating a chatbot that is both effective in combating disinformation and user-centric in its approach.

19 ANNEX 3: ARTICLES AND CORRESPONDING SAMPLE DIALOGUES

In this section, we will provide an overview of the examples of articles chosen by our user partners and the disinformation signals identified within them. These articles were selected as the basis for training our rule-based chatbot, which plays a crucial role in combating disinformation. Additionally, we will present a sample of the dialogues created by our project partners, which serve as valuable inputs for the rule-based system. Although we present one specific example, it is important to note that a total of 26 dialogues were created, corresponding to the analysis of 10 articles.

Example Articles and Disinformation Signals:

Our user partners actively contributed to the project by selecting articles that are known to contain disinformation signals. These articles were chosen for their relevance and representativeness of disinformation techniques prevalent in various contexts. By analyzing these articles, we were able to identify key disinformation signals, including logical fallacies, misinformation, manipulation tactics, and distorted facts.

Through a rigorous analysis process, our team meticulously scrutinized the selected articles, highlighting the specific disinformation signals present within them. This thorough examination allowed us to build a solid foundation for training our rule-based chatbot, enabling it to detect and counter disinformation.

In particular, the owners of each use case have provided up to five articles each, and with the assistance of the relevant partners, several disinformation signals have been identified throughout each article and have been highlighted with colors where each color represents a different and distinct disinformation signal. Below, we quote one of the articles shared by use case 1 and provide the identified disinformation signals highlighted in corresponding colors.

The following article was published in TheTruthAboutTechnology.org which is not a known trustworthy source and the readers in general should be careful about sources claiming to have the truth.

The following disinformation signals have been identified and are marked in the article with the respective colors.

- Emotional text / provocative information / hate speech (PURPLE)
- Non-cross-checked information / unverified source (YELLOW)
- fallacies (GREEN)
 - slippery slope
 - implicit assumption
 - red

herring

The government is spying on you through LED-lighting!!

You ever noticed how cheap LED-lighting is? You cannot even find regular, normal lighting anymore. The government and big businesses are pushing out regular lights and forcing us to have as many LED-lights as possible. They say this is because of "Climate Change", but Climate Change is still unstoppable, so this is proven to be total bullshit.

We all know that LED came out of nowhere and is taking over everything now. What very few people know is that unlike regular lights, LED is connected to the Internet of Things. The Internet of Things is a huge network in which all electronic devices are connected and share information and data on people with the government and big multinationals.

LED are the eyes of this system. You ever seen an LED-light flickering? That's right. That means its monitoring you and sharing this data. This is how the government is controlling us and making big multinational companies ever RICHER! This is total and unacceptable ABUSE!

Moreover, university research all over the world have shown that LED-lighting is bad for your eyes. Therefore, if this goes on, it will not be before long that it makes us totally blind. Bastards politicians and CEO's are actively PIOSONING us to silent us.

It is scientifically proven and now the truth is out! Share this with as many people as possible, because you care for them.

We must STOP the power of the Government and the RICH.

Sample Dialogues for the Rule-Based System:

In addition to analyzing the articles, our project partners collaborated to create a series of Socratic dialogues. These dialogues were designed to enhance users' critical thinking skills and serve as input for the rule-based system. By engaging users in thoughtful and interactive conversations, we aimed to empower them with the necessary tools to recognize and challenge disinformation effectively.

Although we showcase one example dialogue, it is important to note that a total of 26 dialogues were meticulously crafted by our project partners. Each dialogue was tailored to address specific disinformation signals identified within the selected articles, providing users with a comprehensive understanding of disinformation techniques and strategies. These dialogues serve as valuable inputs to the rule-based system, enriching its knowledge base and enhancing its ability to deliver accurate and informative responses.

Below, a sample dialogue is provided assuming that the participant's critical thinking level is low (only low and medium levels of critical thinking were considered), and it corresponds in the article provided above. The dialogue incorporates the six steps of the Socratic method that were analyzed in depth in Table 4 of Section 4.3.3.

Socratic Sample Dialogue

Author: IPT

Article: "The government is spying on you through LED-lighting!!"

Critical thinking level: Low

Disinformation signal: Logical Fallacy - Appeal to Authority (argumentum ad auctoritatem)

AIS = AI Socratic (Intelligent Agent)

C = Citizen

[Step 1: Clarification]

AIS: Hello! Let me introduce myself, my name is TITAN, your AI Intelligent Coach against Disinformation

C: Hello!

AIS: You've selected the article: "The government is spying on you through LED-lighting!!"

C: Yes.

AIS: What is the main issue in the article?

C: The article claims that governments and large corporations are using LED lights to monitor individuals. This is conducted through the Internet of Things.

[Step 2: Challenging assumptions]

AIS: What caught your attention about this article?

C: The topic is extremely interesting and raises an issue of major concern.

AIS: What do you already know about the issue?

C: This is the first time I've heard about this issue.

AIS: Would you like to learn more about the quality of reasoning in this article?

C: Certainly, yes!

AIS: Good reasoning or sound reasoning refers to avoiding logical fallacies or errors in an argument.

C: That's helpful.

AIS: Would you like me to assist you in identifying logical fallacies in this article?

C: Yes, but I don't know what a logical fallacy is exactly.

AIS: A logical fallacy is a flaw in reasoning that makes an argument invalid or unsound. A logical fallacy can make an argument potentially misleading. [Suggestion: Microlesson on Sound Reasoning and Logical Fallacies]

C: That's useful.

[Step 3: Evidence and reasoning]

AIS: We can focus on a logical fallacy called Appeal to Authority (argumentum ad auctoritatem).

C: Fine!

AIS: Would you like to learn more about this logical fallacy?

C: Yes. [Suggestion: Microlesson on Appeal to Authority logical fallacy]

AIS: Appeal to Authority (argumentum ad auctoritatem) occurs when a claim is supported by relying on the expertise of an authority figure, or an expert in a field, rather than on sound reasoning or objective evidence.

C: Can you provide an example?

AIS: Here is an example "You should take vitamins A and C because experts in healthcare said they are the best." Can you identify the logical fallacy in this claim?

C: Not really, this claim is based on the expertise of healthcare specialists.

AIS: Yes, but is there any evidence provided to support the claim?

C: No.

AIS: This is the reason why this example is a case of an Appeal to Authority fallacy.

C: Can you clarify?

AIS: The claim in the example urges us to use vitamins because experts in healthcare said so, without providing sound reasoning or objective evidence.

C: That's informative.

AIS: Can you identify a similar practice in the article?

C: Yes, I think the following statement "university research all over the world have shown that LED-lighting is bad for your eyes" [p]

AIS: What led you to that decision?

C: The expression "university research" appeals to the authority of university researchers

AIS: Yes, but can we admit that university research is conducted by experts on a field with careful scientific research that is published in reputable and reliable academic journals?

C: Yes, we can admit it.

AIS: So, why do you suggest that this could be the case of an Appeal to Authority fallacy?

C: There is no evidence provided to support the claim, and without any cited sources.

AIS: Brilliant! So, the statement [p] can be called into question.

C: Yes

[Step 4: Alternative viewpoints]

AIS: Are there any alternative viewpoints?

C: Perhaps, no.

AIS: What about claiming that despite the unspecified evidence of the assertion, the conclusion might be correct.

C: How is this possible? I'm not sure.

AIS: Your question is valid. We need further reasons to doubt the validity of the statement [p].

C: Yes.

AIS: Is there any additional reason to doubt the validity of the statement [p]?

C: I am not sure.

AIS: Consider the assertion in statement [p] that university research is globally applicable, “all over the world”

C: That’s interesting.

AIS: The phrase is a generalization, a broad expression suggesting that something is globally or universally true.

C: That’s helpful

AIS: Drawing a conclusion based on limited or insufficient evidence can result in another logical fallacy, known as Hasty Generalization.

C: Could you please explain further? [Suggestion: Microlesson on Logical Fallacies / Hasty Generalization]

AIS: A Hasty Generalization occurs when conclusions are drawn about an entire group or category based on insufficient or limited evidence, often relying on stereotypes.

C: Can you provide an example?

AIS: Consider the claim that “A group of teenagers sabotaged the exhibition. There is no doubt that teenagers are reckless and vicious”.

C: That’s informative.

AIS: So, do you think that statement [p] is also subject to a Hasty Generalization fallacy?

C: Yes.

[Step 5: Implications and consequences]

AIS: Are there any consequences of accepting statement [p]?

C: I don’t think there are any serious implications.

AIS: Could a fallacious claim, which relies on Hasty Generalization and Appeals to Authority, cause unwarranted anxiety or unnecessary fear in the reader?

C: Yes, if it is taken seriously.

AIS: Moreover, do you agree that the statement [p] might bring mistrust to LED-lighting technologies, and their possible advantages?

C: Yes, I agree.

AIS: Would you like to learn more about the implications and consequences of disinformation?

C: Yes > Go to [Microlesson on Disinformation Implications and Consequences]

C: No > Go to Step 6

[Step 6: Challenging the question]

AIS: Was it appropriate to check the article for the Appeal to Authority fallacy?

C: Yes, definitely.

AIS: Do you agree that a claim which relies on unsupported authority and broad generalizations, is not only invalid but also carries potential negative consequences?

C: Yes, indeed.

AIS: So, considering the Appeal to Authority fallacy, the claim should be doubted, and the article is in question.

C: Yes.

AIS: We can also assert that the argument is not valid, as far as the evidence provided is not sufficient to support the claim.

C: Correct.

AIS: What would be a recommended course of action once an Appeal to Authority fallacy has been identified?

C: I am not sure.

AIS: I would suggest examining the validity of statement [p], by finding reliable evidence and unbiased sources.

C: That’s a good suggestion.

AIS: Would you recommend any additional evaluation of the article?

C: Perhaps, the article should be checked for more logical fallacies.

AIS: Would you like further assistance?

C: Yes [= Go to Step 3]

C: No

AIS: Thank you for using TITAN!

Conclusion:

In this section, we have provided a glimpse into the examples of articles chosen by our user partners, highlighting the disinformation signals identified within them. Through the analysis of these articles, we have developed a robust understanding of disinformation techniques, allowing us to train our rule-based chatbot.

Furthermore, we have showcased a sample dialogue created by our project partners, which is part of a collection of 26 dialogues crafted to foster critical thinking and feed the rule-based system. These dialogues serve as essential inputs, enabling the chatbot to engage users in meaningful conversations and equip them with the necessary skills to combat disinformation.

By combining the insights gained from analyzing the articles and the power of the Socratic dialogues, our project takes a comprehensive approach to address disinformation. The examples and dialogues presented in this section provide a glimpse into the depth and breadth of our efforts to develop a chatbot that is well-equipped to identify, counter, and educate users about disinformation in a proactive and informed manner.