



DELIVERABLE REPORT D2.2

DELIVERABLE

SUBMISSION DATE	NAME OF THE DELIVERABLE	WORK PACKAGE
30.10.2024	D2.2: TITAN Methodologies and Models	WP2
NATURE	AUTHOR(S)	LEAD BENEFICIARY
Public	Maria Amata Garito (UTIU), Gian Vittorio Caprara (UTIU), Giuseppe Corbelli (UTIU), Andrea Falegnami (UTIU), Francesco Saverio Nucci (UTIU), Marinella Paciello (UTIU), Alessandro Pollini (UTIU), Andrea Tomassi (UTIU) P. Kapetanakis, (NCSR), I. Paparrigopoulos (NCRS), I. Papachristou (NCRS), D. Katsamori (NCRS), G. Petasis, (NCSR) John Lawrence (UNIVDUN)	UNINETTUNO

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
V1.1	12/15/23	Structure and first version	Alessandro Caforio	UTIU
V1.2	12/18/23	Introduction and Section 3 added	Giuseppe Corbelli, Andrea Falegnami, Andrea Tomassi	UTIU
V1.3	12/21/23	Revised version and added figures and citations	Giuseppe Corbelli, Andrea Falegnami, Andrea Tomassi	UTIU
V1.4	04/04/24	Revised version	Giuseppe Corbelli, Andrea Falegnami, Andrea Tomassi	UTIU
V2.0	07/05/24	Contextualization and Introduction revision	Massimo Buscema, Marco Breda, Giuseppe Corbelli, Andrea Falegnami, Andrea Tomassi	UTIU
V2.4	09/07/24	Revised version	Francesco Saverio Nucci	UTIU
V3.0	10/09/24	New Contribution on dialogues	P. Kapetanakis, I. Paparrigopoulos, D. Papachristou, D. Katsamori, G. Petasis	NRCS
V4.0	30/09/24	Final structure definition and document revision to include new contributions	Francesco Saverio Nucci, Andrea Falegnami, Andrea Tomassi	UTIU
V5.0		Peer Review	John Lawrence	UNIVDUN
V5.2		Conclusion paragraph added	Francesco Saverio Nucci	UTIU
V5.3		Minor changes, Final version	Francesco Saverio Nucci	UTIU

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 authorized 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

DELIVERABLE.....	1
PROJECT DETAILS	1
QUALITY CONTROL ASSESSMENT	2
DISCLAIMER.....	2
STATEMENT OF ORIGINALITY	2
ACKNOWLEDGEMENT	2
TABLE OF CONTENTS	3
EXECUTIVE SUMMARY	5
SECTION I SOCRATIC METHOD AND DIALOGUES	8
1 SOCRATIC METHOD FOR LOGICAL ASSESSMENT OF INFORMATION RELIABILITY	8
1.1 THE SOCRATIC METHOD	8
1.2 THREE KINDS OF PROPOSITIONS	9
1.3 FURTHER APPROACHES	11
2 CRITICAL THINKING AND MEDIA LITERACY METHODOLOGY.....	11
2.1 INTRODUCTION.....	11
2.1.1 <i>The Challenge of Navigating Media Complexity & Information Disorder.....</i>	<i>11</i>
2.1.2 <i>Disinformation & Misinformation: Types and Tactics.....</i>	<i>12</i>
2.1.3 <i>Why Critical Thinking and Media Literacy are important.....</i>	<i>13</i>
2.1.4 <i>What Media Literacy Is</i>	<i>13</i>
2.1.5 <i>What Critical Thinking Is: A Common Ground Definition.....</i>	<i>13</i>
2.2 CRITICAL THINKING PRINCIPLES USED BY TITAN	14
2.2.1 <i>Critical Thinking requires Generic Critical Thinking Abilities.....</i>	<i>14</i>
2.2.2 <i>Critical Thinking requires a Critical Mindset/Dispositions.....</i>	<i>15</i>
2.2.3 <i>Critical Thinking requires Domain-specific Knowledge</i>	<i>16</i>
2.2.4 <i>Navigating Media Requires Media Critical Thinking (MCT)</i>	<i>16</i>
2.2.5 <i>Prebunking & Inoculation: “Prevention is the best cure”</i>	<i>18</i>
2.2.6 <i>Media Critical Thinking Rests on the Socratic Method</i>	<i>19</i>
3 CRITICAL THINKING ASSESSMENT (CTA) METHODOLOGY	20
3.1 MAIN OBJECTIVE AND SCOPE	21
3.2 DEFINITION OF THE CRITICAL THINKING ASSESSMENT MODEL AGAINST DISINFORMATION....	21
3.3 PROPOSAL FOR THE TECHNICAL IMPLEMENTATION.....	28
3.3.1 <i>Objective CT assessment</i>	<i>28</i>
3.3.2 <i>Dispositional self-assessment of Critical Thinking</i>	<i>34</i>
3.3.3 <i>Individual variables related to self-regulation</i>	<i>34</i>
3.3.4 <i>Data elaboration and recognition of users’ characteristics.....</i>	<i>37</i>
3.3.5 <i>Matching users’ characteristics and TITAN system’s recommendation.....</i>	<i>38</i>
3.4 RESEARCH DIMENSIONS	38
3.4.1 <i>TITAN CTA in Living Labs</i>	<i>38</i>
3.4.2 <i>Dimensions for TITAN Pilots.....</i>	<i>38</i>
4 THE DIALOGUE IMPLEMENTATION METHODOLOGY	39
4.1 THE FOUR BLOCK APPROACH	39
4.1.1 <i>Rationale of the Four-Block Approach.....</i>	<i>39</i>
4.1.2 <i>Block 1: TITAN’s Psychological Inoculation prebunking method</i>	<i>41</i>

4.1.3	<i>Block 2 - Item-Focused Dialogues, Short Description</i>	44
4.1.4	<i>Block 3 - person (user)-focused dialogues</i>	48
4.1.5	<i>Block 4 dialogues</i>	50
4.2	THE SOCIOLOGICAL DIMENSION	54
SECTION II SUPPORTING METHODOLOGIES AND TOOLS		60
5	EVALUATION OF DISINFORMATION SIGNALS	60
5.1	TRAINING METHODOLOGIES	60
5.1.1	<i>Traditional Learning, Sentenced-based Classification</i>	60
5.1.2	<i>Sequential Labeling Classification</i>	60
5.1.3	<i>Few-Shot learning</i>	61
5.1.4	<i>Zero-Shot Learning</i>	61
5.2	EVALUATION METHODOLOGIES	62
5.2.1	<i>Comparison with Gold Standard Corpus</i>	62
5.2.2	<i>User Feedback Collection</i>	62
5.2.3	<i>Active Annotation</i>	63
5.2.4	<i>TITAN Training & Evaluation Methodologies</i>	63
6	PROPAGATION IMPACT MEASUREMENT VIA SOCIAL MEDIA MONITORING	64
6.1	MAIN FACTORS	64
6.2	PROPOSAL FOR TECHNICAL IMPLEMENTATION	65
6.3	EVALUATION	65
7	PERSONALIZED MICRO-LESSONS FOR MEDIA LITERACY AND CRITICAL THINKING	66
7.1	PERSONALIZED MICRO-LESSONS STRUCTURE	66
7.2	MEDIA LITERACY AND CRITICAL THINKING	66
7.3	METHODOLOGY OF MICRO-LESSONS COLLECTION	68
8	USER ENGAGEMENT THROUGH GAMIFICATION METHOD	70
8.1	LITERATURE REVIEW ON WAYS TO ENHANCE ECOSYSTEM SUSTAINABILITY AND USER ENGAGEMENT	70
8.2	MOTIVATION THEORIES & MOTIVATION TYPES	70
8.3	MOTIVATION TYPES- INTRINSIC & EXTRINSIC MOTIVATION THEORIES	72
8.3.1	<i>Intrinsic motives</i>	72
8.3.2	<i>Extrinsic motives</i>	73
8.3.3	<i>Gamification: an emerging motivational factor</i>	74
8.3.4	<i>Gamification Mechanisms & TITAN</i>	76
9	CONCLUSION	81
10	REFERENCES	83
ANNEXES		87
ANNEX 1: TYPES OF INFORMATION DISORDER		88
ANNEX 2: TACTICS OF DISINFORMATION		90
ANNEX 3: ITEM-FOCUSED DIALOGUES		93
ANNEX 4: MEDIA LITERACY PROTOCOL		120
ANNEX 5: DIALOGUES EXAMPLES FOR CT DIMENSIONS		125
ANNEX 6: DIALOGUES EXAMPLES FOR SELF REGULATION DIMENSION		128

EXECUTIVE SUMMARY

The TITAN project is a comprehensive initiative aimed at addressing the growing concern of disinformation in today's media landscape. Recognizing the critical need for tools that enhance critical thinking, the project is dedicated to developing an innovative solution that provides guidance and coaching to citizens, enabling them to navigate the complex and often misleading information environment. This initiative is part of a broader effort funded by the Horizon Europe Programme of EU and UK Research and Innovation, emphasizing its significant and far-reaching impact. The project's main objective is to equip individuals with the skills necessary to critically evaluate information and counter disinformation effectively. At the heart of the TITAN project is the adaptation of the Socratic Method, a classical approach to fostering critical dialogue and rigorous questioning. This method encourages users to delve deeper into the information presented to them, challenging assumptions, identifying biases, and evaluating the logical consistency of the content. By engaging in structured questioning, individuals can reach their conclusions regarding the reliability and truthfulness of the information they encounter.

This deliverable (**D2.2**) presents the outcomes of **Task T2.3**, which builds on previous tasks focused on scientific analysis and technical requirements. It outlines key methodologies that will continue to evolve through collaboration with other work packages, technical developments, and piloting efforts.

The report covers eight essential socio-technical methodologies and components, divided into two sections for clarity of reader. This forms the backbone of TITAN's approach, the methodologies have been divided into these two subsets to mirror the importance of the different methodologies inside TITAN system.

The first section: Socratic Method and Dialogue, is fully dedicated to Socratic Method and Critical Thinking Methodology, including a deep analysis of the methodology to assess this specific kind of Human feature, the so-called Critical Thinking Assessment, this section is composed by the following different methodologies:

1. **Socratic Method for Logical Assessment of Information Reliability:** TITAN adapts the Socratic method, a classical questioning technique, to help users critically evaluate the reliability of statements and information. This method encourages deep, logical inquiry rather than relying solely on fact-checking. It enables individuals to explore the credibility, bias, and logical coherence of the information they encounter, fostering a more nuanced understanding. Some preliminary indications on this methodology is already presented in the deliverable D2.1, however in this deliverable a more detailed analysis has been conducted also after the first pilot analysis and its associated results in order to specify better how to implement and include this methodology inside the TITAN system
2. **Critical thinking and media literacy methodology:** These methodologies are central to combating disinformation. Media literacy teaches individuals how to recognize manipulation and assess the authenticity of media content. Critical thinking, on the other hand, involves evaluating the accuracy and logic of the information encountered. Together, these skills enable users to navigate today's media landscape effectively. The modern media environment, rife with misinformation and disinformation, necessitates robust critical thinking and media literacy interventions
3. **Critical Thinking Assessment (CTA) Methodology:** A key feature of TITAN is its Critical Thinking Assessment (CTA) tool, which evaluates users' abilities across cognitive domains such as

analysis, interpretation, evaluation, inference, and self-regulation. The CTA methodology is designed to assess and improve users' skills in recognizing and countering disinformation. By tailoring the assessment to cognitive profiles, TITAN can provide personalized interventions to strengthen critical thinking abilities.

4. **The Dialogue implementation methodology** The dialogue system employs a modern adaptation of the Socratic Method, which uses structured, open-ended questions to stimulate deep critical inquiry. This method encourages users to question assumptions, evaluate evidence, and reflect on their reasoning processes, thereby enhancing both media literacy and critical thinking. The dialogue implementation method will use a four-block approach, each serving distinct goals: **Block 1** focuses on psychological inoculation by prebunking disinformation tactics through passive approaches like infographics. **Block 2** guides users through critical evaluation of specific media items using a Socratic method, prompting them to assess the credibility of authors and media sources. **Block 3** encourages reflective, person-focused dialogues, helping users develop critical thinking dispositions, following the first five dispositions of our Critical Thinking Model. **Block 4** integrates self-regulation strategies (the last dimension of the CT model) and fosters a holistic approach to critical thinking. Finally the implementation of the dialogues also takes into consideration some important sociological dimensions described in the section.

The second section: supporting methodologies and tools, is dedicated to supporting methods and tools which facilitate the user in navigating into the disinformation context, but also to integrate more information and inputs to produce the dialogue, during the implementation phase of the TITAN system. Overall, the TITAN project represents a multifaceted and forward-thinking approach to combating disinformation. By integrating a range of methodologies and tools, the project not only addresses the immediate challenges posed by false information but also fosters a culture of critical engagement and informed decision-making. In this section the following supporting methodologies are described:

5. **Evaluation of Disinformation Signals:** TITAN's approach to combating disinformation includes developing advanced tools for detecting false information. These tools are based on evaluating disinformation signals, using methodologies that incorporate both traditional machine learning and newer techniques such as few-shot and zero-shot learning. These methods allow for precise detection and categorization of misleading content in various contexts. Tools and techniques are developed to identify and assess false information. By understanding the patterns and signals associated with disinformation, users can better recognize and counteract deceptive content.
6. **Propagation Impact Measurement via Social Media Monitoring:** Recognizing the vital role that social media plays in the spread of disinformation, TITAN monitors how false information propagates across networks. This component analyzes the reach and influence of disinformation and assesses the effectiveness of different strategies to counteract its spread. By doing so, TITAN provides insights into the dynamics of information dissemination and helps identify the most effective countermeasures.
7. **Personalized Micro-Lessons for Media Literacy and Critical Thinking:** TITAN offers users tailored micro-lessons designed to improve their media literacy and critical thinking skills. These lessons are customized based on users' cognitive profiles, ensuring that the content is both relevant and impactful. By providing targeted education, TITAN equips individuals with the knowledge they need to critically engage with information and identify disinformation.

8. **User Engagement through Gamification method:** To ensure the long-term success and user engagement of the TITAN Citizen Intelligent Ecosystem, the project integrates gamification strategies to enhance user engagement. By incorporating motivational theories and gamification mechanisms, TITAN incentivizes users to remain active participants in critical thinking activities. This approach not only keeps users engaged but also helps create a community focused on combating disinformation. Ensuring the sustainability of the TITAN Citizen Intelligent Ecosystem is a critical goal of the project. This involves creating an incentives framework that encourages continuous user engagement and participation.

Methodologies and Technical Approach:

The report details how the TITAN project integrates advanced techniques, such as **social media monitoring, disinformation signal detection**, and **machine learning models** for classifying and understanding disinformation. It explores the use of **few-shot learning** and **zero-shot learning** models to adaptively detect new disinformation patterns with minimal prior training data, making TITAN adaptable to evolving misinformation trends.

A significant aspect of the TITAN methodology is its application in **Living Labs** and pilot studies, where the assessment tools will be tested in real-world scenarios. Feedback from these studies has been to refine the tools, ensuring they are practical, effective, and user-friendly across diverse populations. The deliverable also emphasizes **user engagement** through personalized feedback and continuous improvement of tools based on user interactions.

Broader Integration:

The TITAN project is more than just a set of tools—it represents a comprehensive ecosystem that fosters a culture of critical thinking and informed decision-making. By integrating methodologies that range from cognitive assessment to social media monitoring, TITAN addresses both the cognitive and social dimensions of disinformation. The continuous refinement of these methodologies through real-world testing in Living Labs and pilot programs ensures that TITAN will remain adaptable and relevant in an ever-changing media landscape.

Through its multi-faceted approach, TITAN aims to create a resilient population equipped to navigate the challenges of the digital information age. The project's focus on personalized interventions and community engagement ensures that its impact will be both deep and lasting.

Implications and Future Directions:

TITAN represents a multi-dimensional and sustainable approach to addressing disinformation by equipping individuals with the cognitive tools needed for critical engagement. Through innovative methodologies, personalized learning, and gamified incentives, the project aims to create a resilient, informed public capable of defending against the growing threat of disinformation.

As TITAN progresses, its tools and models will be continuously refined based on pilot study outcomes, ensuring that the ecosystem remains robust, adaptable, and relevant in countering evolving disinformation challenges. The long-term vision of the project is to contribute to a **informed digital society** where individuals are better equipped to discern fact from falsehood in the media environment.

1 SOCRATIC METHOD FOR LOGICAL ASSESSMENT OF INFORMATION RELIABILITY

1.1 THE SOCRATIC METHOD

While deliberate misinformation, disinformation, and deception are by no means new societal phenomena, the recent rise of fake news (Lazer et al., 2018) and information silos (Flaxman et al., 2016) has become a growing international concern, with politicians, governments and media organisations regularly lamenting the issue. Efforts to combat such disinformation dressed up as genuine news focus too often exclusively on the factual correctness of the claims made. Whilst the truth of purported facts is clearly of crucial importance, there are other, often overlooked, aspects to consider here. It is, after all, very possible to argue from true factual statements to blatantly false or misleading implications by applying skewed, biased, or otherwise defective reasoning. Furthermore, the categorical corrections on factual impropriety delivered by fact-checkers can both alienate readers who believe they are being told what to think (Nyhan & Reifler, 2010) and raise questions around the impartiality of the fact-checkers themselves (Dotsnuanon, 2022).

The Socratic method is a form of cooperative argumentative dialogue, based on asking and answering questions to stimulate critical thinking and to draw out ideas and underlying presuppositions. The Socratic method can be an effective tool to evaluate the truthfulness, bias, and credibility of information, when applied to the logical assessment of information reliability. Such a dialogue has the benefit of not telling the reader what to think, or what decision to come to about a given piece of information, but instead of collaboratively deepening understanding of the material in question, thus allowing the reader to reach their own conclusions regarding trustworthiness. Integrating the unitary stages of the Socratic method into a logical assessment of information reliability involves a detailed, systematic approach to critical evaluation.

Initially, the process demands clarification, where insightful questions are posed to unravel complex or vague assertions, ensuring a common understanding and eliminating ambiguities. This foundation allows for a deeper examination of underlying assumptions, where the validity and origin of these presuppositions are questioned, potentially revealing alternative, more valid assumptions and fostering new perspectives. The scrutiny extends to the rationale behind claims, focusing on the logic and evidence presented to identify any weaknesses or inconsistencies, thereby assessing the argument's soundness.

Further exploration into individual viewpoints encourages a detailed articulation of opinions, revealing biases or subjective influences through requests for further explanation and justification. The method also emphasizes the importance of considering the broader implications and potential consequences of beliefs or claims, prompting a holistic understanding of their impact on various stakeholders.

Finally, challenging the questioning process itself encourages a meta-cognitive reflection on the motives and methods of inquiry, enhancing understanding of the dynamics of critical thinking. This comprehensive approach forms a robust framework for assessing information reliability, as this method encourages deep engagement with content, fostering a habit of critical thinking and skepticism that is crucial in navigating the vast amounts of information encountered daily. The focus then turns to what form these questions should take, what makes them insightful with respect to the material under consideration, and how or when they should be introduced in a structured dialogue.

As a starting point for the types of questions that can be asked, we can consider the distinction made by (Visser & Lawrence, 2022). In this work the authors consider three levels of questioning: questions relating to individual statements or propositions; questions regarding validity of common reasoning patterns; and, questions relating to the reasoning structure as a whole.

1.2 THREE KINDS OF PROPOSITIONS

At the statement or propositional level, three different proposition types can be considered: *Factual* propositions, *Value* based propositions, and those concerning the expression of a *Policy*. A proposition is classified as a proposition of *Fact* if its veracity can be verified through empirical observation, for example “investment in solar energy has increased by 76.4%”. It is classified as a proposition of *Value* if it contains some evaluation (such as, ethical (e.g. right/wrong), aesthetical (e.g. beautiful/ugly), legal (e.g. guilty/innocent), or logical (e.g. true/false) evaluations). An example of a value proposition would be the statement “investing in solar energy is a good idea”. Finally, it is classified as a proposition of *Policy* if it expresses a plan of action or an act to be carried out, such as “the UK should invest in solar energy”. Each of these proposition types in turn give rise to questions that can be asked about them, immediately offering an expansion on fact-checking by broadening the range of proposition types that can be checked. Where the veracity of facts can be questioned, policies could be questioned for consistency; although two policies may individually be reasonable, if they are inconsistent then clearly both could not be used in the same argument. Similarly, values could be questioned for popularity; supporting a policy with a value may be reasonable, but it all depends on how widely shared that value is. Proposition type classification can also be used as an indicator of various types of fallacious argument. For example, an argument where a value is being used to support a fact, or a policy is being used to support a fact, as in the example below:

“It won’t rain [**fact**], because it would be bad to have to cancel our picnic [**value**]”.

- Example adapted from (Van Eemeren & Henkemans, 2016).

Presented with such a situation, it would be reasonable to ask if the premise is a sufficient reason for accepting the veracity of the conclusion. This consideration of patterns in proposition types can be expanded considerably through the consideration of argumentation schemes and their associated critical questions.

The Argumentation Scheme model of human argumentation proposed by Douglas Walton (Walton et al., 2008) provides a framework to effectively model more than 60 stereotyped patterns of argumentative reasoning (i.e., argumentation schemes), and a wide set of critical questions to challenge their validity from both factual and logical viewpoints. Due to their versatility and usefulness (Reed and Walton, 2001), argumentation schemes have recently gained great popularity in the field of computational argumentation: in argument mining (Lawrence and Reed, 2016), in argument generation (Saha and Srihari, 2023), and in argument-based Human-Computer Interaction (Ruiz-Dolz et al., 2023) among others. Therefore, the Argumentation Scheme model represents a promising starting point from which to design misinformation detection systems that overcome the limitations of existing work. With this model of argumentation, and specifically the critical questions, we can address both factual and logical issues in natural language arguments containing misinformation. Furthermore, the critical questions themselves allow for the user to explore the reasoning in greater depth, giving not just a verdict on the validity of the reasoning, but an explanation as to why this is the case.

Let us consider the *Argument from Position to Know* scheme as an example, Walton et al. (2008) define this argument as:

Major Premise: Source **s** is in position to know about things in a certain subject domain **f** containing proposition **p**.

Minor Premise: **s** asserts that **p** is true (false).

Conclusion: **p** is true (false).

An argumentation scheme thus provides a set of abstract variables (i.e., **s**, **f**, and **p** in this case) that can be replaced with natural language text, together with the connections (in natural language) between these variables required to make a specific argumentation scheme instance. Furthermore,

Walton provides the set of predefined critical questions, the answers to which determine the validity of the argument from both factual and logical viewpoints:

CQ1: Is *s* in position to know whether *p* is true?

CQ2: Is *s* an honest/trustworthy/reliable source?

CQ3: Did *s* assert that *p* is true?

Being unable to provide a satisfactory answer for all the critical questions will result in a potential piece of misinformation or a weaker argument, and we will know exactly the reasons why. In this case, either if *s* is not in position to know about *p*, *s* is not a reliable source, or *s* has not asserted that *p* is true, as it is said in the argument, this will undermine its credibility.

The final category of questions put forward by (Visser & Lawrence, 2022), considers the reasoning structure as a whole, identifying areas of bias where only one side of an argument is being exposed. For example, given a claim which has a number of supporting statements, or a detailed structure of support, but no evidence provided to the contrary, we may wish to ask whether there is any possible counter-argument that could be given. With such a proliferation of potential questions that could be asked, attention must turn to additional aspects, namely: when are particular questions *relevant*? When are relevant questions *important*? How can a dialogue be structured in order to efficiently pose these questions?

The first of these questions involves determining whether a question is applicable given a specific text, and relies on identification of the situations mentioned, namely the proposition types, instances of given argumentation schemes, and determination of the overall argument structure. These tasks all fall under the scope of *Argument Mining*.

Argument Mining (Lawrence & Reed, 2020) broadly refers to the automatic identification and extraction of argument components and structure in natural language text. Lawrence and Reed (2020) break down the related tasks involved in argument mining as shown in Figure 1.

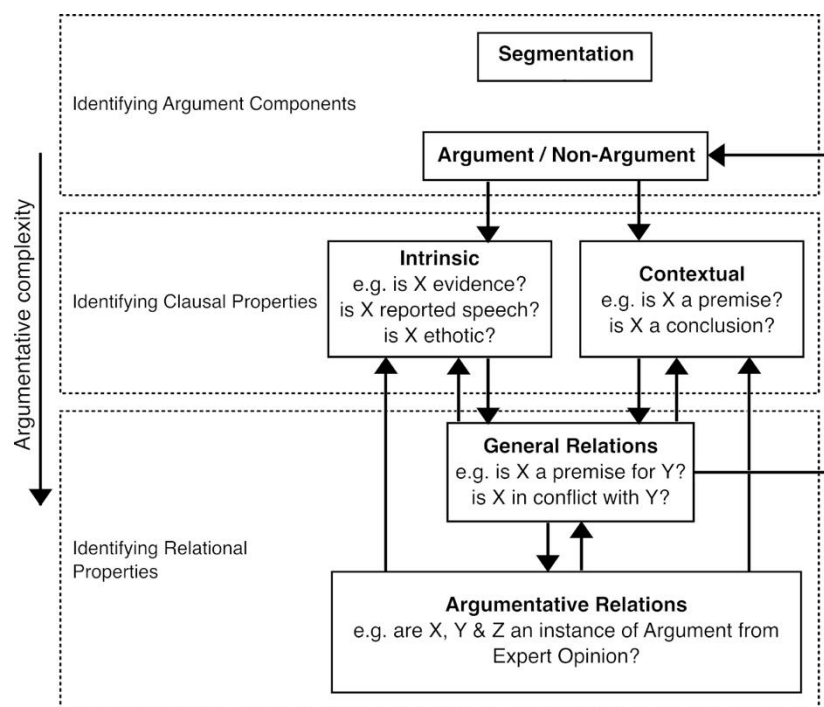


Figure 1. The tasks and levels of complexity in argument mining techniques

Starting from the identification of argument components by segmenting and classifying these as part of the argument being made or not (these tasks are sometimes performed simultaneously, sometimes separated and sometimes the latter is omitted completely), tasks move down through levels of increasing complexity: first considering the role of individual clauses (both intrinsic, such as whether the clause is a policy statement, and contextual such as whether the clause is the conclusion to an argument); and, secondly considering argumentative relations from simple premise/conclusion relationships to whether a set of clauses forms a complex argumentative relation, such as an instance of an argument scheme Figure 1.

Whilst the majority of these tasks have received a high level of attention in the community, the identification of argument schemes falls somewhat behind. Preliminary work supported by the TITAN project (Ruiz-Dolz & Lawrence, 2023) has been carried out in this regard and has achieved solid results in both identifying scheme instances. This work has also explored the ability of natural language processing techniques to approximate the answers to the associated critical questions. In so doing, the work points the way towards narrowing the range of questions to be asked to those around which there is some doubt.

1.3 FURTHER APPROACHES

Further approaches to narrowing the list of potential questions to be asked to those which are important, are presented by (Visser & Lawrence, 2022) employing a combination of properties based on network graph theory (Barabási, 2016) to determine the criticality of each component for the structure as a whole. In particular, using the centrality and divisiveness of each proposition to rank their associated potential questions. In this conceptualization, *central* issues are those that play a particularly important role in the argumentative structure, as can be determined using a variation of the Pagerank algorithm (Brin & Page, 1998) developed by Google to rank search results by importance. Meanwhile, *divisive* issues are those that split opinion and which have points both supporting and attacking them (Konat et al., 2016). The Divisiveness of an issue measures how many others are in conflict with it and the amount of support which the two sides have.

2 CRITICAL THINKING AND MEDIA LITERACY METHODOLOGY

2.1 INTRODUCTION

2.1.1 The Challenge of Navigating Media Complexity & Information Disorder

The modern media landscape is fraught with the unique challenges of Information Disorder, i.e. “a condition in which truth and facts coexist in a milieu of misinformation and disinformation—conspiracy theories, lies, propaganda, and half-truths (USAID, 2021, p.65; see also McIntyre, 2018; Kovach & Rosenstiel, 2022; Wardle & Derakhshan, 2017). Information and cognitive overload overwhelm individuals with the sheer volume of content available (Bawden & Robinson, 2019; Pauley & Gormley, 2023), while misinformation and disinformation are propagated through sophisticated rhetorical, psychological, and digital manipulation tactics (Harjani et al., 2022; Kozyreva et al., 2020; CISA, 2021; Donovan et al., 2023). Social media algorithms create personalized content feeds, reinforcing existing beliefs, ideological segregation and echo chambers (e.g. Guess, Nyhan & Reifler, 2023). Media outlets use sensationalism and clickbait prioritizing eye-catching headlines over factual accuracy to drive engagement, often leading to the distortion of information rather than its verification (Hobbs, 2021; Staender et al., 2021; Jung et al., 2022). Technological manipulation tactics skew public opinion and spread false information through coordinated inauthentic behavior, sockpuppeting, astroturfing, brigading, troll farms and bot networks (Murero, 2023; USAID, 2021; Chan, 2024; Schoch et al., 2022). The rapid spread of information often outpaces the ability to fact-check, the promotion of pseudoscience and conspiracy theories, the influence of economic and

political interests, the role of social media influencers, infotainment, and citizen journalism further distort the accuracy of information and factual reporting (Lewandowsky et al., 2017; Hobbs, 2021; Kovach & Rosenstiel, 2022). Addressing these issues requires a strong focus on media literacy and critical thinking to effectively navigate and interpret today's complex information environment.

2.1.2 Disinformation & Misinformation: Types and Tactics

TITAN focuses on disinformation and low-quality media content that both contribute to the spread of inaccurate and false information. Disinformation' is "information that is (i) false, (ii) communicated as true, and (iii) intentionally communicated as true in order to influence people's beliefs or behavior" (Kwek et al., 2023, p.2; see also van der Linden, 2024). Misinformation refers to the unintentional spread of false, misleading, or incomplete information (Wardle & Derakhshan, 2017). Media Information Disorder (i.e. malinformation, disinformation and misinformation) is problematic content that comes in various types (e.g. Wardie & Derekshan, 2017; Kapantai et al., 2020) (See Figure 2).



Figure 2. The 7 types of Mis- and Disinformation (Wardie & Derekshan, 2017)

Misinformation often occurs due to errors, biases and limitations of the media system (Kovach & Rosenstiel, 2022; Hobbs, 2021). So, while reputable media outlets aim for high journalistic standards (e.g. accuracy and objectivity; Newman et al., 2021), they still face challenges that can affect reporting quality and trust. Over-reliance on press releases and commercial pressures can affect news selection and depth (Lewis et al., 2008; Bunce, 2015). The rapid news cycle may result in hasty and less accurate reporting (Lewis & Cushion, 2009). Infotainment can dilute important issues (Thussu, 2007; Hobbs, 2021), and misleading use of statistics can distort information. False balance and lack of transparency can undermine media credibility.

On the other hand, the term 'tactics of disinformation' refers to the methods or strategies used to disseminate and amplify disinformation intentionally, i.e. it describes how disinformation is spread. Each tactic uses various techniques/methods to intentionally disseminate false information. The most common tactics are the following: Discredit, Emotional Manipulation, Polarization & False Dichotomies/Dilemmas, Impersonation, Conspiracism and Trolling (Harjani et al., 2022; Roozenbeek, Traberg & van der Linden, 2022; Donovan et al., 2023; Kozyreva et al., 2020; European External Action Service, 2024). In addition, there are tactics related to manipulation of evidence, e.g. Cherry-picking, Pseudoscience, Science/Facts Denial (Diethelm & McKee, 2009; Cook, 2015; Hansson, 2008, 2017;) (See Annex 2, Table 2. Tactics of Disinformation).

2.1.3 *Why Critical Thinking and Media Literacy are important*

Recent research strongly supports the idea that critical thinking and an analytic mindset are essential in reducing susceptibility to dis/misinformation. Global initiatives aimed at combating disinformation, such as the Digital Education Action Plan (EC, 2021-27), emphasize the importance of promoting multiple forms of literacy—information, media, and digital literacy—with critical thinking as a central component. Key organizations, including UNESCO, the National Association for Media Literacy Education (NAMLE), the European Association for Viewers Interests (EAVI), The News Literacy Project, the European Commission, the Poynter Institute, the Council of Europe, the Global Alliance for Media and Information Literacy (GAPMIL), and the International Federation of Library Associations and Institutions (IFLA), lead various initiatives, research, and educational programs designed to enhance critical thinking and media literacy. Experts, too, design educational strategies to enhance media literacy and critical thinking (CT) skills, enabling individuals to engage with information meaningfully, promote global citizenship, and counter disinformation (e.g. Hobbs, 2021; Wineburg & McGrew, 2017; Caulfield, 2017; Grizzle et al., 2013).

Several studies demonstrate that media literacy, inoculation, and priming individuals to analytical and critical thinking are the most effective interventions against information disorder. Inoculation strategies are based on the principle that "prevention is the best cure": by presenting counterarguments before participants are exposed to conspiratorial media, they become significantly more resistant to adopting conspiracy beliefs (e.g. van der Linden & Roozenbeek, 2024). Also, participants who were primed to think analytically exhibited a lower likelihood of believing in conspiracy theories compared to those in the control group, especially when the interventions provided explicit instruction in critical thinking skills for evaluating conspiracy theories or pseudoscience (O'Mahony et al., 2023; Swami, 2014; Lutzke et al., 2019; Horn and Veermans, 2019; Lantian et al., 2021; Dyer & Hall, 2019; Wilson, 2018). Moreover, media literacy programs that focus on critical thinking (e.g. showing how to evaluate scientific evidence or how to discern real science from pseudoscience) have been shown to reduce the spread of misinformation (e.g. Lu et al., 2024; Jones-Jang et al., 2021;). Despite some reservations (e.g. Večkalov et al., 2024), the above studies highlight that critical thinking enhances media literacy, i.e. the ability to assess the credibility of authors and news outlets, evaluate evidence and sources, detect media biases, become aware of personal bias, understand narrative framing, recognize logical fallacies and rhetorical tricks, etc. The studies also show that critical thinking, especially, reflective thinking, serves as a crucial defense mechanism against cognitive bias and other psychological tactics leveraged by disinformation agents (e.g. Stanovich & West, 2008; Wooley & Joseff, 2020; Ecker et al., 2022).

2.1.4 *What Media Literacy Is*

Media literacy involves understanding how media messages are constructed, recognizing biases and propaganda, and assessing the impact of media on individuals and society (Hobbs, 2021; Potter, 2018). It equips individuals with the tools to identify the tactics used in misinformation, such as narrative framing, logical fallacies, rhetorical tricks, psychological tactics, identity fraud, and digital manipulation techniques. Additionally, media literacy offers specific skills for verifying information in the digital age. It teaches individuals to cross-check facts, assess the authenticity of images and videos, and identify deepfakes—tasks that are beyond the scope of generic CT alone (Wardle & Derakhshan, 2017). Also, prebunking and critical thinking are interconnected in media literacy efforts. Prebunking provides the foundation of knowledge needed to recognize and reject misinformation by informing individuals about common disinformation tactics and how they work. This knowledge enhances critical thinking skills by offering specific strategies and examples that individuals can use to evaluate and question the information they encounter.

2.1.5 *What Critical Thinking Is: A Common Ground Definition*

Critical thinking is a multifaceted concept with numerous definitions across disciplines, including psychology, philosophy, and education, yet several core elements consistently emerge. The key common elements of most definitions are purposefulness, rigorous evaluation and analysis, logical

reasoning, and self-regulation (e.g. Paul & Elder, 2006; Ennis, 1991; Halpern, 1998, 2014; Scriven & Paul, 1987; APA Delphi Report, 1990; Facione, 1990). Purposefulness guides the direction of thinking, evaluation and analysis ensure the accuracy and reliability of information, logical reasoning structures the thought process, and self-regulation maintains the overall quality of thinking. First, CT is defined as "purposeful, reasoned, and goal-directed thinking" (Halpern, 1998, p. 4) and "reasonable, reflective thinking focused on deciding what to believe or do" (Ennis, 1991, p. 11). Second, evaluation and analysis are crucial for verifying the accuracy and reliability of information. As the American Philosophical Association (1990) explains, critical thinking is "purposeful, self-regulatory judgment" that involves "interpretation, analysis, evaluation, and inference." This rigorous approach helps ensure that judgments are well-supported by credible evidence. Third, effective evaluation and analysis of information and arguments relies on reasoning and logic, i.e. critical thinkers must apply logical principles and structured reasoning (Paul & Elder, 2006). This involves using deductive reasoning to ensure conclusions logically follow from premises, inductive reasoning to assess the strength of arguments based on evidence, and identifying and avoiding logical fallacies for maintaining the validity of arguments. Fourth, reflective thinking and self-regulation are meta-cognitive strategies that allow individuals to actively monitor their thought patterns, recognize their biases, and make necessary adjustments (e.g. Facione, 2015). These strategies ensure that thinking remains clear, coherent, and aligned with intellectual standards.

2.2 CRITICAL THINKING PRINCIPLES USED BY TITAN

2.2.1 Critical Thinking requires Generic Critical Thinking Abilities

The seminal critical thinking models, including Halpern's Model (Halpern Critical Thinking Assessment, 2007), Facione's Model (California Critical Thinking Skills Test, 1990, 1994), the Paul-Elder Framework (2001), Watson-Glaser Critical Thinking Appraisal (1941; 1980 and 1994), RED Model (Foundation for Critical Thinking, 2009), and Cornell Critical Thinking Test (Ennis & Millman, 1961; 1985; 2005), share the following common abilities (See Table 1):

Table 1. Critical Thinking Dimensions

Critical Thinking Dimension	Definition
Argument Analysis	The ability to analyze the structure of arguments and evaluate their validity, focusing on their logical coherence
Inference	The ability to draw logical conclusions based on the available evidence, evaluate the strength of inferences, and consider other possible alternative explanations.
Decision Making	The ability to choose the best option by weighing options, evidence and predicting what might happen
Identifying Assumptions	The ability to recognize and question the hidden assumptions behind statements or arguments to see if they hold up

Evaluation of Evidence	The ability to assess the quality, relevance, and credibility of evidence used to support a claim or conclusion
Metacognition	Being aware of one's thoughts, including reflecting on and improving one's own reasoning process

2.2.2 Critical Thinking requires a Critical Mindset/Dispositions

There is a general agreement among psychologists and educators that critical thinking encompasses not only a range of abilities and skills but also a specific mindset of intellectual dispositions that influence thinking performance (Paul & Elder, 2006; Halpern, 2014; Hitchcock, 2022; Facione, 2015; Butler & Halpern, 2020; Ku & Ho, 2010). In particular, critical thinking skills determine how well a person can manage a thinking task, while a critical mindset determines whether one is willing to engage thoroughly in the thinking process. Studies show that simply possessing critical thinking skills without actively seeking the truth or investing cognitive effort offers no benefit compared to those who lack these skills (Butler & Halpern, 2020; Facione, 2015; Rand & Pennycook, 2019; Kwek et al., 2023) (see Table 2).

Table 2. Critical Thinking Disposition

Critical Thinking Disposition	Definition
Inquisitiveness	The tendency to be curious and eager to learn more, asking questions and seeking out new information to gain deeper understanding.
Judiciousness	The ability to make well-considered decisions, carefully weighing evidence and potential outcomes before arriving at a conclusion.
Truth Seeking	The commitment to seek out and evaluate evidence impartially, and to be willing to revise one's beliefs in the face of new evidence
Confidence in Reasoning	The belief in one's own ability to reason effectively and make sound judgments, while remaining open to others' perspectives
Open-Mindedness	The willingness to consider and evaluate alternative viewpoints and perspectives, and to be receptive to new and diverse ideas
Analyticity	The ability to critically analyze information, identify logical connections, and assess the validity and relevance of arguments.
Systematicity	The tendency to approach problems and decisions in a structured and methodical manner, ensuring thoroughness and consistency.
Autonomy	The capacity to think independently, make decisions based on one's own judgment, and be self-directed in one's reasoning and learning

Intellectual Courage	The willingness to face and address challenging ideas and viewpoints, even when it involves personal discomfort or risk.
Intellectual Humility	The recognition and acceptance of the limits of one's knowledge and being open to others' viewpoints and corrections
Intellectual Integrity	The commitment to hold oneself to the same standards of evidence and reasoning that one applies to others, maintaining honesty and fairness in the pursuit of truth

2.2.3 Critical Thinking requires Domain-specific Knowledge

Generic critical thinking skills are foundational and transferable across various fields (Ennis, 1987; Halpern, 1998, 2014; Butler & Halpern, 2020). These skills improve with practice in diverse contexts (Butler & Halpern, 2020). However, domain-specific knowledge and understanding specific domain-related issues and biases are crucial for effective reasoning and decision-making (McPeck, 1981, 1990; Bailin et al., 1999; Facione, 2015). Bailin et al. (1999) argue that critical thinking involves integrating domain-specific knowledge rather than relying solely on generic skills. As Facione (2015) underscores, "critical thinking abilities are not a magic elixir that can be applied to any issue whatsoever by somebody who has no knowledge of the facts relevant to exploring that issue." For instance, understanding medical principles is essential for effective critical thinking in healthcare. A doctor must not only apply general analytical skills but also have in-depth knowledge of medical conditions, treatments, and procedures to make sound clinical decisions (McPeck, 1990). Therefore, effective critical thinking requires a blend of general skills, dispositions, and substantial domain-specific knowledge (Bailin et al., 1999; Facione, 2015).

2.2.4 Navigating Media Requires Media Critical Thinking (MCT)

Navigating the modern media landscape requires more than just generic critical thinking (CT) skills due to the unique and complex challenges posed by today's media environment. While generic CT provides a foundation for logical reasoning, evidence evaluation, self-reflection and questioning assumptions, it falls short in addressing the overwhelming volume and rapid dissemination of information, the challenges of misinformation and disinformation, the algorithmic biases, emotional manipulation, deep fakes, and the complexity of user-generated content (Wardle & Derakhshan, 2017). These challenges necessitate domain-specific knowledge, which is where media literacy comes in.

Critical thinking for media, or Media Critical Thinking (MCT) hereafter, addresses these limitations by integrating generic CT skills with media literacy and prebunking strategies, creating a more robust approach to evaluating media content. In MCT, media literacy and critical thinking are deeply interconnected. Critical thinking is an indispensable part of media literacy programs, and vice versa (Hobbs, 2021; Jowett & O'Donnell, 2018; Caulfield, 2017; Rheingold, 2012; Wineburg & McGrew, 2017; boyd, 2014;). Media literacy provides the tools to access, analyze and evaluate media messages, while critical thinking ensures that these processes are carried out with rigor and a focus on logic/rationality and evidence. For example, generic CT skills can help individuals assess the logical consistency of an argument and media literacy provides the context needed to evaluate the credibility of the source and the potential biases involved. Also, a critical thinker trained in media literacy can scrutinize a news article not just for its logical structure but also for how it might be influenced by corporate interests or political agendas and how reputable sources/outlets differ from unreliable media. In addition, media-literate users can recognize when a story uses emotionally charged language to sway public opinion or when it presents information in a misleading way to attract clicks (Hobbs, 2011;). Moreover, when encountering a viral social media post, a media-literate individual will

use CT skills to analyze the evidence and check the source's credibility, while also applying media literacy knowledge to recognize potential manipulation techniques and misinformation.

2.2.4.1 Media Critical Thinking (MCT) Skills & Dispositions

MCT is a blend of generic critical thinking (CT) and Media Literacy. Generic CT includes skills in interpretation, argument analysis, inference, identifying assumptions, evaluation of evidence and metacognition (e.g. Facione, 1990, 2015). These are the core skills required to (1) recognize, analyze and evaluate media messages, evidence and arguments, in order (2) to detect the most common logical fallacies and rhetorical ploys used in dis/misinformation. However, the critical media consumer also needs a set of further critical thinking skills required (3) to judge the reliability of experts, sources, and evidence, (4) to analyze narrative framing and disinformation narratives, (5) to detect deceptive persuasion tactics, including digital manipulation tactics, and (6) to recognize media bias, agendas, and political economic influences (see Table 3).

Table 3. Generic & Media CT Skills

Socratic questions & Facione's CT skills	MCT skills	MCT Subskills
Clarification / INTERPRETATION	Interpreting/making sense of media messages by categorizing and comprehending the meaning of information, concepts and ideas within contexts.	Grasping the meaning of words, phrases, terms and concepts (including connotations, metaphors, analogies, subtext, wordplay) within the context in which they are presented; interpreting data/statistics, visual elements (e.g., graphs, charts, images), tone and style, etc.
Probing Assumptions / ANALYSIS	Breaking down media information into its component parts, such as reasons, claims, arguments, data, and assumptions.	Analyze a media message/narrative to identify the foundational assumptions and narrative techniques behind it, and understand how these assumptions and other elements contribute to the overall argument or narrative.
Probing Rationale, Reasons, Evidence / EVALUATION	Assessing the credibility of statements and the quality of reasoning/arguments, determining if claims are supported and arguments are logically valid.	Identifying fallacies or errors in reasoning that may undermine the argument, claim or evidence; recognize narrative framing, such as selective emphasis/omission, storytelling techniques, emotional appeals, imagery, language, agenda-framing, contextualization.
Questioning Viewpoints and Perspectives / INFERENCE	Drawing reasonable conclusions based on evidence, recognizing what follows logically from the available data.	Assessing the validity, completeness, accuracy, relevance of evidence/data, distinguishing opinions from facts.

Questioning the Question / SELF-REGULATION (METACOGNITION)	Reflecting on and adjusting one's thinking and biases when evaluating media content	Using metacognition to recognize the psychological drivers of disinformation, reflecting on why information feels convincing, awareness of personal habits, thoughts and biases, improving reasoning processes.
Media-specific CT skills and dispositions	Identify deceptive persuasion tactics (mainly rhetorical tactics)	Identifying emotional triggers in word choice; recognizing framing; detecting hyperbole, irony, sarcasm, rhetorical questions,
	Evaluate the credentials of authors and outlets	Assessing track record, reputation, expert credentials, online profiles, financial aspects, purpose, agenda, biases
	Understand media agendas/media bias, political and economic influences	
	Identify digital manipulation tactics	Recognizing signs of digital manipulation, identity fraud and impersonation, coordinated inauthentic behavior, sock puppets, troll farms, manufactured amplification, clickbait, flooding.
	Constructive Skepticism	Engaging in a reasoned critique of the media without falling into cynicism, which can lead to distrust and nihilism.

2.2.5 Prebunking & Inoculation: "Prevention is the best cure"

Researchers have found that preventing disinformation from spreading is more effective than trying to debunk it after it has been circulated (van der Linden & Roozenbeek, 2024). Disinformation science calls this preventive approach prebunking (or 'preemptive debunking'). This is a broad concept that involves taking proactive measures to remind individuals to think carefully before sharing their thoughts publicly (van der Linden, 2022). Prebunking grew from dissatisfaction with the 'therapeutic' corrections of debunking because of the danger of "continued influence of misinformation" after individuals had been exposed to it (van der Linden & Roozenbeek, 2024). Inspired by the social psychologist William McGuire (1964), scientists experimented and devised prebunking as an alternative strategy against disinformation. McGuire's Inoculation Theory draws from immunology to create a psychological "vaccine" against manipulative persuasion and brainwashing. Similar to medical vaccines that provide physiological resistance to future infections, psychological inoculation offers resistance to attempts at manipulating attitudes. This protection shields individuals from being deceived or persuaded by misleading content (Harjani et al., 2022). Studies indicate that individuals can be mentally immunized against disinformation, in a manner akin to vaccination against viral infections (Traberg et al., 2024). In sum, prebunking aims to counteract potential disinformation before it becomes firmly established (Cook, Lewandowsky & Ecker, 2017).

2.2.5.1 How Inoculation works

In practical terms, inoculation involves exposing individuals preemptively to a weakened version or small 'dose' of a disinformative argument/content—strong enough to activate defensive responses but not strong enough to persuade (Compton, 2013). This is followed by refuting the argument, thereby fortifying resistance against future, full dose disinformative persuasion attempts (Papageorgis & McGuire, 1961). Inoculation messages equip people with counter arguments that can potentially confer resistance to disinformation, even if the disinformation aligns with their existing beliefs (Cook et al., 2017). The process of inoculation (Figure 3) operates effectively in two stages: first, forewarning individuals about upcoming exposure to misleading messages, and second, providing tools for preemptive refutation such as counter arguments and a simplified example of the misleading content. Prebunking interventions can take passive forms like videos, audio, infographics, or text, as well as active forms such as quizzes or games, each with its own strengths and weaknesses (Harjani et al., 2022). However, McGuire suggested early on that a more active inoculation approach, where individuals generate their own arguments both for and against, could be particularly more effective (van der Linden & Roozenbeek, 2021).

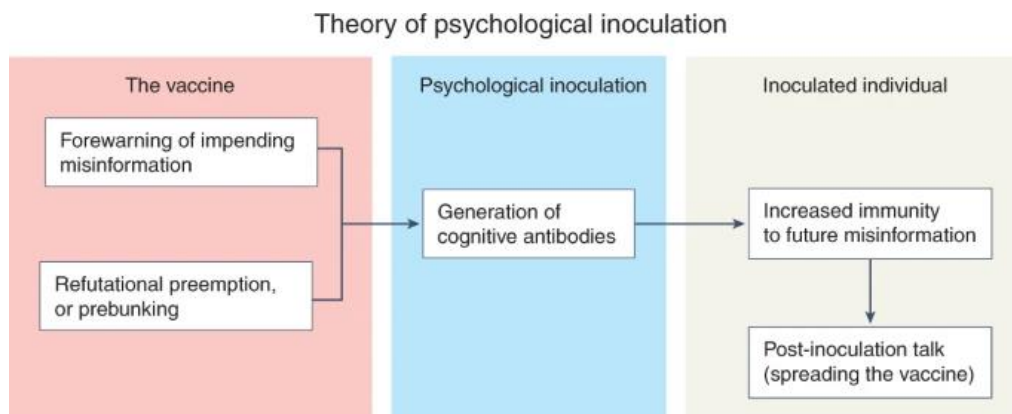


Figure 3. Theory of psychological inoculation (Maertens, R. 2022)

2.2.5.2 How Prebunking Supports Critical thinking and Media Literacy

Prebunking and critical thinking are interconnected in evaluating media information (e.g. Cook, Lewandowsky & Ecker, 2017). Prebunking is a proactive application of critical thinking. It involves recognizing misinformation tactics (e.g., emotional manipulation, logical fallacies). It fosters analytical thinking and skepticism, i.e. questioning and analyzing the information presented. Also, it encourages cognitive reflection on why certain information is false. In sum, it is a targeted approach that proactively builds readiness to apply critical thinking before encountering deceptive content. Also, prebunking equips individuals with specific insights into how disinformation is constructed and spread and how it can be critically refuted. Thus, people are better prepared to apply their critical thinking skills in order to question and evaluate media information. Finally, prebunking helps develop media literacy by addressing specific vulnerabilities to dis/misinformation. It teaches individuals what to look out for and how to anticipate potential deceptive strategies.

2.2.6 Media Critical Thinking Rests on the Socratic Method

Media Critical Thinking is essentially Socratic since it uses self-reflective thinking, media-specific and generic critical thinking skills to analyze and evaluate media content, thus addressing challenges like dis/misinformation, bias, and the complexity of digital media. Modern critical thinking education uses Socratic questioning to develop these skills through open-ended questions that encourage exploration and reflection. Paul and Elder's (2016) framework are a significant modern adaptation, emphasizing disciplined questioning to foster deep, critical inquiry, similar to Socrates' original method.

The original Socratic Method, as depicted in Platonic dialogues, was a critical inquiry technique which sought to uncover truths about morality and the good life (Morrison, 2011). Socrates sought to restore objectivity in knowledge and ethics, challenging the relativistic views of the Sophists. Socratic questioning was a dialectical approach. Using a question-and-answer format, he examined interlocutors' moral knowledge, encouraging them to define concepts, such as justice and courage, and pushing them to develop their understanding through rigorous questioning. This method involved focusing on definitions to clarify key concepts, probing the assumptions underlying the interlocutor's beliefs, and exposing contradictions in their arguments.

Socrates likened himself to a midwife, helping others discover their own knowledge through this questioning process, known as the Maieutic Method. He also employed the technique of Elenchus, where he would challenge answers that led to contradictions, often resulting in a state of *aporia*, or puzzlement, thus indicating the need for further inquiry and deeper thinking. Central to this method was Socratic Ignorance, where Socrates professed his own lack of knowledge to encourage deeper reflection and understanding. Through Socratic Irony and humility, Socrates claimed to know nothing, creating a non-threatening environment that allowed interlocutors to reflect openly on their beliefs.

In CT pedagogy, the modern adaptation of the Socratic method is a dialogical method that focuses on practical critical thinking, promoting questioning, analysis, and self-reflection, with a less adversarial tone. Teachers act as facilitators, guiding discussions and encouraging deep thinking.

The modern approach involves open questions, follow-up queries, and debate to help students refine their understanding. For instance, discussions on justice begin with definitions and evolve through probing questions aiming to reveal possible inconsistencies. Paul and Elder (2020) have developed a Socratic framework for media critical thinking ensuring relevance in today's information world. Their aim is to systematically analyze and evaluate media information and sources using core questions to find out if they meet essential intellectual standards, such as logic, relevance, clarity and accuracy. In particular, the questions are designed to spark critical thinking by examining underlying media biases, assumptions, disinformation techniques often embedded in news stories, advertisements, and political discourse. Their framework includes questions for clarification to ensure understanding, questions that probe assumptions to examine underlying beliefs, questions that probe reasons and evidence to assess support for claims, questions about viewpoints and perspectives to explore alternative views, questions that probe implications and consequences to consider the impact of information. It further includes questions about the origin and purpose of information to uncover motives, questions that challenge the status quo to question commonly accepted beliefs, and questions that encourage self-reflection to address personal biases.

3 CRITICAL THINKING ASSESSMENT (CTA) METHODOLOGY

The main objective of the critical thinking assessment methodology within the TITAN ecosystem was reached following several operational steps, described in the following workflow chart (Figure 4).

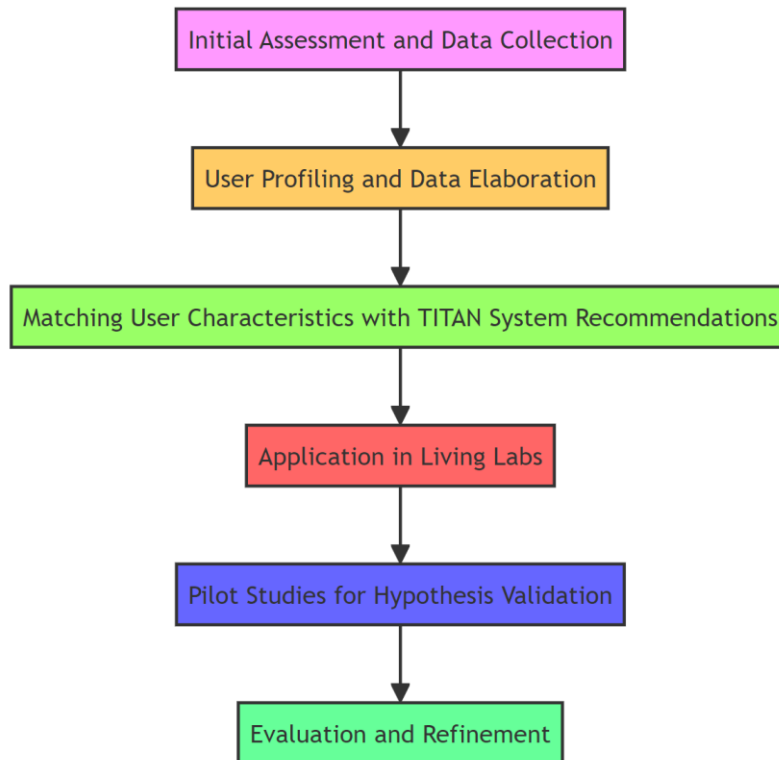


Figure 4. The overarching research process followed in the project

3.1 MAIN OBJECTIVE AND SCOPE

The main objective of the Critical Thinking Assessment (CTA) methodology within the TITAN project is to develop a comprehensive framework to systematically evaluate the critical thinking skills of individuals in the context of disinformation. The scope of this methodology includes the identification of cognitive profiles for each of the six CT skills (i.e., the six assessment dimensions).

The measurement of these skills allows the development of interventions to improve them. The CTA aims to address the cognitive factors that influence critical thinking, ensuring a fine-grained approach to understanding and improving how individuals interact with disinformation. Facione's six-dimension model for critical thinking (CT) assessment (1990) and the Critical Thinking Assessment (CTA) methodology within the TITAN project complement each other effectively. Facione's model, which provides detailed measurements of CT across multiple facets, not only offers an overall CT score but also breaks down specific competencies integral to critical thinking. This detailed approach is in line with the CTA's objective in the TITAN project, which is to develop a thorough framework for assessing critical thinking, particularly in the context of disinformation. The CTA focuses on identifying individual cognitive profiles based on the five CT skills, paralleling Facione's dimensions. This alignment suggests that integrating Facione's model with the CTA methodology could lead to a comprehensive, multidimensional tool for assessing critical thinking. Such a tool would measure critical thinking skills and provide insights into their application in understanding and countering disinformation.

3.2 DEFINITION OF THE CRITICAL THINKING ASSESSMENT MODEL AGAINST DISINFORMATION

In 1990, Peter A. Facione conducted a pioneering Delphi study, which has since been recognized as the gold standard in the field for evaluating and assessing the construct of critical thinking. This seminal research established a foundational framework and set forth critical benchmarks for the systematic analysis and measurement of critical thinking abilities (Facione, 1990). Critical thinking is:

“purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. CT is essential as a tool of inquiry. As such, CT is a liberating force in education and a powerful resource in one's personal and civic life. While not synonymous with good thinking, CT is a pervasive and self-rectifying human phenomenon”. Then, the ideal critical thinker is “habitually 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. Thus, educating good critical thinkers means working toward this ideal. It combines developing CT skills with nurturing those dispositions which consistently yield useful insights and which are the basis of a rational and democratic society” (Facione, 1990).

The scientific consensus has focused on six dimensions of critical thinking, namely critical thinking cognitive skills and sub-skills (Figure 5). These are (quoting from the original article):

[1] Interpretation

To comprehend and express the meaning or significance of a wide variety of experiences, situations, data, events, judgments, conventions, beliefs, rules, procedures or criteria.

[1.1] Categorization

To apprehend or appropriately formulate categories, distinctions, or frameworks for understanding, describing or characterizing information; to describe experiences, situations, beliefs, events, etc. so that they take on comprehensible meanings in terms of appropriate categorizations, distinctions, or frameworks.

[1.2] Decoding Significance

To detect, attend to, and describe the informational content, affective purport, directive functions, intentions, motives, purposes, social significance, values, views, rules, procedures, criteria, or inferential relationships expressed in convention-based communication systems, such as in language, social behaviors, drawings, numbers, graphs, tables, charts, signs and symbols.

[1.3] Clarifying Meaning

To paraphrase or make explicit, through stipulation, description, analogy or figurative expression, the contextual, conventional or intended meanings of words, ideas, concepts, statements, behaviors, drawings, numbers, signs, charts, graphs, symbols, rules, events or ceremonies; to use stipulation, description, analogy or figurative expression to remove confusing, unintended vagueness or ambiguity, or to design a reasonable procedure for so doing.

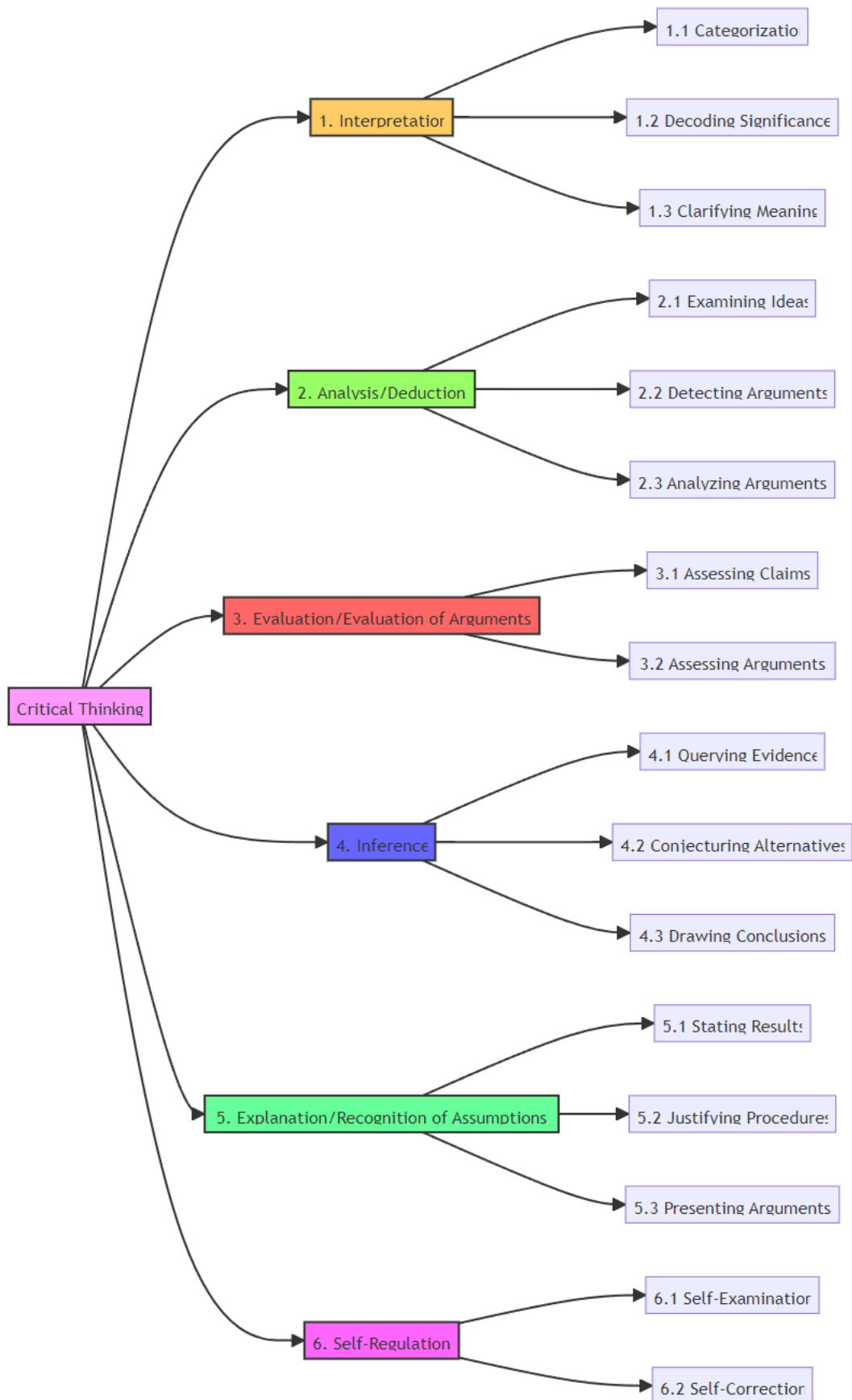


Figure 5. The six dimensions of critical thinking, namely critical thinking cognitive skills and sub-skills.

[2] Analysis

To identify the intended and actual inferential relationships among statements, questions, concepts, descriptions or other forms of representation intended to express beliefs, judgments, experiences, reasons, information, or opinions.

[2.1] Examining Ideas

To determine the role various expressions, play or are intended to play in the context of argument, reasoning or persuasion; to define terms; to compare or contrast ideas, concepts, or statements; to identify issues or problems and determine their component parts, and also to identify the conceptual relationships of those parts to each other and to the whole.

[2.2] Detecting Arguments

Given a set of statements, descriptions, questions or graphic representations, to determine whether or not the set expresses, or is intended to express, a reason or reasons in support of or contesting some claim, opinion or point of view.

[2.3] Analyzing Arguments

Given the expression of a reason or reasons intended to support or contest some claim, opinion or point of view, to identify and differentiate: [a] the intended main conclusion, [b] the premises and reasons advanced in support of the main conclusion, [c] further premises and reasons advanced as backup or support for those premises and reasons intended as supporting the main conclusion, [d] additional unexpressed elements of that reasoning, such as intermediary conclusions, unstated assumptions or presuppositions, [e] the overall structure of the argument or intended chain of reasoning, and [f] any items contained in the body of expressions being examined which are not intended to be taken as part of the reasoning being expressed or its intended background.

[3] Evaluation

To assess the credibility of statements or other representations which are accounts or descriptions of a person's perception, experience, situation, judgment, belief, or opinion; to assess the logical strength of the actual or intended inferential relationships among statements, descriptions, questions or other forms of representation.

[3.1] Assessing Claims

To recognize the factors relevant to assessing the degree of credibility to ascribe to a source of information or opinion; to assess the contextual relevance of questions, information, principles, rules or procedural directions; to assess the acceptability, the level of confidence to place in the probability or truth of any given representation of an experience, situation, judgment, belief or opinion.

[3.2] Assessing Arguments:

To judge whether the assumed acceptability of the premises of a given argument justify one's accepting as true [deductively certain], or very probably true [inductively justified], the expressed conclusion of that argument; to anticipate or to raise questions or objections, and to assess whether these point to significant weakness in the argument being evaluated; to determine whether an argument relies on false or doubtful assumptions or presuppositions and then to determine how crucially these affect its strength; to judge between reasonable and fallacious inferences; to judge the probative strength of an argument's premises and assumptions with a view toward determining the acceptability of the argument; to determine and judge the probative strength of an argument's intended or unintended consequences with a view toward judging the acceptability of the argument; to determine the extent to which possible additional information might strengthen or weaken an argument.

[4] Inference

To identify and secure elements needed to draw reasonable conclusions; to form conjectures and hypotheses; to consider relevant information and to deduce the consequences flowing from data,

statements, principles, evidence, judgments, beliefs, opinions, concepts, descriptions, questions, or other forms of representation.

[4.1] Querying Evidence

In particular, to recognize premises which require support and to formulate a strategy for seeking and gathering information which might supply that support; in general, to judge that information relevant to deciding the acceptability, plausibility or relative merits of a given alternative, question, issue, theory, hypothesis, or statement is required, and to determine plausible investigatory strategies for acquiring that information.

[4.2] Conjecturing Alternatives

To formulate multiple alternatives for resolving a problem, to postulate a series of suppositions regarding a question, to project alternative hypotheses regarding an event, to develop a variety of different plans to achieve some goal; to draw out presuppositions and project the range of possible consequences of decisions, positions, policies, theories, or beliefs.

[4.3] Drawing Conclusions

To apply appropriate modes of inference in determining what position, opinion or point of view one should take on a given matter or issue; given a set of statements, descriptions, questions or other forms of representation, to educe, with the proper level of logical strength, their inferential relationships and the consequences or the presuppositions which they support, warrant, imply or entail; to employ successfully various sub-species of reasoning, as for example to reason analogically, arithmetically, dialectically, scientifically, etc.; to determine which of several possible conclusions is most strongly warranted or supported by the evidence at hand, or which should be rejected or regarded as less plausible by the information given.

[5] Explanation

To state the results of one's reasoning; to justify that reasoning in terms of the evidential, conceptual, methodological, criteriological and contextual considerations upon which one's results were based; and to present one's reasoning in the form of cogent arguments.

[5.1] Stating Results

To produce accurate statements, descriptions or representations of the results of one's reasoning activities so as to analyses, evaluate, infer from, or monitor those results.

[5.2] Justifying Procedures

To present the evidential, conceptual, methodological, criteriological and contextual considerations which one used in forming one's interpretations, analyses, evaluation or inferences, so that one might accurately record, evaluate, describe or justify those processes to one's self or to others, or so as to remedy perceived deficiencies in the general way one executes those processes.

[5.3] Presenting Arguments

To give reasons for accepting some claim; to meet objections to the method, conceptualizations, evidence, criteria or contextual appropriateness of inferential, analytical or evaluative judgments.

[6] Self-Regulation

Self-consciously to monitor one's cognitive activities, the elements used in those activities, and the results educed, particularly by applying skills in analysis and evaluation to one's own inferential judgments with a view toward questioning, confirming, validating, or correcting either one's reasoning or one's results.

[6.1] Self-Examination

To reflect on one's own reasoning and verify both the results produced and the correct application and execution of the cognitive skills involved; to make an objective and thoughtful meta-cognitive self-assessment of one's opinions and reasons for holding them; to judge the extent to which one's thinking is influenced by deficiencies in one's knowledge, or by stereotypes, prejudices, emotions or

any other factors which constrain one's objectivity or rationality; to reflect on one's motivations, values, attitudes and interests with a view toward determining that one has endeavored to be unbiased, fair-minded, thorough, objective, respectful of the truth, reasonable, and rational in coming to one's analyses, interpretations, evaluations, inferences, or expressions.

[6.2] Self-Correction

Where self-examination reveals errors or deficiencies, to design reasonable procedures to remedy or correct, if possible, those mistakes and their causes.

The sixth dimension, "Self-Regulation," is so general and psychologically impactful that it is considered a fundamental prerequisite for all the others and cannot be measured in the same way (Nosich, 2012). This dimension requires a different type of assessment; concurrently, preliminary research is being conducted specifically to better understand the impact of this variable, which is more closely related to the individual (Machete & Turpin, 2020).

Critical thinking plays a pivotal role in discerning the truth amid a deluge of information, some of which may be deliberately false or misleading. For example, in the context of social media, critical thinking helps individuals to evaluate the credibility of sources, question the validity of arguments, and distinguish between facts and opinions. In the realm of news consumption, it enables people to critically assess the reliability of various news outlets and discern biases or hidden agendas. Furthermore, in everyday decision-making, critical thinking aids in recognizing and rejecting fallacious reasoning and propaganda, thereby fostering informed and rational judgments. Therefore, it was necessary to develop a new tool for evaluating critical thinking based on the dimensions (just presented) validated by scientific literature, making the items openly accessible to the scientific community. As a result, it became necessary to avoid commercial instruments in order to maintain control over every step of the research project, its models, and results. The objective, in this specific case, is to focus the assessment on critical thinking in relation to information disorders. This represents a crucial degree of innovation; never before has something similar been realized, which consequently meant that we could not rely on previous scientific precedents. Within the framework of the research project, we will attempt to test the designed approach through living labs and two iterations of piloting. This will allow for the measurement of a "specific" type of critical thinking.

A fundamental premise of the project is that the Socratic method, known for its dialogic and question-centric approach, encourages deep analysis, reflection, and the challenging of assumptions, all of which are key components of critical thinking (Spector & Ma, 2019). By fostering an environment where questioning and interactive dialogue are central, the method aims to cultivate a higher level of cognitive engagement and analysis, thereby enhancing an individual's ability to critically assess and interpret information (Overholser & Beale, 2023). This aligns with the project's goals of developing and reinforcing critical thinking, particularly in contexts that demand rigorous evaluation and discernment of information.

The general six steps of the Socratic method are (see Figure 6):

[1] Clarification: This involves posing insightful questions to gain a comprehensive and unambiguous understanding of the ideas or statements presented. The aim is to dissect complex or vague assertions, thereby unraveling any ambiguities or misunderstandings that might exist. This step is crucial for ensuring that all participants in a discussion have a common understanding of the subject matter, thus laying a solid foundation for meaningful dialogue.

[2] Probing Assumptions: This step is centered on critically examining the underlying assumptions or presuppositions that form the basis of a person's beliefs or arguments. It involves questioning the validity and origin of these assumptions and considering the possibility of alternative, perhaps more valid, presuppositions. This exploration not only challenges the foundational aspects of arguments but also opens up avenues for new perspectives and insights.

[3] Probing Rationale: In this stage, the focus is on delving into the reasoning and evidence that support a person's claims. By questioning the logic and examining the evidence presented, this approach seeks to uncover any weaknesses, inconsistencies, or gaps in the argument. This critical examination is pivotal in assessing the soundness and credibility of the arguments being made.

[4] Questioning Viewpoints: This entails asking individuals to further elaborate on their opinions and viewpoints. It encourages a deeper exploration of their perspectives, prompting them to provide more detailed explanations and justifications. This not only aids in a better understanding of their viewpoints but also helps in identifying any biases or subjective influences that might be shaping their opinions.

[5] Probing Consequences: Here, the emphasis is on asking individuals to consider the broader implications and potential consequences of their beliefs or claims. This includes contemplating the short-term and long-term effects, as well as the impact on different stakeholders or scenarios. By doing so, it promotes a more holistic understanding of the issues at hand, encouraging a foresighted approach in thinking and decision-making.

[6] Challenging the Question (Questions on the Question): This involves inviting individuals to reflect on the entire process of questioning and examination itself, including the reasoning and motives behind the responses given. It encourages a meta-cognitive approach where individuals not only consider the content of the discussion but also the methods and approaches used in the questioning process. This reflection can lead to a deeper understanding of the inquiry process and the dynamics of critical thinking.

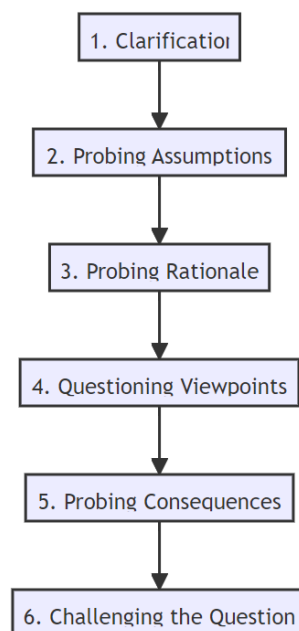


Figure 6. The Socratic method flow

Innovation operates here on two levels: on one hand, our focus is not merely on measuring critical thinking in general, but specifically critical thinking in relation to misinformation. This approach is directed towards aligning educational content precisely with the distinct cognitive needs and aptitudes of each user, thereby enhancing the effectiveness of the learning experience. In this case, the identification of specific profiles of individuals variably susceptible to misinformation or challenges

in applying critical thinking necessitated the measurement of three more general socio-cognitive dimensions. On these dimensions, it is possible to identify groups of citizens by analogy. These variables extend beyond fixed demographic attributes, which may pose risks of profiling, and instead concentrate on individual psychological and behavioral dimensions. Collectively, these variables offer a comprehensive perspective on an individual's critical thinking and information processing abilities. They facilitate a more detailed understanding of how individuals interact with, interpret, and share information, thus providing valuable insights into their susceptibility to misinformation and their proficiency in critical thinking. In detail, the variables used to refer to different user groups in a customized way are the **Propensity to engage in Analytical Reasoning**, for assessing an individual's inclination to approach problems reflexively, avoiding impulsive decisions evaluating the tendency to reflect on issues in a deep, analytical manner, thus determining the likelihood of an individual engaging in thoughtful, reasoned decision-making rather than relying on intuitive, immediate responses. Then, the **Value placed on Truthfulness (VT)** assesses how much importance an individual attributes to knowledge, learning, and truthfulness, especially in challenging situations. It acts as a barometer of a person's dedication to accurate and critical thinking. VT is pivotal in understanding an individual's commitment to discerning and upholding truth, which is especially relevant in the context of information evaluation and decision-making processes. The **Regulatory Self-Efficacy in Sharing Misinformation (RSSM)** is aligned with social cognitive theories, notably those of Bandura. The RSSM assessment encompasses both inhibitory and active factors in managing online disinformation. The inhibitory aspect, consisting of four items, assesses a person's ability to refrain from sharing news for personal gain or under impulsive tendencies. The active component, also with four items, evaluates the perceived capacity for proactively scrutinizing the accuracy of news items. This dual approach reflects an individual's overall self-efficacy in regulating their interactions with and dissemination of information in a digital context.

3.3 PROPOSAL FOR THE TECHNICAL IMPLEMENTATION

In the next sections the proposed steps to be followed to implement the critical thinking socio-cognitive assessment for disinformation are listed.

3.3.1 Objective CT assessment

The objective Critical Thinking Assessment against Misinformation instrument was specifically designed, drawing inspiration from the scoring methodology of the most suitable commercial tools for measuring and evaluating critical thinking. However, in this instance, the focus of the statements is twofold: they address both the spread of misinformation as a distinct phenomenon, and the content of some of the most prevalent online hoaxes. The propositions that follow the statement for each of the five dimensions are distinctly binary; they can only be right or wrong. The scoring is thus binary as well (i.e., a correct response, indicated in brackets after each item, is scored as 1, while an incorrect response is scored as 0). This approach allows for calculating a cumulative score (sum score) for each dimension independently, and an overall critical thinking skill score (general sum score) by summing each response's values. The reliability of each dimension and of the entire scale can, therefore, be calculated using the Kuder-Richardson 20 (KR-20) coefficient (Kuder & Richardson, 1937), which is specifically designed for similar variables. Presented here are eight exemplary different proposed equivalent tests that assess the five dimensions, each tailored to accurately evaluate the respective aspects of critical thinking as it pertains to misinformation and its dissemination.

1. INTERPRETATION ASSESSMENT

Instructions: Below is a statement taken from an online article. Several conclusions are drawn based on this statement. Your task is to determine whether each conclusion logically follows the statement provided. Choose 'Conclusion follows' if you believe the conclusion is a logical outcome based on the statement, and 'Conclusion does not follow' if you believe it isn't.

Statement: A widely shared online article claims that individuals who spend more than 5 hours a day on social media are 70% more likely to experience symptoms of depression than those who spend less than an hour.

Conclusion: Spending time on social media directly causes depression.

- Conclusion follows
- [✓] Conclusion does not follow (**correct**)

Conclusion: There is a suggested correlation between longer social media usage and a higher likelihood of experiencing depression symptoms.

- [✓] Conclusion follows (**correct**)
- Conclusion does not follow

Conclusion: People who spend less than an hour on social media never experiences symptoms of depression.

- Conclusion follows
- [✓] Conclusion does not follow (**correct**)

Conclusion: Individuals who spend exactly 5 hours on social media might fall between the two groups in terms of depression risk.

- [✓] Conclusion follows (**correct**)
- Conclusion does not follow

Conclusion: If one reduces their social media usage from 6 hours to 2 hours, they will not experience any symptoms of depression.

- Conclusion follows
- [✓] Conclusion does not follow (**correct**)

Conclusion: The article implies that very heavy social media usage is linked with a higher chance of depression symptoms.

- [✓] Conclusion follows (**correct**)
- Conclusion does not follow

Conclusion: People who use social media for 3 hours a day are 70% more likely to experience depression symptoms than those who spend less than an hour.

- Conclusion follows
- [✓] Conclusion does not follow (**correct**)

Conclusion: Individuals who use social media for 4 hours are still less likely to experience depression symptoms than those who use it for over 5 hours, based on the article's claim.

- [✓] Conclusion follows (**correct**)
- Conclusion does not follow

2. ANALYSIS ASSESSMENT:

Instructions: Below is a set of statements taken from an online article. Several conclusions are drawn based on these statements. Your task is to determine whether each conclusion logically follows the provided statements. Choose 'Conclusion follows' if you believe the conclusion is a logical outcome based on the statements, and 'Conclusion does not follow' if you believe it isn't.

Statement: All blue fruits are rich in antioxidants. Some apples are blue. No antioxidant-rich foods are bad for heart health.

Conclusion: All apples are rich in antioxidants.

- Conclusion follows
- [✓] Conclusion does not follow (**correct**)

Conclusion: Some apples are good for heart health.

- [✓] Conclusion follows (**correct**)
- Conclusion does not follow

Conclusion: There exist fruits that are both blue and not rich in antioxidants.

- Conclusion follows
- [✓] Conclusion does not follow (**correct**)

Conclusion: All foods that are bad for heart health lack antioxidants.

- [✓] Conclusion follows (**correct**)
- Conclusion does not follow

Conclusion: There is no blue fruit that is bad for heart health.

- [✓] Conclusion follows (**correct**)
- Conclusion does not follow

Conclusion: If a fruit is not blue, it does not contain antioxidants.

- Conclusion follows
- [✓] Conclusion does not follow (**correct**)

Conclusion: Blue apples are not bad for heart health.

- [✓] Conclusion follows (**correct**)
- Conclusion does not follow

Conclusion: All antioxidant-rich foods are fruits.

- Conclusion follows
- [✓] Conclusion does not follow (**correct**)

3. EVALUATION ASSESSMENT:

Instructions: Below is a statement taken from a digital article. Several arguments related to this statement are presented. Your task is to evaluate the strength of each argument, regardless of how much you personally agree or disagree with each topic. Choose 'Strong argument' if you believe the argument strongly supports or refutes the statement, and 'Weak argument' if you believe the argument doesn't effectively support or refute the statement.

Statement: Online platforms should implement stricter regulations on the content they host to prevent the spread of misinformation.

Argument: Yes, because unregulated content can lead to the spread of false information, which can harm society.

- [✓] Strong argument (**correct**)
- Weak argument

Argument: No, because many popular figures believe that online content should not be regulated.

- Strong argument
- [✓] Weak argument (**correct**)

Argument: Yes, because platforms that have implemented stricter regulations have seen a decrease in misinformation-related incidents.

- [✓] Strong argument (**correct**)
- Weak argument

Argument: No, because implementing regulations might require hiring more staff, which could increase operational costs for the platform.

- Strong argument
- [✓] Weak argument (**correct**)

Argument: Yes, because with regulations, users can trust the content they come across, fostering a safer digital environment.

- [✓] Strong argument (**correct**)
- Weak argument

Argument: No, because online platforms are just tools, and it's up to individuals to discern between truth and misinformation.

- Strong argument
- [✓] Weak argument (**correct**)

Argument: Yes, because without regulations, platforms might be held liable for any damages caused by misinformation on their sites.

- [✓] Strong argument (**correct**)
- Weak argument

Argument: No, because content regulations could lead to a reduction in the amount of content and fewer active users.

- Strong argument
- [✓] Weak argument (**correct**)

4. INFERENCE ASSESSMENT:

Instructions: Below is a statement taken from an online article. Several inferences drawn from this statement are presented. Your task is to evaluate whether each inference logically follows from the

statement or not. Based on the statement, decide whether each of the following inferences is 'True or Probably True' or 'False or Probably False'.

Statement: Recent research has shown that individuals who spend more than 5 hours a day on social media platforms are more likely to feel isolated and report higher levels of anxiety than those who spend less than an hour.

Inference: People who avoid social media altogether will never feel isolated or anxious.

- ☒ False or Probably False (**correct**)
- True or Probably True

Inference: Spending more than 5 hours a day on social media is correlated with feelings of isolation.

- ☒ True or Probably True (**correct**)
- False/Probably False

Inference: All individuals who spend less than an hour on social media is free from anxiety.

- ☒ False or Probably False (**correct**)
- True or Probably True

Inference: It's possible that excessive usage of social media may contribute to heightened levels of anxiety.

- ☒ True or Probably True (**correct**)
- False/Probably False

Inference: Individuals who spend 4 hours on social media won't feel any anxiety or isolation.

- ☒ False or Probably False (**correct**)
- True or Probably True

Inference: The research suggests there's a relationship between the amount of time spent on social media and feelings of isolation and anxiety.

- ☒ True or Probably True (**correct**)
- False or Probably False

Inference: Spending time on social media always leads to negative feelings and outcomes.

- ☒ False or Probably False (**correct**)
- True/Probably True

Inference: Limiting social media usage may be beneficial for mental well-being.

- ☒ True or Probably True (**correct**)
- False/Probably False

5. EXPLANATION ASSESSMENT:

Instructions: Below is a statement taken from a blog post. Several underlying assumptions are presented based on the statement. Your task is to evaluate whether each proposed assumption is actually made in the statement or not. Choose 'Assumption made' if you believe the assumption is

implicitly suggested by the statement, and 'Assumption not made' if you believe the statement doesn't necessarily imply the assumption.

Statement: Due to the increasing number of wildfires every summer, families should invest in better home insurance that covers natural disasters.

Assumption: Wildfires are a type of natural disaster.

- [✓] Assumption made (**correct**)
- Assumption not made

Assumption: There will be a wildfire near every family's home every summer.

- Assumption made
- [✓] Assumption not made (**correct**)

Assumption: All home insurance plans exclude coverage for natural disasters.

- Assumption made
- [✓] Assumption not made (**correct**)

Assumption: Better home insurance plans provide coverage for natural disasters.

- [✓] Assumption made (**correct**)
- Assumption not made

Assumption: Families currently do not have any home insurance.

- Assumption made
- [✓] Assumption not made (**correct**)

Assumption: The frequency of wildfires is a reason for families to consider better home insurance.

- [✓] Assumption made (**correct**)
- Assumption not made

Assumption: Wildfires are the only natural disaster families should be concerned about.

- Assumption made
- [✓] Assumption not made (**correct**)

Assumption: Families can afford better home insurance.

- Assumption made
- [✓] Assumption not made (**correct**)

In the initial stage of the study, a threshold was established to distinguish between high and low scores in the five essential dimensions of critical thinking. This threshold is set at half of the maximum score possible S [cutoff = floor($S/2$)]. For instance, if three items are selected for each dimension of critical thinking, the cutoff is determined to be 1. Conversely, with four items per dimension, the cutoff is set at 2. This approach enables a detailed assessment of critical thinking levels across each of the five dimensions. It identifies areas where individuals exhibit weaker skills, indicating a need for targeted enhancement and strengthening of critical capacities to effectively confront misinformation.

3.3.2 *Dispositional self-assessment of Critical Thinking*

In choosing a dispositional assessment framework for critical thinking, we adopted a self-assessment test structured once again around Facione's model, through the "Critical Thinking Assessment Scale Short Form" devised by Payan-Carreira (2022). This instrument, already validated through empirical research, dissects critical thinking into six distinct subscales, each corresponding to one of the five cognitive dimensions, plus self-regulation. For each subscale, we identified the three items with the highest factor loadings (λ s), signifying their statistical significance as the most representative of each dimension. These items were subsequently simplified for greater clarity, reducing the original scale to 18 items that were slightly rephrased to enhance comprehension.

Simplified Critical thinking Self-assessment Scale Short form (CTSAS)

1. To what extent do you agree or disagree with the following statements? (1=strongly disagree; 5= strongly agree):
2. I try to figure out the content of the problem.
3. I classify information in a systematic way.
4. I examine the values that underlie information.
5. I examine how opinions relate to each other.
6. I figure out the reasoning process behind what others say or write.
7. I figure out whether the reasoning of others builds on claims for which they give no proof.
8. I seek the truthfulness of the evidence that someone provides for their claim.
9. I search for additional information that might support or weaken an argument
10. I examine the reasoning of an objection to a claim
11. I collect abundant evidence to back up opinions.
12. I figure out the merits and demerits of a solution while comparing them to other alternatives in order to make decisions.
13. I arrive at conclusions that are supported with strong evidence.
14. I can describe the consequences of a problem by thinking logically.
15. I can logically present results to tackle a certain problem.
16. I can explain a key concept to clarify my thinking.
17. I review sources of information to ensure important information is not overlooked.
18. I continually revise and rethink strategies to improve my thinking.
19. I reflect on my thinking to improve the quality of my judgment

Concentrating on individual self-assessments, the objective is here to approach critical thinking evaluation as encompassing a range of variable dimensions that should not be viewed as fixed traits inherent within individuals. Therefore, if critical thinking against misinformation is a set of skills that can be developed rather than immutable characteristics, this assessment strategy is designed to accommodate such peculiar differences. This perspective not only aligns with contemporary views on cognitive development, but also offers a practical framework for measuring critical thinking against misinformation in a way that is both malleable and respectful of individual differences.

3.3.3 *Individual variables related to self-regulation*

Parallel to this, the study also focuses on collecting data related to the three aforementioned psychosocial variables. This is a key aspect of the research, as it aids in categorizing different clusters of citizens. This categorization's purpose is to later customize interactions, ensuring that engagement methods are tailored to each group's unique characteristics and needs. In order to mitigate the

potential risks of profiling that can arise from relying on fixed demographic characteristics such as gender identity, religion, or socioeconomic status, the approach adopted in this study concentrates solely on psychological and behavioral variables at the individual level. This focus ensures that the assessment is based on dynamic aspects of a person's psychology and behavior, which are more reflective of their current state and potential for change.

1. Propensity to engage in Analytical Reasoning - Cognitive Reflection Test-II (PAR)

The instrument in question is composed of four open-ended questions (Thomson & Oppenheimer, 2016). As a result, the CRT-II (Cognitive Reflection Test-II) generates four raw textual responses. To analyze these responses effectively, a response coding schema is essential. This schema is applied to convert the textual answers into numerical values. These numerical values are indicative of an individual's tendency to engage in analytical thinking as opposed to opting for impulsive responses that are more readily cognitively available. Furthermore, the instrument delineates three distinct response options. Each of these options is explicitly defined, ensuring clarity in understanding and interpretation of the responses.

Please answer these questions:

1. *If you're running a race and you pass the person in second place, what place are you in?* (intuitive-incorrect answer: first (value = 0); correct answer: second (value = 1); other-incorrect answer (value = 0)).
2. *A farmer had 15 sheep and all but 8 died. How many are left?* (intuitive-incorrect answer: 7 (value = 0); correct answer: 8 (value = 1); other-incorrect answer (value = 0)).
3. *Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?* (intuitive-incorrect answer: June (value = 0); correct answer: Emily (value = 1); other-incorrect answer (value = 0)).
4. *How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long?* (intuitive-incorrect answer: 27 (value = 0); correct answer: none (value = 1); other-incorrect answer (value = 0)).

Upon completion of the CRT-II, the results are processed by summing the numerical values assigned to each response. This sum score falls within a range: at the lower end, a score of 0 indicates a difficulty in giving reflective responses, while at the upper end a score of 4 signifies exclusively correct responses. This scoring spectrum effectively measures an individual's inclination towards either impulsive or reflective thinking. The scoring scheme can be implemented automatically through a rule-based scoring algorithm utilizing regular expressions for pattern matching and categorization, or by using functions coded within specific libraries through R software for statistical computation. At the conclusion of this scoring process, the test yields a singular value; this value represents a quantifiable measure of an individual's propensity for reflective thinking as opposed to impulsive decision-making.

2. Value placed on Truthfulness (VT)

The assessment is designed with four straightforward items that probe into the individual's valuation of knowledge, learning, and truthfulness, even when faced with personal costs. These costs might include exerting extra effort, overcoming situational challenges, or the potential for a radical shift in their own perspective. The significance of this assessment lies in its ability to shed light on an individual's commitment to attaining a certain degree of truthfulness. It reveals insights about their motivation to pursue accurate and critical thinking, particularly in scenarios that present various challenges. This evaluation is crucial as it not only assesses the individual's current stance on these values but also their willingness to uphold them in less-than-ideal circumstances.

Please think about a topic that is particularly crucial to you. Rate the importance of the following statements on a scale from 1 (Not at all important) to 10 (Extremely important). If you can, avoid giving exactly the same rating for all four items.

1. How important is it for you to learn new things about that topic, even when it requires a very big effort?
2. To you, how important is truthfulness concerning that topic, even when it might lead to uncomfortable situations?
3. How important is it for you to question your knowledge concerning that topic, even when it may involve a major change in your current perspective?
4. To you, how important is acquiring new knowledge concerning that topic, even when it may involve a major commitment?

The four values obtained in this way (one for each item) are averaged to return a single value that assesses the value placed on truthfulness by the person.

3. Regulatory Self-Efficacy in Sharing Misinformation (RSSM)

Self-efficacy is defined by Bandura as “a person’s particular set of beliefs that determine how well one can execute a plan of action in prospective situations. To put it in more simple terms, self-efficacy is a person’s belief in their ability to succeed in a particular situation”. Self-efficacy is a central mechanism in exercising human agency (Bandura, 1982), reflecting the individual's perceived self-regulatory capability to control thoughts, motivations, and actions toward a specific goal.

It's crucial in various behavior contexts as it influences how people think, motivate themselves, and act; particularly in the context of misinformation, regulatory self-efficacy involves the perceived capability to critically analyze and validate online information sources and to refrain from sharing unverifiable news, especially under social pressures. This self-regulatory capacity is important in preventing the spread of misinformation, as it enables individuals to resist impulsively sharing unreliable content and to critically evaluate information before sharing.

The assessment of the perceived ability of citizens to cope with online disinformation considers an inhibitory and an active factor, in line with social cognitive literature (Bandura, 1991). The inhibitory factor (4 items) assesses the perceived ability of subjects to refrain from sharing a piece of news even when it would benefit them or they feel the urge to do so, while the active factor (4 items) is aimed at measuring the perceived ability to take action themselves by inquiring and investigating the veracity of a piece of news.

When facing a piece of news that seems dubious or ambiguous to you, how well do you think you can perform the following activities? (1: Not at all - 10: Completely)

1. Avoid spreading it even if not doing so might lead to disagreements within your relevant group. **(inhibitory factor)**
2. Choose not to share it even if it could enhance your popularity or influence. **(inhibitory factor)**
3. Refrain from sharing it even when it seems to be a prevailing trend among your important peers. **(inhibitory factor)**
4. Express your skepticism about the information even if your important peers seem to accept it. **(active factor)**
5. Find and share credible and verified sources even if they may challenge the prevailing narrative in your group. **(active factor)**

6. Refrain from sharing it even if it aligns with your pre-existing beliefs. **(inhibitory factor)**
7. Search for different perspectives and sources on the same news story, even if doing so could question your beliefs. **(active factor)**
8. Take the time to verify the information's sources even when it feels boring or time-consuming. **(active factor)**

The four values obtained regarding the inhibitory factor are averaged, and the same is done for the active factor. This gives two indicators of the two dimensions by which regulatory self-efficacy toward disinformation (inhibitory and active) is assessed.

3.3.4 Data elaboration and recognition of users' characteristics

The initial phase of the assessment involves collecting data on the three socio-cognitive dimensions of interest, utilizing the measurement tools previously outlined. The primary goal of this data collection is to identify different groups of citizens by taking into account assumed nonlinear interactions among these socio-cognitive variables, which is a distinctive trait of the person-centered approach (Magnusson & Stattin, 2007; Howard & Hoffman, 2018). This approach recognizes the complex, interdependent nature of these dimensions and how they uniquely combine in different individuals. To illustrate the potential outcomes of this clustering process, an example using simulated data is provided in Figure 7.

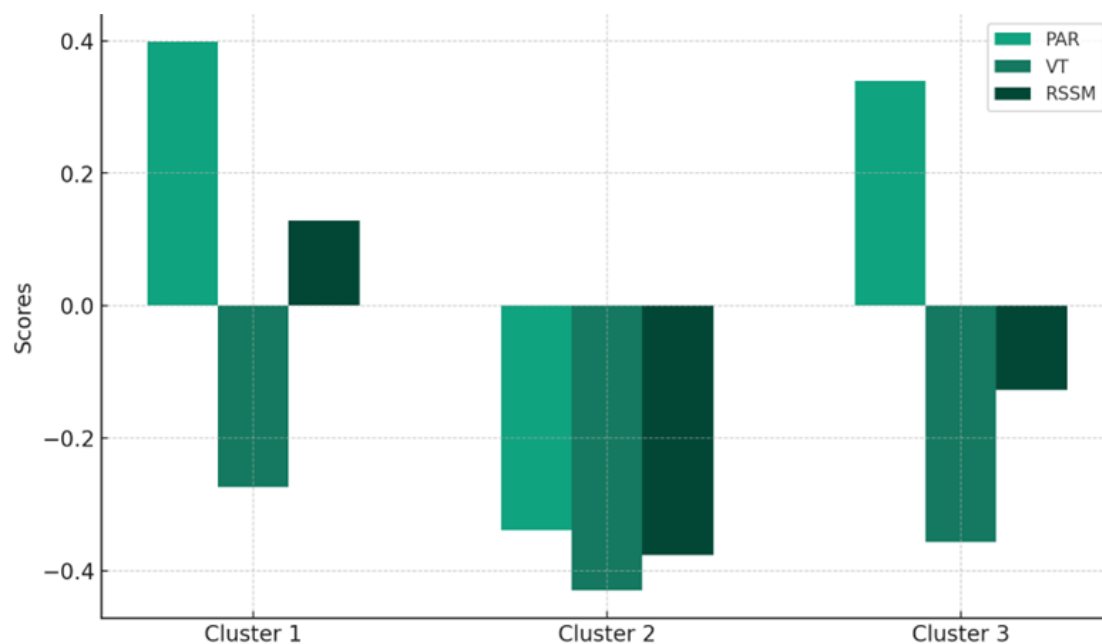


Figure 7. Example from simulated data, assuming three clusters are identified based on the proposed dimensions.

This example serves as a visual representation of how individuals might be categorized into distinct groups, offering insights into the varied socio-cognitive profiles that may emerge from the analysis.

This methodology enables the creation of diverse user profiles, each reflecting the unique characteristics of individual participants. By identifying these distinct profiles, it becomes feasible to tailor the interaction strategies of the TITAN system specifically to each individual. This personalized approach is key to enhancing the system's effectiveness. By adapting both conversation style and interaction techniques to align with the specific traits and needs of each person, the TITAN system can offer a more targeted and relevant experience. This customization is not only conducive to a more engaging user interaction but also ensures that the content and manner of communication are optimally suited to each user's unique socio-cognitive profile.

3.3.5 Matching users' characteristics and TITAN system's recommendation

Initially, the assessment will measure an individual's critical thinking regarding misinformation using the tool outlined in section **Error! Reference source not found.**. This process determines users' proficiency in the five key dimensions of critical thinking. Consequently, the technical implementation's concluding step involves matching user profiles with personalized recommendations from the TITAN system. Depending on each individual's specific characteristics and needs, the system will propose appropriate modules, exercises, and resources. An initial procedural hypothesis involved direct matching between skills related to critical thinking dimensions and the stages of Socratic dialogue (Figure 8), later expanded and extended to incorporate the complexity of critical thinking assessment and the non-linearity of the Socratic method.

user_id	male	age	int1_T1	int2_T1	int3_T1	an1_T1	an2_T1	an_T13	eval1_T1	eval2_T1	eval3_T1	inf1_T1	inf2_T1
1	1	23	0	0	1	0	0	1	0	1	0	1	0
2	0	32	1	1	0	0	0	0	0	1	1	1	1
3	1	78	0	1	1	0	0	0	1	1	0	0	1
4	1	22	1	1	0	1	0	0	1	1	0	0	0
5	0	18	0	1	0	1	0	1	0	1	1	1	1
6	1	47	1	1	0	0	1	0	0	0	1	1	1

Table 1. Values for the first assessment.					Table 2. Values for the second assessment.				
user_id=1	T1				user_id=1	T2			
interpret:	1	Dialogue on Clarification	Low		interpret:	1	Dialogue on Clarification	Low	
analysis:	1	Dialogue on Probing Assumpt.	Low		analysis:	2	NONE	High	
evaluat:	1	Dialogue on Probing Ration.	Low		evaluat:	2	NONE	High	
infer:	2	NONE	High		infer:	2	NONE	High	
explan:	2	NONE	High		explan:	3	NONE	High	
TOTAL CT	7		↑ associated Socratic module	LOW CT	TOTAL CT	10		↑ associated Socratic module	HIGH CT

Figure 8. A snapshot of a simulated dataset depicting the initial data elaboration.

3.4 RESEARCH DIMENSIONS

3.4.1 TITAN CTA in Living Labs

The application of the TITAN CTA in Living Labs involves real-world testing and refinement of the assessment model. In these controlled environments, participants will engage with the TITAN tools, allowing researchers to observe and measure the effectiveness of the CTA in diverse settings. Feedback from these sessions will be integral to fine-tuning the assessment methodology.

The living labs will be utilized to verify the effort required by a user to complete the entire questionnaire. This primarily addresses hypotheses related to user acceptance and user requirements rather than research-focused inquiries. The aim is to assess the practicality and feasibility of the questionnaire from the user's perspective, ensuring that it aligns with their capabilities and expectations. This evaluation is crucial for determining the suitability and user-friendliness of the tool in real-world settings, thus ensuring its effectiveness and applicability in diverse user scenarios.

3.4.2 Dimensions for TITAN Pilots

For the TITAN pilot programs, the research will focus on several key dimensions. These include the effectiveness of the CTA in different demographic groups, the impact of various teaching and learning methods on critical thinking skills, and the scalability of the assessment model across different regions and cultures. The pilots aim to validate the CTA methodology in varied contexts and gather comprehensive data to inform future iterations of the TITAN project.

Utilization of the pilot studies is intended for hypothesis validation. The methodology involves testing the proposed mapping against a control group, with critical thinking skills assessed both pre- and post-pilot. Enhanced critical thinking in dimensions initially identified as weak, if significantly more pronounced in the experimental group compared to the control group, would affirm the efficacy of the mapping approach. In instances of non-conformance, the mapping will undergo recalibration and be re-applied in a subsequent pilot study. Concurrently, the pilot studies will facilitate the optimization of item number and typology. Data collection is critical in determining the most predictive items for each dimension, leading to the refinement of the questionnaire for the ensuing pilot. The research design encompasses the hypothesis (mapping) being subjected to empirical examination, involving both an experimental and a control group. This involves phases of pre-pilot administration, profiling, dialogue engagement, system utilization, pilot conclusion, and final assessment. Subsequent data analysis aims to ascertain the mapping's efficacy, followed by refinement processes, a second iteration of the pilot study, and subsequent re-validation. In preparation for the pilot studies and to achieve preliminary validation, the questionnaire is being disseminated to a student cohort, providing early insights into data trends and enabling adjustments ahead of the main study.

4 THE DIALOGUE IMPLEMENTATION METHODOLOGY

4.1 THE FOUR BLOCK APPROACH

4.1.1 Rationale of the Four-Block Approach

In order to implement a proper dialogue system a methodology to create a dialogue system based on the Socratic Method that would encourage critical thinking and engagement has been developed. The dialogue system was structured into four blocks: (1) psychological inoculation through prebunking, (2) critical evaluation of media items, (3) reflective dialogues, and (4) self-regulation strategies. Inside other WPs these blocks has been integrated into the TITAN platform, allowing for interactive engagement. The dialogue system was successfully implemented and tested. It has been refined based on pilot user feedback, ensuring that it stimulates critical inquiry and media literacy in a structured and user-friendly manner.

4.1.1.1 Requirements

The Four-Block Structure of the chatbot/user interaction was designed to meet the updated set of 'practical' requirements / recommendations. First, **(a)** the length/level of engagement and **(b)** coherence/clarity of the dialogues had to be taken into account, **(c)** the structure to become more transparent and, finally, a few technical requirements aiming to facilitate the modeling and incorporation of the dialogues into the TITAN system.

(a) Length/level of engagement: Striking a balance between length and user engagement level appears to be a complex issue that every chatbot has to address. According to recent findings, while LLM-powered chatbots like ChatGPT-4 might have the ability to alter conversation lengths, *'this does not necessarily lead to improved conversation quality'* (Shih-Hong Huang et al., 2024). Therefore, our initial effort was based on the premise that our dialogues (at least at the beginning) had to be short and coherent in order to keep the user focused, engaged, as well as facilitate their incorporation into the TITAN system. Assuming that *'conversation quality does not vary drastically across different conditions'* (Shih-Hong Huang et al., 2024), we introduced **closed questions** (limiting our original *open-ended* questions) in order to facilitate the answering process for the user, keep our dialogues as short as possible, and encourage a more strictly-guided procedure of conversing with the user always in line with the 'Socratic' method of questioning. Closed-questions could therefore (also) be answered in a Yes or No/I am not sure manner making sure to directly address the potentially disinformative content of the article.

Example:

Examining Evidence and Rationale (*1/6 Types of Socratic Questions*)

- What evidence is there to support the main claim of the author? (*Open Question*)
- Does the evidence offered by the author appear to support the (i.e. conspiracy) theory and further feed your suspicion? (*Closed Question*)

Finally, by *accompanying our questions with additional information* (in the form of explanations, tips or warnings), we aimed to facilitate the answering process for the users and offer them a clearly defined context for a discussion revolving around disinformation. *'To effectively achieve a conversational goal, chatbots should be proactive in providing users with appropriate responses and guide them on the information they need by prompting specific questions on the request criteria'* (Alexandra Rese & Pauline Tränkner, 2024). This way, in the event of a long discussion we aspire to keep user satisfaction levels high by providing the necessary help in order to facilitate answering to more complex, high-conversability questions (Shih-Hong Huang et al., 2024).

(b) Coherence/Clarity: Our concerns regarding the coherence and clarity of these dialogues also played a key role in their design. Our questions had to be conceptually clear and more specific in order to elicit a relevant response from users. Moreover, the language and expressions used should create a sense of familiarity for users thus further encouraging engagement. Complicated terminology had to be simplified or clearly defined/explained, and seemingly repetitive questions had to be avoided since extended conversation length could potentially result in *'repetitive information within messages and it may not enhance message clarity'* (Shih-Hong Huang et al., 2024). Another concern referred to the user's ability to answer broad, generic questions without a certain form of guidance/necessary feedback (i.e. what you think the author is trying to achieve. In your opinion which would be the best way to address [this]?, etc.). As we already mentioned, for this purpose, we made sure to introduce tips, examples and/or warnings to help users better contextualize their answers. Accordingly, Morrissey and Kirakowski (2013) *'identified four attributes for a more natural conversation by a chatbot: conscientiousness (appropriate response behavior), manners (welcoming and closing texts, apologies), thoroughness (accurate grammar and spelling), and originality (specific information regarding the topic, initiative)'* (Alexandra Rese & Pauline Tränkner, 2024). Since a conversation's structure can include different types of *adjacency pairs* (i.e. 'QUESTIONS set up an expectation for an ANSWER, PROPOSALS are followed by ACCEPTANCE (or REJECTION)', etc.) (Daniel Jurafsky & James H. Martin, 2023) it is only logical that information relevant to disinformation and its individual tactics should be taken under consideration by the chatbot in an effort to provide users with material relevant to the discussion and *'help the system decide what actions to take'* (Daniel Jurafsky & James H. Martin, 2023).

(c) Structure: From the very beginning of designing dialogues, the importance of a clear structure became apparent. Introducing a concrete structure in the dialogues would result in facilitating their incorporation into the system while, at the same time, result in avoiding completely random dialogues. Additional research helped determine which types of questions could be asked at the beginning of a dialogue, which were essential to be considered by all users regardless of their particular interest in an article or post, which questions had the potential to be common for all disinformation tactics/dialogue templates, and, finally, which were essential for the proper closure a dialogue (i.e. questions asking the users take, questions helping the user sum up, etc.). Additionally, clarification questions (i.e. questions able to accommodate guided/specific requests and subsequent responses) were taken under consideration in the event users needed additional information and/or potentially help the chatbot initiate any ensuing sub dialogues.

4.1.1.2 Goals

The four-block structure was designed to serve different goals in order to meet the challenges of information disorders, the logic and benefits of prebunking, and the principles of media critical thinking. **Block 1** is the prebunking and inoculation interaction with the user through an infographic.

The purpose of Block 1 is to improve the effectiveness of TITAN's dialogues in Blocks 2 and 3, ensuring the user's understanding of dis/misinformation and the motivation to do analytical thinking. In short, by informing and exposing the user to small 'dosages' of disinformation before the dialogues, the users acquire the domain-specific knowledge/concepts and terms of mis/disinformation tactics, thus preparing them to discuss with the chatbot the domain-specific challenges. Also, the prebunking interaction is a media literacy material which guarantees that those unwilling to engage in lengthy dialogues with the chatbot will be able to acquire valuable information regarding the most common disinformation tactics/techniques. **Block 2** is a new form of dialogue called the "Item-Focused Dialogues". The purpose of these dialogues is to guide the users to evaluate the credibility of media items, i.e. the source and informational content of the item. This is an in-text, domain-specific approach aiming to discuss with the user a particular media item and its, potentially, misinformative or mis/disinformative content. **Block 3** is the user-focused (or person-focused) dialogues. These dialogues focus on the user's dispositions, beliefs, personal bias and media use habits responsible for shaping their perception and interpretation of media sources and informational content. This is an out-of-text approach, i.e. a discussion with the user independently of the content of the media item. The dialogues of **Block 4** are based on the original approach of TITAN. They guide the users to critically evaluate media items using *generic* critical thinking skills instead of the *domain-specific* critical thinking used in Block 2.

4.1.2 Block 1: TITAN's Psychological Inoculation prebunking method

The purpose of BLOCK 1 is to offer a toolbox with the main disinformation tactics and to inoculate the users. "If you can recognize it, you can resist it." The Inoculation Theory guided the creation of Item-Focused Dialogues by determining their goal (rationale), logic and structure. Based on the state-of-the-art scientific logic/basis, we blended inoculation with media critical thinking. In particular, Block 1 incorporates psychological inoculation in order to guide the dialogue design of Block 2 which does media-specific critical thinking against disinformation. In other words, the Item-Focused Dialogues are structured with the logic of inoculation in order to be tactic-specific: if the user learns to recognize the tactic through the dialogues, they can resist it! In effect, Block 1 informs users of the most common disinformation tactics, their patterns and their underlying logic and mechanisms. This pre-exposure equips users with a toolbox that contains the main signs of disinformation with examples of the most common disinformation tactics, fostering familiarity and ease of detection thus enhancing the effectiveness of the Item-Focused Dialogues (Block 2).

The inoculation concept aligns with our concept of 'media critical thinking' (MCT), which aims to identify disinformation tactics, influencing our approach - and aligns with the Socratic approach and Facione's critical thinking dimensions. First, exposure to examples prompts individuals to analyze and critically evaluate these tactics and arguments. Second, it enhances their media critical thinking abilities because individuals learn to interpret the information presented, analyze the arguments and evidence provided, evaluate the sources, make logical inferences, explain their reasoning and regulate their thinking processes—core components of critical thinking. Lastly, it encourages reflective thinking by prompting individuals to reflect on their own susceptibility to misinformation and biases. This reflective practice supports ongoing improvement in critical thinking skills.

At the same time, following the logic of inoculation, Block 1 fortifies user's immunity and cognitive resistance against disinformation: if the user learns to recognize the tactic through the dialogues, they can resist it! The upshot of the inoculation logic is that the dialogues themselves enhance the users' cognitive/mental resistance to disinformation.

4.1.2.1 Methodology of the dialogue's design of Block 1

TITAN'S Prebunking form

The purpose of Block 1 is to offer a toolbox with the main disinformation tactics and to inoculate the users. Since active prebunking requires investing time and cognitive effort, we selected the passive

approach of prebunking (which is informative) in order to make the prebunking process simple and fast. Passive inoculation formats (e.g. infographics, videos) require less engagement, hence the users will not feel burdened with/pressed to perform a task that requires cognitive effort, time and concentration. They mainly have to read 'sticky' information bullets. In addition, we selected the technique-based prebunking approach instead of the narrative-based prebunking because the goal of our Item-Focused Dialogues is to use critical thinking in order to identify techniques/tactics of disinformation (Roozenbeek et al., 2022; Harjani et al., 2022; see Block 2). Also, Narratives are numerous, dynamic and ever changing, hence, we "can hardly inoculate people against every single myth that might present itself in the future" (Lewandowsky, Ecker & Cook, 2017). On the other hand, technique-based inoculation has much broader application because it encompasses more instances of disinformation than narratives (Cook et al., 2017).

TITAN'S Infographic: Intended outcomes and Structure

We created a tactic-based prebunking infographic, informing users about the main manipulation tactics that are commonly used in online disinformation (e.g. conspiracy theory, polarization, trolling and discredit). The **intended outcomes** of the tactic-based prebunking infographic are the following: First, the infographic prebunks users by providing key, memorable information about warning signs and exposing them to a real-life example. This offers users a 'toolbox' of key concepts, signs and examples for use in Item-Focused Dialogues, thereby speeding up the identification process. Second, and related to the previous point, the infographic provides the necessary knowledge (i.e. media literacy material) for effective media critical thinking. Critical thinking is most effective when based on domain-specific knowledge. Finally, the infographic contributes to broader media literacy training efforts.

The infographic provides concise descriptions of the unique warning signs of each tactic (guidance on recognition) and a real-life example illustrating each tactic in action.

For example, three key (Figure 9), unique warning signs of Conspiracy Theory are the following:

1. A **secret group** of powerful people manipulate our society behind the scenes
2. Conspirators have **evil motives**
3. They divide the world into **good** ("us") and **bad** ("them")

EXAMPLE: "Vaccines are just a way for billionaires to track us with their microchip vaccines! Who's really in control of our bodies here?"

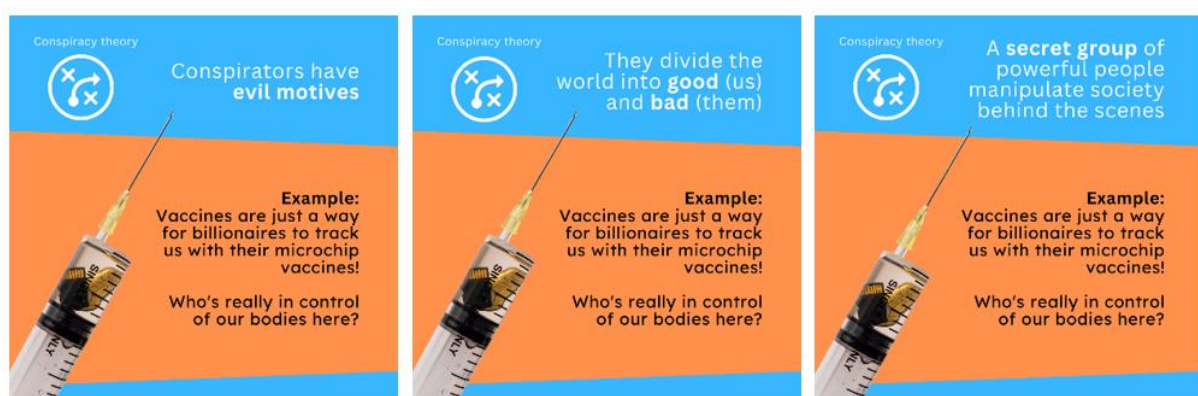


Figure 9. Three keys to interpret the example on vaccines

TITAN'S Infographics about Conspiracy Theory

The infographic does not fully align with the typical structure of Inoculation interventions warning signs plus preemptive refutations. Due to TITAN's requirement for quicker, less intensive engagement, we opted for a simpler structure which presents the key signs of each tactic and provides a single example of each tactic. Thus, the infographic does not include the preemptive refutations of the standard structure. This infographic introduces users to key signs and provides a single example of each tactic without incorporating counterarguments or refutations, which are key components of inoculation. Our aim with Block 1 (Prebunking) was to inform users about the existence of disinformation tactics and teach them how to spot these tactics in the dialogues of Block 2 examining specific media items. However, we have not overlooked the second part of the inoculation method. That is, we placed in Block preemptive refutations for each sign (i.e. fallacy, rhetorical ploy or problematic narrative), in order to against the equip users with the critical thinking tools to counteract manipulative tactics.

FUTURE RECOMMENDATIONS: Need for a. review of the unique signs of each tactic, b. different examples for each sign, illustrating the specific sign.

The suggested format for TITAN's inoculation prebunking consists of the following:

- Greater condensation of the unique characteristics of the tactic. For example, the unique warning signs of conspiracy theory are the following:
 - There is a secret plot by powerful groups with nefarious intent/agenda against the public.
 - Narrative of persecuted victims or heroes.
 - Narrative of "suppression of truth" by the alleged conspirators.
 - "Something must be wrong" with the official accounts.
 - Unfalsifiable/self-sealed claims, i.e. no evidence can prove them wrong because any evidence against them is dismissed as part of the conspiracy or deliberately hidden.
 - The author has inside knowledge of the secret plot or relies on anonymous insiders.
- A distinct example for each unique sign should be provided to make them clearer and easier to identify. The following examples clarify each unique sign:
 - "He claims the injections were part of a plan by the World Economic Forum (WEF), the United Nations and the World Health Organization (WHO), and Big Pharma to massively reduce world population." (SIGN 1, There is a secret plot by powerful groups with nefarious intent/agenda against the public)
 - *"One of the few insiders who blows the whistle on the truth."* (SIGN 6, Author has inside knowledge of the secret plot or relies on anonymous insiders.)
 - "Let us hope that We, the People, will enter a stage of conscious awakening, able to resist and boycott their plan. We are People of the Light." (SIGN 2, Narrative of persecuted victims or heroes.)
 - "The average person has no clue. But their brain capacity and emotion-sensitive pineal gland are being gradually diminished and eventually destroyed." (SIGN3, Narrative of "suppression of truth" by the alleged conspirators.)
 - "However, Dr. Nehls, a leading expert with inside knowledge of the plot, says the plan may not be realized in time, because the "shots were created as a ticking time bomb, that would continue to eliminate the vaccinated for several years after they received the shots"." (SIGN 5, Unfalsifiable/self-sealed claims, i.e. no evidence can prove them wrong because any evidence against them is dismissed as part of the conspiracy or deliberately hidden.)

The examples above are from the article selected by the TITAN project:

4.1.3 Block 2 - Item-Focused Dialogues, Short Description

TITAN's **item-focused** dialogues aim at the critical evaluation of potentially disinformative media items. By abiding to the principles of generic Socratic questioning and media critical thinking, TITAN's critical thinking coach guides users to reflect on the accuracy and trustworthiness of an article/post. In particular, item-focused dialogues prompt users to **(a)** examine the trustworthiness of the author/source (i.e. credibility, reputability, timeliness, etc.), **(b)** identify signs of disinformation tactics/techniques used in an article, as well as **(c)** understand the motives, purpose and implications of an author/item. In contrast to TITAN's person/user-focused dialogues, *item-focused dialogues are specifically created to examine the content and credentials of a particular media item*, an in-text approach prompting users to carefully examine their source, topic and concerns relating to a - potentially- disinformative article/post.

Accordingly, our item-focused dialogues are divided into two parts: **(a)** the 'WHO', and **(b)** 'WHAT' part.

(a) The 'WHO' Part

This part of the dialogue appeals to the interpretation, analysis and evaluation skills (i.e. important Critical Thinking dimensions) of the user by probing questions regarding the author and the source. Mainly tackling problems relating to the trustworthiness of the author/source, this part of the item-focused dialogues aims at the critical evaluation of the credentials and reputability of online sources and individual authors.

By encouraging users to investigate 'WHO' is behind a media item/article we aim guide users to pose fundamental questions relating to the credentials of their source (*critical thinking training*), as well as help them collect important information in order to determine whether an author/source can be considered trustworthy (*media literacy training*). As already mentioned, by questioning the source, assumptions and potential implications of who is behind a particular media item, the user is guided in a 'Socratic' manner to learn to evaluate, analyze and interpret important information that might relate to the potentially disinformative character of a media item¹.

In particular, in this section, users will learn to **(a)** investigate an author's credentials, **(b)** determine whether an expert's opinion is needed in order to discuss a particular topic, **(c)** evaluate a website's appearance (i.e. design, attention to detail, text/content appearance, etc.) and credentials (i.e. policy & purpose, funding, editorial team, etc.), as well as, **(d)** the timeliness of the information they are provided with (i.e. outdated content, functional references & links, etc.) (*see Annexes*).

Questions incorporated in the 'WHO' part of the item-focused dialogues mostly rely on media literacy training protocols (i.e. the *CRAAP Test*, *SIFT Method*, *SMART Check*, *A.S.P.E.C.T.*, etc. please see Annex 3) aiming to train users recognize and resist disinformation by enhancing their critical thinking reflexes and providing them with the necessary information to do so (i.e. how to investigate an author's background, navigate through a journalistic website, etc.). While some of these questions can be traced back to the abovementioned media literacy training protocols, questions are specifically designed to aid users perform these 'tasks' without having to rely exclusively on their pre-existing knowledge/background but rather by arming them with all necessary tools and information to do so (*see Annexes*). As we will see, questions posed by TITAN's chatbot (both in the 'WHO' & 'WHAT' part) are accompanied with **explanations** (*common for all users*), **tips** (*lower levels of media literacy*), **examples** (*lower levels of media literacy*), as well as **warnings** (*users appear to answer questions inconsistently and/or are in a hurry to skip questions and proceed with the dialogue*) and relevant

¹ [1] Problems relating to the authorship/source can be considered an important indicator of potentially disinformative content as well.

micro-lessons (*additional information needed to answers a question, additional interest in a particular aspect of the question, etc.*).

(b) The 'WHAT' Part

Like the first part of the structure of the item-focused dialogues, the second part ('WHAT') adheres to certain 'Socratic' and critical thinking principles (i.e. clarification, interpretation, evaluation, etc.). Unlike the first part of these dialogues, the 'WHAT' part encourages users to delve deeper and engage more critically with the content of the article/post. Once more, by guiding users to investigate whether the content/information they are presented with (in terms of generic and unique signs of disinformation techniques and narratives) can be considered trustworthy/reliable, this part of the item-focused dialogues aims at the critical evaluation of the item's content.

By dividing this part into a series of questions regarding disinformation tactics/techniques, a distinction was made between certain *generic* and *unique* questions on a variety of disinformation signs/signals. Therefore, in the first part of this section, users are called to **(a)** clarify and evaluate arguments, **(b)** identify emotional manipulation/sensationalist tactics, **(c)** attacks on credibility of opponents (efforts to discredit), **(d)** polarizing techniques and binary thinking, as well as, **(e)** matters regarding the manipulation of evidence and relevant fallacies/ploys (*see Annexes*). The second part of the 'WHAT' section refers to questions unique to a series of disinformation signs/signals. Having identified unique signs of disinformation tactics (i.e. characteristics of conspiracy theories, trolling, efforts to discredit or polarize, etc.), users are gradually acquainted with the individual tactics of disinformation, their unique characteristics, usual purpose/agendas, implications, etc.

To answer these questions, users are provided with all necessary tools and information, while, at same time, their critical thinking ability/skills can be further cultivated. In particular, users have the opportunity to enhance their **(a)** ability to recognize, analyze and evaluate arguments, **(b)** to detect common logical and rhetorical fallacies, **(c)** identify biases and problems with evidence offered, **(d)** distinguish elements of the author's reasoning, as well as **(e)** determine the overall quality of a media item.

In conclusion, in this part of the dialogue users are encouraged to determine the trustworthiness (*accuracy*) of the content by learning to discern signs of individual disinformation tactics/techniques and are presented with an opportunity to summarize their conclusions/findings of the overall quality of a media item.

As previously discussed, TITAN's item-focused dialogues were divided into two main parts: **(a)** the 'WHO', and **(b)** the 'WHAT' part, both consisting of questions specifically designed to prompt users to critically evaluate their source ('WHO') and the item's content ('WHAT'). Using a mix of open-ended and closed questions, users are guided to effectively evaluate potential disinformation, as well as the overall quality of the media item. In an attempt to encompass/cover many potential answers, we introduced **three hypothetical branches**, each representing a different type of user, a distinction based on the assumption that a user's ability to answer questions may differ depending on their personal knowledge regarding disinformation. In particular:

User No.1: This branch refers to the average, media literate user who is cooperative and able to effectively answer questions. These users -presumably- have the ability to navigate easier through TITAN's dialogues, therefore concentrating on questions (and consequent sub dialogues) of personal interest revolving around i.e. a particular tactic/technique and/or aspect of disinformation.

User No.2: This branch refers to -presumably- low media literacy users who are having difficulty answering the chatbot's questions and are in need of additional guidance/help (and/or users who are unsure of how to answer a particular question and voluntarily request additional information/guidance). In this particular case, we make sure to provide these users with additional information in the form of *explanations*, *tips*, *definitions* and *examples* in order to facilitate the answering process and arm them with the necessary information to continue the dialogue. Additionally, potential sub-dialogues initiated by these users, were designed to include suggestions for further reading/training in the form of *microlessons*.

User No.3: The final branch refers to users who appear unwilling to cooperate and constantly challenge the chatbot. This branch is addressed to users who might (deliberately or not) dismiss the questions posed to them or even defend certain aspects/forms of disinformation (i.e. ‘*why not troll? It’s fun*’, ‘*I don’t need you in order to determine whether that’s potentially disinformative, I already know*’, etc.). In these cases, we made sure to introduce additional *warnings* for users to contemplate on, as well as certain *refutations* (wherever applicable) addressing specific aspects/characteristics of unique disinformation tactics/techniques as part of a ‘semi-debunking’ procedure.

These additions, along with a series of other amendments, aimed at the alignment of our dialogues with the methodological commitments and critical thinking goals of TITAN.

4.1.3.1 Methodology of the dialogue’s design - Structure of TITAN’s Questions (Block 2)

As already stated, TITAN’s questions were designed to adhere to the principles of **(a)** ‘Socratic’ questioning, **(b)** critical thinking, as well as **(c)** media literacy training against disinformation. Also, we made sure that every question was followed by a short explanation describing its rationale to further encourage critical engagement, facilitate conceptualization and clarify its importance and purpose. Follow-up questions were designed to function as entry-points for *microlessons*, *generic refutations* (i.e. pointing out inconsistencies, addressing potential fallacies and/or rhetorical ploys) and additional *warnings* (some common for all users) in an effort to further help users and therefore facilitate the answering process. Moreover, certain *descriptions/definitions* of individual disinformation signs/signals (whether *unique* or *generic*) were also introduced for similar reasons.

Structure of Item-Focused Dialogues: PART 1

As previously explained, the first part of the item-focused dialogues (‘WHO’) refers to questions investigating the author and the source by tackling problems relating to their trustworthiness and credibility/reputability. It is comprised by **(a)** two indicative questions regarding the author/expertise and **(b)** six questions regarding the source. The number and content of these questions may vary depending on the nature of the media item (i.e. article, post, etc.) but, in every case, it aims to cover a broad range of credibility issues relating to the author/source, one of the most important parts of discerning potentially disinformative content (see Table 4). Depending on their answers, users can either quickly navigate the first part of the dialogues or delve deeper into relevant issues.

Table 4. Indicative questions for Part 1 (‘WHO’)

Indicative Questions (About)	Description	Type
Author's credentials	Evaluating the credibility and qualifications of the author.	Closed
Need for expert opinion	Assessing whether an expert perspective is essential and evaluating the expertise of the author.	Open
Website appearance	Evaluating the professionalism and design of a website.	Closed
URL purpose and bias	Analyzing the intent and potential bias of a source based on its URL.	Open
Funding disclosure	Examining the financial backing of a website and its potential influences.	Closed
Policy statements	Evaluating the alignment of a website's policies with standards.	Closed

Editorial team details	Assessing transparency and editorial standards of a website.	Closed
Timeliness of information	Considering the relevance and currency of information on a website.	Open

Structure of Item-Focused Dialogues: PART 2

The second part of the item-focused dialogues ('WHAT') addresses a number of issues regarding the identification of **(a)** *generic signs* of disinformation tactics/techniques (i.e. emotional manipulation, binary thinking, efforts to discredit, etc.), **(b)** *unique signs* of individual disinformation tactics (i.e. conspiracy theories, trolling, smear campaigns, etc.), as well as, **(c)** the potential *purpose/agenda* of the author/media outlet. After answering **(i)** five questions (*common in all dialogues*) aiming to prompt users employ their critical thinking skills in order to identify *generic signs* of disinformation, users will have the opportunity to **(ii)** discuss the *unique characteristics/signs* of a particular (or more should they choose so) disinformation tactic/technique based on their take of the article/post (number of questions varying depending on the tactic). With the first five generic questions functioning as a bridge to the individual tactics, users are called to critically examine the media item by identifying signs of potential disinformation (Table 5).

Table 5. Indicative Generic Questions for PART2 ('WHAT')

Indicative Generic Questions (About)	Description	Type
Clarification and Argument Evaluation	Does not explicitly address a specific fallacy or rhetorical ploy but focuses on distinguishing the important distinction in critical thinking for media, the distinction between factual claims and personal speculation/opinion.	Open
Emotional Manipulation and Sensationalism / Disruption, Provocation and Manipulation tactics	Addresses common fallacies and rhetorical tactics used by trolls, conspiracists, propagandists, marketeers, scammers, to provoke strong emotional reactions (Rage Farming, Fearmongering, Stirring the Pot, Hyperbole, Hasty Generalizations, Clickbait Headlines, etc).	Closed
Credibility Attacks & Misleading / Diversion Tactics	Explores common fallacies and rhetorical ploys used in Credibility Attacks by trolls, conspiracists, propagandists (e.g. ad hominem attacks, smear), distorting or misrepresenting opponents' positions (straw man arguments), mislead or divert attention (Whataboutism and selective quoting).	Closed

Polarizing language and Binary Thinking / Division Tactics	Explores common fallacies and rhetorical ploys such as False dichotomies, false dilemmas, Polarizing language, Demonization, Oversimplification, etc.	Closed
Evidence Fallacies / Ploys / Misleading Tactics	Explores common fallacies and rhetorical ploys related to the interpretation/ and evaluation of data/evidence, such as cherry-picking, reliance on pseudoscience, anecdotal evidence, and questionable sources.	Closed
Agenda/motives of the user/post	Explores the agenda and motives of the author	Open

4.1.4 Block 3 - person (user)-focused dialogues

In this part of the document, we focus on the person focused dialogue not on the item-focused one, this means we are not focused on the specific news. This section outlines the methodology employed in our European project to enhance critical thinking skills among citizens through personalized dialogues. Our approach is divided into two main blocks: Block 3 and Block 4. Each block targets different dimensions of critical thinking and incorporates validated questions and self-regulation strategies to personalize and refine the dialogues. This narrative aims to provide a comprehensive and detailed explanation of the scientific methods and processes involved in the development and implementation of these personalized dialogues. Blocks 3 and 4 relate to the person and not to the item. For this reason, it is not possible to define simple closed questions, but it is necessary to stimulate the user with open-ended questions, precisely because the system should only serve as a prompt, while it is the user themselves who must find and grow in their own path.

4.1.4.1 Why a Person-Focused Dialogue?

Person-focused dialogues are designed to engage individuals in reflective questions that enhance their self-efficacy in handling misinformation and affirm the importance they place on truthfulness. Unlike item-focused dialogues that target specific types of news articles, person-focused dialogues are tailored to improve self-regulatory aspects of critical thinking. These dialogues are introspective, guiding users to evaluate their beliefs and behaviors and their broader impact on the social environment, rather than merely relying on some external information.

4.1.4.2 Critical Thinking Framework for Block 3 and 4. The Theoretical Approach

After a detailed analysis of the literature, the Facione approach was chosen. Based on this approach, two distinct blocks of dialogue construction were defined: Block 3, which works on the first 5 dimensions, and Block 4, which works on the last (sixth) dimension. Here are the 6 dimensions with the figure Figure 5 (for sake of simplicity, reported also here in order to save the reader from going up and down in the document) that illustrates how they combine into sub-elements.

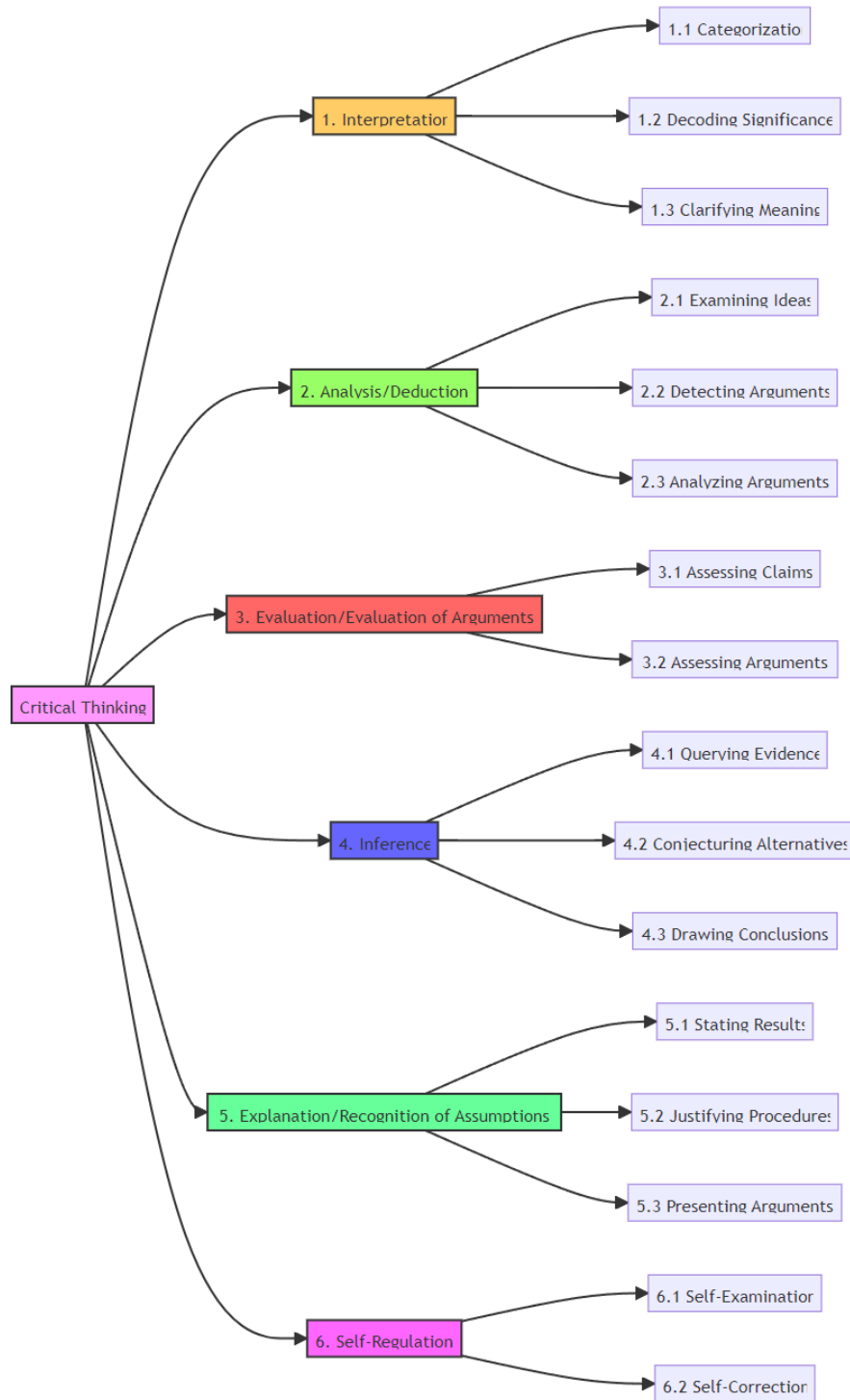


Figure 5. The six dimensions of critical thinking, namely critical thinking cognitive skills and sub-skills.

4.1.4.3 Methodology design of Block 3

The dialogue related to blocks 3 and 4 should therefore not be considered as an alternative or sequential to that of blocks 1 and 2, but rather it should be integrated with them. From a sequential point of view, the first step to take is to provide an assessment of the skills and dimensions related to critical thinking. This assessment is defined in section 2.2, and it is equally indicated in the deliverable. Block 3 forms the foundation of our personalized dialogue approach, focusing on the five critical dimensions of critical thinking: Interpretation, Analysis, Evaluation, Inference, and Explanation. Each

dimension encompasses specific cognitive skills essential for individuals to effectively navigate and interpret complex information landscapes. The methodology begins with a rigorous assessment of these dimensions. Utilizing robust tools such as the Critical Thinking Assessment Measure (CTAM) and the Payan-Carreira Critical Thinking Test, we systematically evaluate participants' cognitive abilities. These instruments are chosen for their proven reliability and validity, presenting participants with objective, right-or-wrong scenarios that accurately gauge their proficiency. The results are meticulously analyzed to identify areas of strength and areas requiring further development. Consider the example of a participant named Alex. Through our assessments, we discover that Alex demonstrates strong abilities in Interpretation and Inference but shows weaknesses in Evaluation and Analysis. This insight becomes the cornerstone for crafting personalized dialogues tailored to Alex's specific needs.

Drawing from an extensive repository of validated questions in the literature, we craft dialogues that address the specific needs of each participant. These questions are not arbitrarily chosen; they are the result of extensive research and are designed to target and enhance each dimension of critical thinking. For Alex, whose assessment indicates weaknesses in Evaluation and Analysis, the personalized dialogue might include questions such as:

- Evaluation: "How credible do you find the sources cited in this article? What factors influence your judgment?"
- Analysis: "Can you identify any assumptions the author is making? How do these assumptions affect the overall argument?"

These questions are seamlessly integrated into the dialogues, creating a dynamic and engaging conversation. The TITAN chatbot acts as a knowledgeable guide, prompting Alex to reflect, question, and analyze the information presented. The dialogues are designed to be adaptive, responding to Alex's inputs and providing tailored feedback that encourages deeper critical thinking.

A key strength of this methodology is its dynamic nature. The dialogues are not static; they evolve based on continuous monitoring and reassessment. Alex's progress is periodically evaluated using follow-up assessments, ensuring that the dialogues remain relevant and effective. This ongoing process of reassessment and adaptation allows us to refine the dialogue content continuously, fostering a cycle of continuous improvement in Alex's critical thinking abilities. This dynamic personalization ensures that participants receive a tailored experience that adapts to their evolving cognitive skills. By maintaining a close feedback loop, we ensure that the interventions remain pertinent and effective, promoting sustained growth in critical thinking proficiency.

The Critical Thinking Assessment Measure (CTAM) and the Payan-Carreira Critical Thinking Test can be employed to assess participants' proficiency levels across the five dimensions. These tools provide objective, right/wrong answers that help identify areas where participants may need further development. Participants should be administered these assessments at the beginning of the study. The results should be analyzed to determine proficiency levels in each of the five dimensions. Scores are categorized into high and low proficiency based on the median split of the dataset.

For examples see Annex 5: Dialogues examples for CT dimensions

4.1.5 Block 4 dialogues

4.1.5.1 Self-regulation dimension

Self-regulation is crucial to functional critical thinking because it involves the ability to control one's cognitive, emotional, and behavioral responses, particularly when dealing with misinformation. Block 4 addresses the sixth dimension of critical thinking: self-regulation. This dimension involves personal attitudes, behaviors, and the ability to monitor and control one's cognitive processes. Self-regulation is crucial for maintaining objectivity and avoiding biases when evaluating information. Effective self-

regulation helps individuals maintain focus on long-term goals and resist the temptation of short-term automatic reactions. In this framework, two specific motivational aspects are emphasized because of their direct link to enhancing self-regulatory capacity in the context of online misinformation.

4.1.5.2 Unpacking the Complexities of Self-Regulation

While cognitive skills are fundamental, the journey towards enhancing critical thinking proficiency would be incomplete without addressing the dimension of self-regulation. Block 4 focuses on this crucial aspect, recognizing that personal attitudes and behaviors play a pivotal role in managing information and resisting misinformation. Self-regulation is assessed through carefully designed self-report questionnaires, which measure two critical aspects: regulatory self-efficacy against misinformation (RSSM) and the value placed on truthfulness (VT). These instruments delve into participants' confidence in handling misinformation and their commitment to truthfulness, providing a nuanced understanding of their self-regulation capabilities. Take, for instance, a participant named Maria. The assessment reveals that Maria has high regulatory self-efficacy but places a low value on truthfulness. This insight sets the stage for personalized dialogue interventions aimed at enhancing her commitment to truthfulness while leveraging her existing confidence.

4.1.5.3 Regulatory Self-Efficacy Against Misinformation

This aspect refers to an individual's perceived ability to effectively manage and resist the spread of misinformation. Self-efficacy, a term coined by psychologist Albert Bandura, describes one's belief in one's ability to succeed in specific situations. In this context, it includes the two factors of confidence to refrain from sharing unverified information, even when doing so might be socially rewarding or easy to do, and the determination to engage in accurate fact-checking, even when it requires significant effort. Dialogues aimed at enhancing this dimension should foster a sense of competence and control, enabling users to critically assess information and respond to it with mastery.

4.1.5.4 Value given to truthfulness

This aspect takes into account the personal value individuals assign to truthfulness in online environments, measuring how much importance a person places on the accuracy and integrity of the information they consume and share. This value is about setting a specific goal, but also reflects a deeper stance towards communication and information exchange. By giving a high importance to truthfulness, individuals inherently support a healthier digital discourse and contribute to a more solid and trusted online community. Dialogue strategies should therefore encourage reflection on the personal and societal impacts of spreading verified versus unverified information, pushing toward a greater appreciation for the role of truth in maintaining meaningful connections and trust within the community.

4.1.5.5 Clustering for Personalized Intervention

Participants like Maria are grouped based on their self-regulation profiles using a median split. This method categorizes participants into distinct clusters, such as low-high RSSM and low-high VT. Each cluster represents a unique combination of self-regulation traits, enabling us to tailor interventions precisely. For Maria, whose profile indicates high confidence but low commitment to truthfulness, the dialogues are crafted to reinforce the importance of truthfulness while leveraging her existing confidence. TITAN might ask questions like:

- "What do you think are the consequences of spreading unverified news? How might it affect individuals or communities?"
- "In what ways do you feel responsible for the accuracy of the information you share online? How does it reflect on you?"

These questions are designed to prompt reflection and emphasize the significance of truthfulness, gradually shifting Maria's attitudes towards a higher commitment to accurate information sharing.

4.1.5.6 Enhancing Regulatory Self-Efficacy in Sharing Misinformation

A pivotal component of our methodology is enhancing regulatory self-efficacy in sharing misinformation. This dimension assesses participants' confidence in their ability to verify and manage the information they encounter, addressing both inhibitory and active factors. Inhibitory factors involve the ability to refrain from sharing dubious information, while active factors involve taking proactive steps to verify the authenticity of information.

Consider a participant named John, who scores low on regulatory self-efficacy. John might struggle with resisting the urge to share sensational news or lack the initiative to verify its credibility. For John, TITAN poses questions aimed at building these skills:

- Inhibitory Factor: "How would it feel to pause and think for a moment more every time you encounter a news story before deciding whether to share it or like it? What might change in your interactions online?"
- Active Factor: "Can you recall a time when you wished you had thought more before sharing something online? What would you do differently now with the right strategies?"

These questions serve as transformative prompts, encouraging John to develop a mindful approach to information sharing. By addressing both inhibitory and active factors, the dialogues empower John to navigate the digital information landscape with greater confidence and discernment.

4.1.5.7 The Block 4: The Self-Regulation Grid

The below image (Figure 10) presents a conceptual framework designed to classify individuals based on two key dimensions of self-regulation in dealing with online misinformation: "Self-confidence in dealing with online misinformation" (rsm) and the "Importance given to truthfulness" (value_truth).

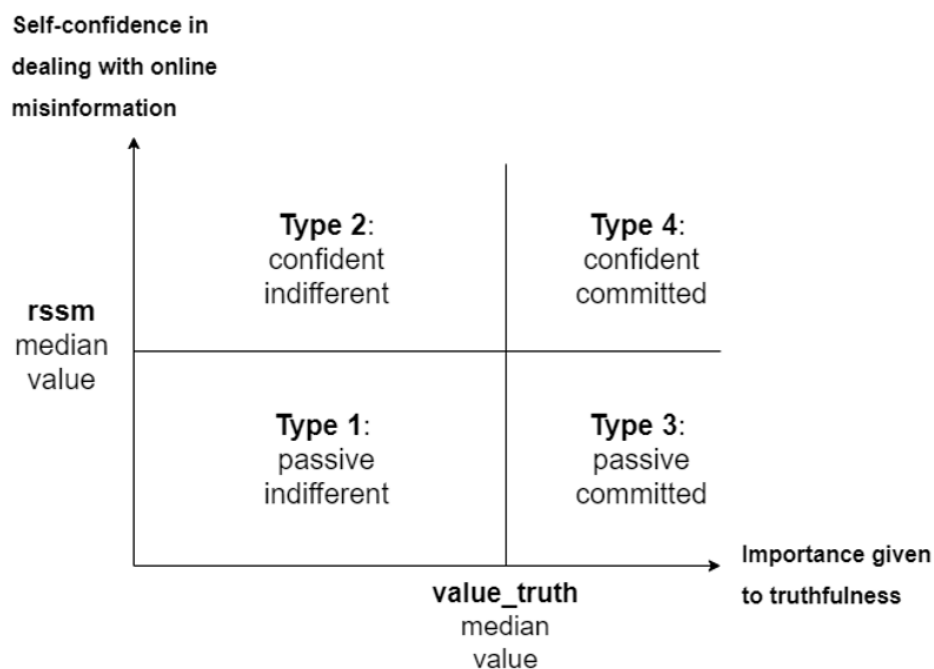


Figure 10. The Self-Regulation Grid

This grid is divided into four quadrants, each representing a different combination of these dimensions, resulting in four distinct types: **Type 1** (passive indifferent), **Type 2** (confident indifferent), **Type 3** (passive committed), and **Type 4** (confident committed).

4.1.5.8 Axes Explained

1. **rssm (Self-confidence in dealing with online misinformation)**

Vertical Axis: This dimension measures an individual's confidence in their ability to identify, verify, and manage misinformation online. The median value on this axis represents the point at which half the population scores above and half scores below, effectively dividing individuals into those with higher self-confidence and those with lower self-confidence.

2. **value_truth (Importance given to truthfulness)**

Horizontal Axis: This dimension assesses the degree to which individuals value truthfulness in the information they encounter and share. It measures how important truth is to an individual, even if it means changing their mind or facing discomfort. The median value on this axis similarly splits the population into those who highly prioritize truthfulness and those who do not.

4.1.5.9 Quadrants Explained

1. **Type 1: Passive Indifferent (Low rssm, Low value_truth)**

Description: Individuals in this quadrant exhibit low self-confidence in managing misinformation and do not place a high importance on truthfulness. They tend to be passive in their approach to information, neither actively verifying nor valuing the accuracy of what they encounter.

Example: Consider John, who frequently scrolls through social media but rarely questions the veracity of the posts he sees. John is neither confident in his ability to detect misinformation nor particularly concerned with the truthfulness of the content he consumes. As a result, he is more likely to passively accept and potentially share misinformation without scrutiny.

2. **Type 2: Confident Indifferent (High rssm, Low value_truth)**

Description: These individuals have high self-confidence in their ability to handle misinformation but do not place much importance on truthfulness. They believe they can detect false information but may not be motivated to prioritize or disseminate accurate information.

Example: Maria is confident in her ability to spot fake news due to her critical thinking skills. However, she is indifferent to the truthfulness of the information as long as it aligns with her interests or biases. This confidence without commitment to truthfulness can lead to selective verification and potential dissemination of misleading information that suits her preferences.

3. **Type 3: Passive Committed (Low rssm, High value_truth)**

Description: Individuals in this quadrant prioritize truthfulness but lack confidence in their ability to manage misinformation. They value accuracy but feel unsure about their skills to identify or counteract false information.

Example: Alex highly values the truth and feels a strong responsibility to share accurate information. However, Alex often feels overwhelmed by the volume of information and doubts his ability to

effectively verify it. This lack of confidence makes Alex hesitant and passive in engaging with online content, leading to potential underutilization of his commitment to truth.

4. Type 4: Confident Committed (High rsm, High value_truth)

Description: This group consists of individuals who are both confident in their ability to manage misinformation and highly value truthfulness. They are proactive in verifying information and are committed to ensuring the accuracy of the content they share.

Example: Emma exemplifies this type; she is both skilled and motivated in dealing with misinformation. Emma regularly fact-checks the information she encounters and feels a strong ethical responsibility to share only verified and accurate content. Her actions contribute to reducing the spread of misinformation within her social circles.

4.1.5.10 Implications of the Framework: understanding the Quadrants

Type 1 (Passive Indifferent): This group is most vulnerable to the spread of misinformation due to their low engagement and lack of commitment to truth. Interventions for this group should focus on increasing both their confidence and their appreciation of the importance of truthfulness. Educational programs and awareness campaigns could be beneficial here.

Type 2 (Confident Indifferent): These individuals could potentially identify misinformation but might not act against it due to their indifference to truth. Strategies for this group should aim to enhance their ethical considerations and the broader implications of spreading false information. Highlighting the social and personal benefits of truthfulness could help shift their attitudes.

Type 3 (Passive Committed): This group is committed to truth but needs support in building their confidence to act effectively against misinformation. Empowering these individuals with tools, strategies, and resources to enhance their verification skills can activate their latent commitment to truth. Training sessions, workshops, and easy-to-use fact-checking tools could be particularly useful.

Type 4 (Confident Committed): These individuals are the most effective in combating misinformation. They not only have the skills but also the motivation to ensure accuracy in their information consumption and sharing. For this group, the focus should be on sustaining and enhancing their capabilities, perhaps by involving them in training others or in developing community standards for information sharing.

The self-regulation grid for managing online misinformation provides a nuanced understanding of how individuals interact with information in the digital age. By categorizing individuals based on their confidence and commitment to truthfulness, we can tailor interventions that address specific needs and behaviors. This approach not only enhances individual critical thinking skills but also promotes a more informed and discerning society. Through targeted educational programs, tools, and resources, we can move individuals towards higher self-regulation, thereby mitigating the spread of misinformation and fostering a culture of truthfulness and integrity in information sharing.

For instructions and examples see Annex 5.

4.2 THE SOCIOLOGICAL DIMENSION

Besides the psychological dimension, the blocks 3 and 4 approach will be complemented also with a sociological dimension. For this dimension, a different methodological approach was implemented. Socratic dialogues should be developed based on an integration of both a literature review, presented in D2.1, and insights from the Co-creation Labs (i.e., Co-creation Phase B). This approach further ensures the integration of relevant works done within the project to obtain, together with the psychological dimension, a holistic approach to stimulating critical thinking through dialogues, fueled by the Socratic method.

The Critical Thinking Dialogue Guidelines have presented above, within the psychological dimension. In this chapter we will elaborate on the Socio-Technical Dialogue Requirements and User Dialogue Requirements.

Socio-Technical Dialogue Requirements

The socio-technical dialogue requirements (Table 6) were constructed based on specific requirements and concepts from Deliverable 2.1. Each requirement is linked to a specific concept and can be found in a specific section of the deliverable. Table 1 shows these different requirements, concepts, and sections, as well as the references, the dialogue requirement they were translated into, and the specific goal of each dialogue requirement.

Table 6. Socio-Technical Dialogue Requirements

Requirement	Concepts	Section	References	Dialogue requirement	Goal
The chatbot should balance critical thinking and trust	Critical Thinking, Trust	To trust or not trust in news	Boyd (2017)	The chatbot should formulate questions that encourage users to critically evaluate information without fostering excessive skepticism	Apply the Socratic method. Express appreciation on critical reflection. Avoid overtly negative framing of sources.
The chatbot should acknowledge the ambiguity of trust in news	Trust, Ambiguity	The ambivalence of trust in news	Strömbäck et al. (2020)	The chatbot should formulate questions that prompt users to reflect on their trust in news sources and how it influences their perception	Inquire about why they find a source trustworthy.
The chatbot should address the role of emotions in news consumption	Emotions, News Consumption	The ambivalent relationship between trust and critical thinking	Beckett and Deuze (2016)	The chatbot should formulate questions that help users recognize the emotional impact of news and how it affects their judgment	Inquire about relationship with news source: levels of attachment to the source, fandom, emotional state post consumption
The chatbot should consider socio-demographic factors	Socio-Demographic Factors	The socio-demographic s of news and information consumption	Edgerly (2015)	The chatbot should formulate questions that explore how users' backgrounds might influence their news consumption and trust levels	Inquire about their history with news use: frequency of use of news (source), start of use of news (source), existing knowledge about

					politics and media
The chatbot should recognize the impact of political orientation	Political Orientation, Media Trust	It's the politics stupid!	Verma et al. (2018)	The chatbot should formulate questions that examine how political beliefs shape trust in media and the acceptance of news	Reflect on how political beliefs influence perception of news.
The chatbot should address news avoidance	News Avoidance, Re-engagement	Zooming in on news avoidance and disengagement	Villi et al. (2021)	The chatbot should formulate questions that investigate reasons for avoiding news and how to re-engage with trustworthy information	Reflect on how news fatigue can affect one's critical stance to news.
The chatbot should help users to differentiate between true and false news	Fake News, Verification Skills	Zooming in on the likeliness to believe fake news and disinformation	Allcott and Gentzkow (2017)	The chatbot should formulate questions that help users develop skills to identify and verify the credibility of news sources	Propose pathways to credible information
The chatbot should help users recognize their biases	Bias Recognition, Critical Thinking	To trust your gut	Boyd (2018)	The chatbot should formulate questions that challenge users to reflect on their biases and how these affect their interpretation of news	Reflect on how one's own bias might affect their interpretation of news.

The chatbot should address alternative ways in which users arrive to critical thinking	Critical Thinking, Epistemological Diversity	The Good, the Bad, and the Trusted	Boyd (2018) , Schwarzenegger (2020)	The chatbot should formulate questions that encourage users to reflect on what underpins their critical thinking, i.e. evidence or rather personal experience, faith or political leaning	Reflect on the diversity of perspectives and epistemological approaches.
The chatbot should address the weaponization of critical thinking	Weaponization of Critical Thinking	The Good, the Bad, and the Trusted	Boyd (2018)	The chatbot should formulate questions that challenge users to reflect on how their critical thinking could be influenced or manipulated	Help the user understand the concept of weaponizing critical thinking and its implications.
The chatbot should highlight the risks of excessive personal responsibility in media literacy	Media Literacy, Personal Responsibility	The Good, the Bad, and the Trusted	Boyd (2018) , Bulger (2018)	The chatbot should formulate questions that prompt users to consider the role of media creators, platforms, and regulators in ensuring trustworthy information	Invite the user to reflect on the role of other societal actors in helping to navigate online information.
The chatbot should be transparent about AI limitations	AI Transparency, Trust	Ceteris paribus in AI solutions or the lack thereof	Følstad et al. (2021)	The chatbot should formulate questions that explain the AI's decision-making process and its limitations to build user trust	Emphasize the limitations of AI in providing news and the importance of human oversight.

Further, to each specific dialogue requirement (DR), some example questions and statements were linked to aid the dialogue development process.

Example 1:

DR: The chatbot should formulate questions that prompt users to reflect on their trust in news sources and how it influences their perception

Example questions:

- *Let's reflect on why you find the source/author trustworthy?*
- *What aspects of this news site make you feel comfortable trusting it?*

Example statements:

- *Many people use sources that they deeply trust. However, that doesn't always mean they are reliable.*

Example 2:

- *DR: The chatbot should formulate questions that examine how political beliefs shape trust in media and the acceptance of news*

Example questions:

- *How do your political beliefs align with this news source?*
- *Do you think your political views impact your perception of this article?*

Example statements:

- *Political beliefs can shape how we interpret information.*
- *Sometimes, our interests and beliefs can shape the way we perceive news.*
- *Being aware of this influence is crucial for critical analysis.*

User Dialogue Requirements

The User Dialogue Requirements (Table 7) were constructed based on specific insights and feedback from the participants that tested the prototype of the TITAN chatbot during the Living Labs in Co-creation Phase B. This feedback specifically relates to the structure, content, and wording of the dialogues. Table 2 presents an overview of the relevant insights that were selected and how they were translated into specific User Dialogue Requirements.

Table 7. User Dialogue Requirements

Insights	Dialogue Requirement
Dialogues need to encourage the user	The chatbot needs to let the user know they are on the right path, e.g. by saying "you are doing great"
The language needs to be accessible	The language the TITAN Coach uses needs to be clear and easy, limiting difficult words and phrasing
Users want to be able to get background information and want to be able to ask for more information	1. The chatbot needs to provide information about a certain concept within the question (e.g. related to a specific news topic, bias, political orientation,...)
	2. The chatbot needs to be able to provide hyperlinks to additional information about article and/or topic
The chatbot needs to take into account the users' answers	The chatbot should repeat a part of the users' answer within its own responses to prove interactivity and engagement
The chatbot needs to feel human	The chatbot needs to act and speak like a human, e.g. using the word 'we' makes users feel like they are doing it together

The dialogue needs to be structured, concise, and to the point	1. The chatbot should have subtitles in the dialogue, to make it more clear what is being evaluated
	2. The chatbot should clearly state and point out what users need to look at and/or think about
	3. The chatbot should highlight the different steps in the evaluation process

SECTION II SUPPORTING METHODOLOGIES AND TOOLS

5 EVALUATION OF DISINFORMATION SIGNALS

5.1 TRAINING METHODOLOGIES

This section presents the training methodologies for developing machine learning models to perform tasks such as disinformation signal detection. The methodologies are divided into traditional learning, which includes sentence-based classification and sequential labeling classification, and few-shot as well as zero-shot learning.

5.1.1 *Traditional Learning, Sentenced-based Classification*

In the realm of machine learning, traditional learning methods, particularly those based on sentence-level analysis, have been widely employed for various classification tasks. These tasks range from binary and multi-class to multi-label classification problems. Numerous traditional learning approaches adopt the fine-tuning of transformer-based models that are readily accessible on the Hugging Face platform, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and mDeBERTa (He et al., 2023). These models are tailored for downstream tasks such as classification. Fine-tuning is a preferred practice in machine learning for deep learning models like transformers. It's advantageous over pre-training because it enables transfer learning and is computationally more efficient. Empirical results show that fine-tuning pre-trained models can lead to better performance, especially in cases where the dataset for the target task is small.

The initial step involves defining the task and selecting one or more publicly available datasets that comprise annotated labeled examples concerning the task, such as disinformation signal detection and hate speech detection. Each example within a dataset comprises a text and its associated label(s) for the task. The models are then fine-tuned using these datasets. During the training phase, the cross-entropy loss function is utilized for the classification task. The objective is to minimize this loss function and update the model's parameters (weights) using the gradient descent optimization algorithm. Sentence-based classifiers analyze an entire piece of text (sentence) as a single unit, and the model's prediction pertains to the whole text. These classifiers are further categorized based on the number of classes an example can have. Multi-Class and Binary Classification: The first category is multi-class classification, where there are multiple classes, but each text example is assigned to one and only one class.

A simpler subcategory of multi-class classification is binary classification, where the text example is assigned to one of two available classes. Multi-Label Classification: The second category is multi-label classification, where each text example can be assigned to multiple labels simultaneously. This task is particularly useful when categories are not mutually exclusive and can be applied cumulatively. A multi-label classification problem could also be approached as several binary classification problems, one for each class label. In conclusion, the choice of classifier and learning method depends largely on the nature of the classification task at hand. Each method has its strengths and weaknesses, and the choice between them should be made based on the specific requirements of the task. The ultimate goal is to develop a model that can accurately and efficiently classify new, unseen examples.

5.1.2 *Sequential Labeling Classification*

Sequential labeling classification, also known as sequential labeling, is a structured prediction problem in which categorical labels are assigned to each element in a sequence. This task is common in Natural Language Processing (NLP) and includes applications such as Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and chunking. Unlike sentence-based classification tasks, which provide a single label for the entire sentence or document, sequence labeling assigns a label to each input token. This approach allows for a more thorough comprehension of the input data since each token is analyzed within its specific context within the

sequence. Deep learning models have significantly advanced the field of sequence labeling. These models can automatically learn feature representations from raw input data, eliminating the necessity for manual feature engineering. These models typically utilize recurrent structures, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, designed to handle sequential data by maintaining a form of 'memory' of past inputs. This feature enables them to capture the dependencies between tokens in a sequence, which is vital for tasks such as sequence labeling. Despite these advancements, sequence labeling remains a challenging task. Issues such as handling long sequences, dealing with label imbalance, and incorporating prior knowledge or external resources are still areas of active research. In conclusion, sequential labeling is a potent tool in NLP, allowing for in-depth, token-level analysis of text. While deep learning has significantly advanced the field, there are still challenges to be addressed, and the work by He et al. (2020) provides a comprehensive survey of these recent advancements and ongoing challenges.

5.1.3 Few-Shot learning

Few-shot learning is a powerful technique for classification tasks that involves using Large Language Models (LLMs), such as Generative Pre-trained Transformers (GPT) models. These models have demonstrated impressive capabilities in general language understanding and can easily adapt to new tasks with one/few-shot learning. LLMs are pre-trained on massive corpora, and fine-tuning with a small number of labeled examples is used in few-shot learning scenarios. Prompt engineering is utilized to provide the necessary instructions along with a few examples to guide the model on the task. Parnami and Lee (2022) proposed several strategies for few-shot learning. These strategies include meta-learning, transfer learning, and hybrid approaches. The goal of these strategies is to make the most of the limited data available in few-shot learning scenarios. Meta-learning involves training models to understand the underlying learning process. This enables models to quickly adapt to new tasks with minimal examples. Transfer learning, on the other hand, entails training models on a large-scale task, and then fine-tuning them on a related, smaller task. The objective is to use the knowledge acquired from the large-scale task to improve the model's performance on the smaller task. Finally, the hybrid approaches integrate elements of both meta-learning and transfer learning to tackle few-shot learning problems. Despite the potential advantages of few-shot learning methodologies, they still face several challenges. One such challenge is the scarcity of data. Since few-shot learning involves learning from a limited number of examples, this presents a significant challenge. Traditional machine learning models typically require large amounts of data to function effectively.

Additionally, deep learning solutions can require significant computational resources and time, which may not be available in all situations. Furthermore, data preparation, such as data collection, preprocessing, and labeling, can be time-consuming. This is particularly challenging in few-shot learning scenarios, as each example can significantly affect the model's performance. Another challenge is the risk of hallucinations, i.e., generating irrelevant or false text. Therefore, it is crucial to provide precise prompts with detailed instructions. Providing examples of the task and specifying the expected output type can significantly reduce the risk of hallucinations. In conclusion, few-shot learning with LLMs offers a promising solution for classification tasks, especially when data is limited. However, it is essential to pay close attention to prompt design and the potential for hallucinations to ensure accurate and reliable model outputs. Further research and innovative solutions are required to overcome the challenges of few-shot learning.

5.1.4 Zero-Shot Learning

Zero-shot learning is an advanced technique in the field of machine learning. It involves training a model to perform a task without exposing it to any task-specific examples during the training phase. This technique is particularly useful when there is a scarcity of task-specific data available. In the case of Large Language Models (LLMs), zero-shot learning takes advantage of the model's pre-existing knowledge. The model is given a well-designed prompt that clearly outlines the task. With this prompt, the model can leverage its existing knowledge to perform the task at hand. LLMs are incredibly versatile, making them suitable for a wide range of tasks, including sentence-based classification and sequence labeling. Recent studies have shown that zero-shot and few-shot learning, combined with instruction-finetuned LLMs, can yield promising results in the field of clinical and biomedical tasks (Labrak et al., 2023). This approach demonstrates the potential of zero-shot

learning in specialized domains. Prompt design is a critical part of zero-shot learning with LLMs. A well-designed prompt can effectively guide the model's in-context learning, leading to improved performance. Recent advancements in this area were discussed at the 14th International Conference on Recent Advances in Natural Language Processing (Yinheng, 2023). In summary, zero-shot learning with LLMs is an innovative approach that enables models to perform tasks without the need for task-specific data. It leverages the model's pre-existing knowledge and a well-constructed prompt to perform the task. This approach has shown success in a broad range of tasks and domains, including clinical and biomedical tasks.

5.2 EVALUATION METHODOLOGIES

This section presents the evaluation methodologies that can be used to assess the performance of machine learning models that were developed for detecting disinformation signals. There are three primary methods for evaluating the reliability and robustness of these models: using a benchmark corpus and calculating metrics, collecting user feedback, and performing active annotation.

5.2.1 *Comparison with Gold Standard Corpus*

To evaluate the effectiveness of disinformation signal models, various techniques can be used to ensure accuracy and reliability. One widely used method is the gold standard corpus approach. This approach involves using a set of data that has been manually annotated and is known to be accurate. The data in the gold standard corpus is not used during the training process of a model and serves as a benchmark for evaluating its performance. Benchmark datasets are crucial in machine learning research. They help coordinate researchers around shared research problems and measure progress towards shared goals (Koch et al., 2021). By comparing the results obtained from a trained model with the annotations in the gold standard corpus, we can determine its accuracy level and identify areas for improvement. To gain valuable insights into the effectiveness of the approach, various metrics such as micro-F1, macro-F1, precision, recall, accuracy, and confusion matrix are analyzed. Micro-F1 measures the harmonic mean of precision and recall, while macro-F1 calculates the weighted average of F1 scores. Precision measures the proportion of true positives among the total number of positive results, while recall measures the proportion of true positives among the total number of actual positives. Through the confusion matrix, the performance of each class is depicted to perform error analysis and identify the strengths and weaknesses of each model on real-world unseen data. The labels used by the metrics are numerical values that represent the index of a specific class (e.g. 0 and 1 for the binary classification instance). Using these metrics, potential biases or inaccuracies in the model can be identified, resulting in refining the techniques accordingly. This helps to improve the accuracy and reliability of the approach, making it more effective in detecting and combating disinformation. Ultimately, the gold standard corpus approach is an essential tool in evaluating the performance of disinformation signal models and ensuring that they function as intended. However, this approach has limitations because it relies on a static dataset and may not capture the ever-evolving nature of disinformation. As Dueben et al. (2022) point out, benchmark datasets provide stable points of comparison and coordinate scientists around shared research problems. Improved performance on benchmarks is considered a key signal for collective progress. Therefore, while the gold standard corpus approach has its limitations, it remains a valuable tool in the evaluation of disinformation signal models.

5.2.2 *User Feedback Collection*

In addition to benchmarking model performance against a gold standard corpus, user feedback collection emerges as another vital assessment method for disinformation detection models. More particularly, this methodology involves developing systems that can explicitly present disinformation signals to users, such as deep learning algorithms or interfaces that highlight suspicious content. These systems must be carefully designed to be effective, user-friendly, and accessible to a broad range of users (Hotjar, 2023). Different levels of feedback can be collected, either explicitly or implicitly. Explicit feedback is provided directly by the user and is clear and straightforward. For example, users can classify information as true/false or correct/false. This can help identify false positives and false negatives. Alternatively, feedback can be more detailed by asking

users to provide additional context, verify the source of information, or relate the information to other known facts. This implicit feedback can provide a more detailed understanding of the disinformation, thereby enhancing the precision of the disinformation signal detection mechanisms. Sahu et al. (2019) propose a new framework for collecting implicit user feedback along with explicit feedback for a movie and video recommender system. They argue that explicit feedback is not always available and not always unbiased, so implicit feedback seems to be a better option for user preference collection. On the other hand, Haruyama et al. (2023) discusses the challenges of collecting explicit feedback. They provide a behavioral account of explicit feedback's sparsity problem by modelling a range of constructs on the rating and review intentions of US food delivery platform users. Their model revealed that standard industry practices for feedback collection appear misaligned with key psychological influences of behavioral intentions. This methodology emphasizes the importance of user involvement in combating disinformation and can provide valuable insights into the effectiveness of disinformation signal detection systems. However, the success of this methodology depends heavily on user participation and the quality of their feedback. Therefore, it is crucial to consider user interface design, engagement strategies, and feedback validation mechanisms. These considerations can help improve the quality of feedback collected and ensure that the detection mechanisms are continuously improving to combat disinformation effectively.

5.2.3 Active Annotation

Active annotation is a process that is employed in the evaluation of machine learning models, particularly when dealing with complex classification problems. This process is similar to active learning, where the most challenging cases are presented to annotators for further examination and feedback. In situations where a classification problem proves to be intricate for models to learn, dataset annotators gather external feedback. This feedback is used to re-evaluate a previously annotated segment that might have been erroneously labeled and could be the reason for the classification difficulty. The concept of active annotation was first introduced by Vlachos in 2006. It is a framework used in semi-supervised learning for creating training material. Initially, an unsupervised approach is used to annotate the data, which detects and corrects errors automatically. Then, a human annotator reannotates the data. This approach has been applied to named entity recognition in the biomedical domain with good results. Active annotation has certain advantages over active learning, as it doesn't require seed annotated data and maintains the reusability of the data. The effectiveness of the framework is measured by the reduction in checked instances needed to achieve a particular performance level and the increase in performance for a fixed number of checked instances (Vlachos, 2006). Through active annotation, difficult-to-learn cases are carefully selected based on various criteria such as ambiguity, subtlety, or similarity to true information. Annotators are presented with these cases and asked to provide their feedback, thereby contributing their expertise and judgment. They can identify and correct potential errors in the dataset, thereby refining and enhancing the performance of the machine-learning models. This process plays a crucial role in ensuring the reliability and robustness of machine learning systems. In conclusion, active annotation is a critical factor in the evaluation of machine learning methodologies. It not only affects the performance of the models but also provides valuable insights into the challenges and strengths of an approach.

5.2.4 TITAN Training & Evaluation Methodologies

The TITAN project has devised three training methods to meet its objective of detecting disinformation signals, which are sentence-based classification, sequential labeling classification, and few-shot learning. By incorporating these techniques, the project seeks to evaluate text context at both the sentence and token level and adapt to new disinformation tactics to ensure that its models are both precise and efficient. To evaluate the performance of its models, the project will utilize two evaluation methodologies, namely, the gold standard corpus approach and the user feedback collection method. The former will serve as a benchmark for the models, allowing for a comparison of results against a manually annotated dataset. The assessment will involve using metrics such as macro-F1 score, precision, recall, accuracy, and confusion matrix. Nevertheless, the gold standard corpus approach has limitations, and the project will overcome these by implementing the user feedback collection method to supplement its evaluation approach. This method

entails collecting user feedback explicitly to present disinformation signals, thereby enhancing the precision of the detection mechanisms in real-world scenarios. The TITAN project aspires to establish robust and adaptable systems for detecting disinformation signals by combining the two evaluation approaches. This system will adjust to new disinformation patterns and evolving user feedback, making it more effective in detecting and combating disinformation. Ultimately, both training and evaluation methodologies will enhance the accuracy and reliability of TITAN's disinformation signal detection mechanisms, making it a valuable tool for the identification and mitigation of disinformation.

6 PROPAGATION IMPACT MEASUREMENT VIA SOCIAL MEDIA MONITORING

Propagation Impact Assessment (PIA) will predict, how efficiently a new posting to social media will spread (i.e. 'propagate'). The purpose is to provide a user an insight of impact, potentially to think twice before submission.

6.1 MAIN FACTORS

The approach used for estimating the propagation impact is based on statistics and supervised machine learning. Based on historical data, a machine learning model is trained to predict the future propagation impact of an article or social media post, i.e. forecast the consequences if a user shares something on social media.

To create the estimate, factors affecting the propagation of a message must be identified. In the field of data science these factors are called features. They describe some characteristics of a dataset in numerical format, e.g. average number of words in a social media post (calculated over a set of posts). The chosen set of features is used as an input for a machine learning model. Besides the input, the output (target) for the model needs to be defined also. In the case of TITAN, the target of the PIA model is relative, meaning that the model will tell if the post will get less or more shares than the user's posts typically.

Petrovic et al. (2011) proposed a machine learning approach that predicts if a message is going to be retweeted on Twitter. They made a division of features into two categories; social and content-based features. Social based features describe the author of a post and they include features such as number of posts made by the author and number of followers the author has. The content-based features describe the characteristics of a single post and include features such as number of hashtags in the post and length of the post. To be able to calculate the features, access to different sources of information is needed. Social features may contain features such as centrality values which describe the role and importance of the user in the community/network of users regarding information flow. To be able to calculate these sorts of features, access to user social graphs (network that describes the connections between users, e.g. "user A is following user B") is needed. This information needs to be provided by an API of a social media service. Petrovic et al. (2011) found that social features have a more important role in retweetability than content-based features. Also using time-related information improved the accuracy of the model. Most important social features (with most weight on the end result) were found to be "listed" (user is followed in lists), number of followers and number of friends. The result is quite expected since the more networked the user is, the more people their tweets will reach.

An important detail mentioned by Petrovic et al. (2011) is also using trending topics and novelty (how much a text differs from other known texts) as features. Another finding was that the hashtag has weight on retweeting behavior; some topics are more likely to be retweeted than others. Related to topics, Jasser et al. (2022) studied the difference of information spreading between controversial and non-controversial information. Their study suggests that the activity around controversial content is higher than with non-controversial ones.

When considering using novelty and trending topics as features, it must be noticed that they change over time, i.e. they have temporal dimension; new topics emerge, some become more popular than others and fade away over time. This means that when creating numerical variables (features) from novelty and trendiness of the topic, they must reflect the current state of the world. When a topic is detected in the text, feedback of

the current state of novelty and trendiness must be obtained from some service that holds information about what is happening in the world (e.g. search via social media API).

Suh et al. (2010) studied the factors impacting retweeting. They calculated several features (similar to Petrovic et al.) and analyzed their effect on retweeting behavior. Principal component analysis was used to find redundant features and e.g. occurrence of retweet and number of followers were found to have a high correlation and presence of hashtag and user mentions a negative correlation. Also retweet rate of different hashtags and URLs mentioned in a post was as studied and differences between rates were found. Suh et al. (2010) found that amongst content features URLs and hashtags have a strong relationship with retweetability. From the contextual features side, number of followers and followers contribute to the retweetability.

Kupavskii et al. (2012) point out the epidemiological point-of-view of retweeting behavior; if the users are considered to be nodes in a network, the retweeted message spreads from one node to neighboring nodes similarly as an infection. Their model predicts the size of retweet cascades over time and they use information that can be obtained when retweeting of a message is already happening in the social network. Besides social and content features they use time sensitive features (e.g. average retweet ratios up to a point in time) and features of the infected nodes (users that have been reached by the retweeted message). Similarly, Zaman et al. (2014) used a Bayesian approach for predicting popularity of a tweet by using the retweet path accumulated in a small period after the tweet has been created.

6.2 PROPOSAL FOR TECHNICAL IMPLEMENTATION

The proposal for technical implementation follows the findings from the literature. Information affecting the propagation of a post in the future can be found from the social characteristics of the user and content-based characteristics of the post. Some of the characteristics, such as presence of URLs and number followers, have more impact on the propagation of the message than for example the total number of messages posted by the user. The propagation impact assessment component in TITAN should use factors identified as important by earlier studies. However, the feasibility for calculating all factors must be analyzed by taking the limitations set by data sources into account. For example, information about social networking of the user is obtained via social media service APIs but the APIs provide only limited set of functionalities and some features identified in the literature might be difficult or impossible to obtain. Thus, the set of factors/features used for assessing the propagation impact is a compromise between the findings of the literature and technical limitations.

Same applies to the output of the propagation impact assessment component. In an ideal scenario an exact number of people who have seen a post would be obtained. However, this is not realistic due to the fact that a (machine learning) model is able to provide an estimate with limited accuracy since all the factors affecting the propagation cannot be taken into account. Also quantizing the output into categorical variables (e.g. less than typical number of shares, more than typical number of shares) should be done to simplify the modelling problem from the machine learning point of view.

The data source used in the studies found in the literature was mostly Twitter. Due to the API pricing changes done by X (former Twitter), accessing the X APIs is not feasible at the moment. Mastodon was chosen as data source for the development and the representativeness of the APIs and data in general is most like more limited compared to X. This also influences the propagation impact assessment model.

The development of the PIA component should be iterative. Experiments using factors found in literature as an input for PIA should be done and new features should be added as input through experimentation.

6.3 EVALUATION

The evaluation of the PIA component can be done in two dimensions. First is the technical evaluation of the model by using a historical dataset with ground truths (numerical values for propagation that happened in real-life). The dataset can be split in temporal plane into training set and testing set. The machine learning model used for PIA is trained with the training set and then the performance of the trained model is evaluated with the training set. This gives an idea how well the machine learning model performs when predicting the propagation impact assessment when considering performance metrics such as accuracy or F1-score.

The second dimension in evaluation is the concept of PIA. Questions such as in which case the users would utilize PIA in their decision making and it would affect their decision on sharing something in social media are relevant for assessing the usefulness of the concept.

7 PERSONALIZED MICRO-LESSONS FOR MEDIA LITERACY AND CRITICAL THINKING

7.1 PERSONALIZED MICRO-LESSONS STRUCTURE

Personalized Learning is the customization and adaptation of educational methods and techniques so that the learning process is better suited for each individual learner, with their own unique learning style, background, needs, and previous experiences. Personalized learning aims to drive learners' engagement in a specific field of learning [1]. Micro-lessons are an educational concept: *“Targeted lessons for students based on their knowledge of the subject matter rather than grade or age.”* [2] Micro-lessons mainly refer to a structured digital resource that uses information technology to present fragmented learning content, process and expanded materials in accordance with cognitive rules. They are mainly designed and provided online through a computer as well as mobile phones so that learners can easily download, watch and learn at anytime, anywhere [3]. Micro-lessons can be associated with the micro-teaching process, which was firstly introduced at Stanford University in the 1960s and derived from the practices in teacher education involved teachers in training, delivering short lessons, recording them for playback, critiquing both individually and as a group, and receiving feedback from their peers [4]. Talking about micro-learning we could focus on its differences compared to macro-learning focusing on specific characteristics [5] as:

- it offers informal learning opportunities that take place outside of the traditional classroom
- it works at the lowest level of learning, not as a summary or new set of skills or behaviors
- mobile-based micro-learning combines elements of mobile learning and micro-learning by delivering small learning units and short-term learning activities through mobile devices in a manner that can be personalized, adaptive, ubiquitous, and context-aware
- micro-content refers to information whose duration is determined by a single topic, idea, or concept and is accessible at a single location online, being suitable for presentation in handheld devices, emails, and web browsers

On the basis of these characteristics, we understand that the micro-lessons offer the learner the opportunity to follow a course according to his/her own learning needs and expectations and at a flexible time. This promotes and encourages knowledge and is in line with the principles of adult education, where the approach is learner-centered. Through the personalized micro-lessons, Titan Socratic dialogues aim to ensure that users, who exhibit pronounced strengths or weaknesses in certain dimensions, receive micro-lessons that are finely tuned to their unique cognitive profiles. To identify a robust and pertinent linkage between the critical thinking dimensions and the sequential stages of Socratic dialogue, a consensus was established through joint discussions among researchers and independent evaluators.

7.2 MEDIA LITERACY AND CRITICAL THINKING

Our rationale regarding the methodology of the micro-lessons is based on TITAN's understanding/definition of critical thinking. Micro-lessons aim to provide the knowledge and skills required to improve on the main components of critical thinking through Socratic dialogues. TITAN's understanding/definition of critical thinking adopts as a starting point the DELPHI REPORT (Facione, 1990): critical thinking is “purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based [...] The ideal critical thinker is habitually 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” (Facione, 1990). We also adopt the Reports's

educational ideal which holds that “educating good critical thinkers [...] combines developing Critical Thinking skills with nurturing those dispositions which consistently yield useful insights and which are the basis of a rational and democratic society” (Facione, 1990).

In particular, TITAN focuses on developing the following **six abilities** of critical thinking (Facione, 2011) (Figure 11):

1. **Interpretation** (categorization, decoding significance, clarifying meaning)
2. **Analysis/ Deduction** (examining ideas, detecting arguments, analyzing arguments)
3. **Evaluation of arguments** (assessing claims, assessing arguments)
4. **Inference** (querying evidence, conjecturing alternatives, drawing conclusions)
5. **Explanation/Recognition of assumptions** (stating results, justifying procedures, presenting arguments)
6. **Self-regulation** (self-examination, self-correction)

In addition, we further focus on the following affective **dispositions** (Facione, 2011):

1. Inquisitive
2. Judicious
3. Truth-seeking
4. Confident in Reasoning
5. Open-Minded
6. Analytical
7. Systematic

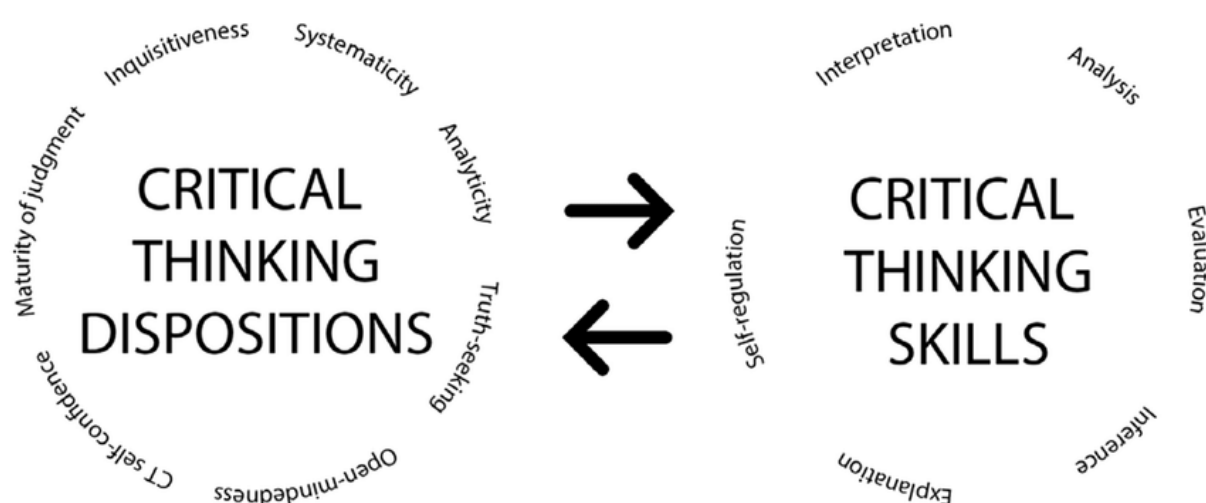


Figure 11. Facione’s Core Critical Thinking Skills & Dispositions, available from: *International Journal of Technology and Design Education*

Since TITAN aims at enhancing critical thinking against disinformation in particular, the micro-lessons and dialogues should help users improve on domain-specific critical thinking, that is, **digital critical thinking** (Paul & Elder, 2020) (Figure 12). We define ‘*digital critical thinking*’ as the ability to evaluate the credibility of a media item and, in particular, the ability to identify disinformation tactics/techniques and narratives (Machete & Turpin, 2020). Following established pre-bunking strategies of media literacy (e.g. Van der Linden, 2023) and conceptions of digital critical thinking, such as Paul & Elder (2020), the evaluation of a media item’s credibility involves Facione’s six generic critical thinking abilities adapted into domain-specific critical thinking skills (van Laar, 2019; Elder & Paul, 2020) focusing on tactics and narratives of disinformation. Specifically, digital critical thinking includes the ability of the user to question and evaluate the following: **(a)** an item’s credentials (evaluation); **(b)** an item’s logic, evidence, and rhetoric (justification); **(c)** an item’s narratives or point of view,

purpose, agenda, concepts, implications, and assumptions (clarification); **(d)** the user's own bias, stereotypes and worldviews (self-regulation). In addition, **(e)** it involves the ability to conjecture and generate alternative viewpoints (novelty; inference).

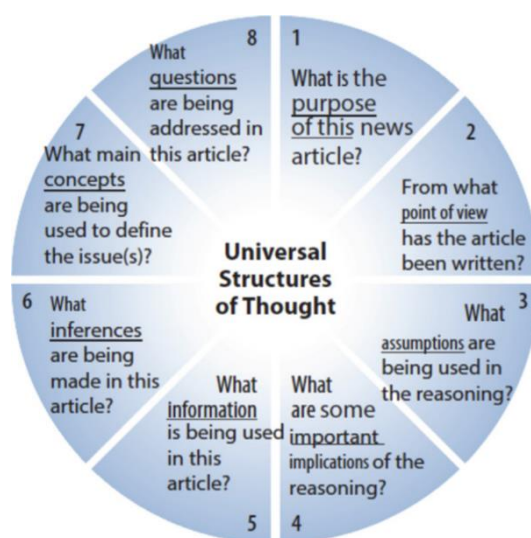


Figure 12. Digital Critical Thinking Skills (Paul & Elder, 2020)

To that end, micro-lessons will also provide the user with the opportunity to learn/explore different ways of evaluating the credibility/appropriateness of digital sources through the introduction of various tests and methods of evaluation. Such examples include M. Caulfield's *SIFT Method*, Constitutional Rights Foundation's *SMART Test* and other similar credibility evaluation tools, e.g. the *CRAAP Test* (Blakeslee, S. (2004). The users will be given the opportunity to acquaint themselves with state-of-the-art media literacy material such as *lateral reading* and other fact-checking processes in an effort to combat mis/disinformation and promote critical thinking. In addition, microlessons will give users the opportunity to familiarize themselves with generic critical thinking skills focusing on argumentation and rhetoric in order to become more able to distinguish *good* from *bad* reasoning (Bowell & Kemp, 2014).

Moreover, TITAN's micro-lessons are based on established prebunking/inoculation techniques which familiarize the users with common, recognized disinformation tactics and narratives, as well as, introduce them to useful media literacy material. In particular, we will recruit *passive* and *active* prebunking techniques in order to hone the user's reflexes and skills, thus building up their resistance against a wide range of common disinformation practices and narratives. *Passive* prebunking interventions will integrate videos, infographics and text (Harjani et al., 2022) into the dialogue process. Micro-lessons will also point users to games created based on inoculation science as *active* forms of prebunking (e.g., the Bad News game created by DROG and University of Cambridge or Harmony Square created by GEC, CISA, Tilt and the University of Cambridge). In addition to these interventions, TITAN's users will have the opportunity to learn how to 'mobilize to respond quickly to misinformation by sharing facts', how misinformation can be 'unstuck' (debunk) through detailed refutation, as well as other key concepts and characteristics of effective debunking techniques.

Finally, we will supply users with material regarding the standards of good journalism, charts of media bias, lists of recognized fact-checking tools/websites, and information about international institutions, non-profit organizations and local governments which specialize in combating mis/disinformation.

7.3 METHODOLOGY OF MICRO-LESSONS COLLECTION

Based on the aforementioned dimensions of digital critical thinking, we have created **five categories** of micro-lessons for media literacy in critical thinking in order to support the dialogue process (Lewandowsky et al., 2020).

1) Micro-lessons on Disinformation Tactics & Techniques

These micro-lessons will offer users the domain-specific information, tools and criteria to evaluate the credibility of media items based on the recognized, common tactics and narratives of disinformation. In particular, the micro-lessons will present the common fallacies, rhetorical ploys, media bias, techniques of manipulating sources, etc. The micro-lessons will enable users to evaluate the logic and rhetoric of a media item, to evaluate the credentials of the item; to think critically over the agenda and bias of the source; to seek alternative sources/viewpoints and correct false beliefs. The structure of the presentation of each tactic will include the following: **(a)** definition and examples of the tactic; **(b)** its warning signs/characteristics; **(c)** its implications; **(d)** a description how the tactic works and affects us; **(e)** interventions to prevent, protect and correct it.

2) Micro-lessons on generic Critical Thinking abilities

These micro-lessons will introduce users to key concepts and techniques from the field of informal logic required for argument evaluation and interpretation of evidence and concepts. The aim is to enhance the generic critical thinking abilities of users which will deepen their ability to evaluate the credibility of an item by focusing on its logic and rhetoric (e.g. understanding arguments, spotting rhetorical ploys, inconsistencies, fallacies, interpreting numerical evidence, explanation, etc.). In effect, the users will sharpen the following generic skills and subskills based on Facione's identification of critical thinking abilities:

1. **Interpretation** (categorization, decoding significance, clarifying meaning)
2. **Analysis/ Deduction** (examining ideas, detecting arguments, analyzing arguments)
3. **Evaluation of arguments** (assessing claims, assessing arguments)
4. **Inference** (querying evidence, conjecturing alternatives, drawing conclusions)
5. **Explanation/Recognition of assumptions** (stating results, justifying procedures, presenting arguments)

Micro-lessons on generic critical thinking abilities will offer understanding of central informal logic concepts, such as what deductive/inductive reasoning is, how a proper argument is structured, how the conclusion follows logically, as well as common fallacies and their respective definitions (Bowell et al., 2014; Sinnott-Armstrong, 2015).

3) Micro-lessons on generic Critical Thinking dispositions

These micro-lessons are designed based on Facione's main affective dispositions. They consist an effort to enhance and explore the critical thinking dispositions of the users against mis/disinformation mainly through a process of self-reflection/self-regulation (e.g. how and why they built confidence on their own reasoning, how open-minded when conversing/engaging with a text they can be, whether they are interested in seeking the truth behind an argument/statement, etc.).

4) Micro-lessons on Disinformation Narratives & Stereotypes:

The micro-lessons also aim at familiarizing the users with popular disinformation narratives (e.g. regarding climate change, immigration, etc.) that make them vulnerable to disinformation (or become "sticky") and their respective characteristics. The structure of the presentation will be based on the following questions: What narratives are out there? How do they affect us and why are we vulnerable to them? Why disinformation narratives become "sticky" and how can we rewrite them?

5) Micro-lessons on Bias & Heuristics

The micro-lessons also aim at familiarizing the users with the inherent cognitive bias that facilitates the spread of disinformation. Cognitive science has established that we understand the world not through direct perception but through inferential processes facilitated by heuristics and bias that we are not aware of (e.g. Kahneman, 2011). We are prone to serious errors of judgment and disinformation which can be avoided to some extent when we understand their basis. We will organize the micro-lessons about bias as to provide information about the following:

1. Two Systems of Thinking (Kahneman, 2011)
2. The illusion of objectivity
3. Heuristic methods

4. Types of bias specific to disinformation: e.g. confirmation and partisan bias; fundamental attribution error
5. How to become aware of these bias
6. How to deal with these bias

8 USER ENGAGEMENT THROUGH GAMIFICATION METHOD

8.1 LITERATURE REVIEW ON WAYS TO ENHANCE ECOSYSTEM SUSTAINABILITY AND USER ENGAGEMENT

As misinformation occurrence and spread increases via social media platforms, there is a parallel emergence of a citizen movement that becomes increasingly conscious of the information that it consumes. The pivotal role of citizens in the fight against disinformation will be enhanced in the course of TITAN, which will support and empower citizens to conduct (individually and collaboratively) effective investigations for understanding whether statements at hand are true.

A critical component of the success of the TITAN ecosystem will be the creation of the necessary incentives framework that will ensure citizen engagement across all ecosystem phases (pre-initiation, initiation, early stages, maturity, etc.). TITAN will design and co-design with citizens, appropriate incentives (incentive theory of motivation, motivational theories, intrinsic and extrinsic motives, gamification, incentives aligned with the Responsible Research and Innovation approach, etc.) to make citizens, agents of change through their direct participation and active, on-going engagement with TITAN. These incentives will ensure, the sustainability of the TITAN ecosystem anchored in Responsible Research and Innovation (RRI) principles (ethically acceptable and socially desirable to benefit the society at large), European values and a trust-by-design approach. Motivation plays a fundamental role in driving human behavior and achieving personal and professional success. The word ‘motivation’ comes from the Latin word mover or ‘to move’ (Alshmemri, Shahwan-Akl, & Maude, 2017) and it can be defined as ‘how to provide something to a person to drive him/her to do something’ (Ruthankoon & Ogunlana, 2003, p. 333). Over the years, numerous motivation theories have been developed by psychologists and scholars to explore the complex factors that influence individual motivation. These theories provide insights into why individuals act, persevere, and strive toward their goals. In the sections that follow we provide a literature review of relevant theories in the area including motivational theories, motives, as well as gamification aspects, relevant for TITAN.

8.2 MOTIVATION THEORIES & MOTIVATION TYPES

Motivation theories enhance our understanding regarding what drives human behavior. There are several different theories about what constitutes and creates motivation. Some of the most prominent ones are described below:

Maslow's Hierarchy of Needs (HON): Proposed by Abraham Maslow (1943, 1954), this motivational theory in psychology, indicates that individuals are motivated by a series of needs - motivated to achieve certain needs - that are grouped as distinct hierarchical levels within a pyramid. Maslow's theory puts forward the idea that “individuals move through a fundamental number of hierarchical motivations, in a unique order, based upon both physiological and psychological needs” (McLeod, 2007, p.54). The five human needs according to Maslow's motivational hierarchy, are ordered having the physiological (survival) needs at the bottom, and the more creative and intellectually oriented ‘self-actualization’ needs at the top. More specifically these needs include: physiological needs at the base of the pyramid, followed by safety-security, love and belongingness, self-esteem, and finally self-actualization needs. According to Maslow, once lower-level needs are satisfied, individuals seek to achieve higher-level needs (Figure 13). Maslow's motivational theory has been widely adapted within educational learning theory (Mittleman, 1991) as it aligns with the principal aim of education to foster learning and make this process meaningful to an individual, impacting upon and becoming useful to their lives (McLeod, 2007).



Figure 13. Maslow's Hierarchy of Needs (Maslow, 1943)

Herzberg's Two-Factor Theory of Motivation (or Herzberg's Dual Factor Theory): In 1959, the American psychologist Frederick Herzberg highlighted the distinction between hygiene factors and motivators separate sets of mutually exclusive factors in the workplace that either cause job satisfaction or dissatisfaction (Herzberg, 1966; 1982). Herzberg's theory identifies these two factors as the key factors as two different continua, each with its own set of factors, that contribute to job satisfaction, relate to self-growth and self-actualization (Table 8). In particular, 'hygiene factors', such as salary, job security, and working conditions, are essential to prevent dissatisfaction ('need to avoid unpleasantness') but do not necessarily lead to motivation. 'Motivators', on the other hand, include factors like recognition, achievement, and responsibility, which contribute to job satisfaction and intrinsic motivation, associated to the need of the individual for self-growth and self-actualisation (Alshmemri, Shahwan-Akl, & Maude, 2017).

Table 8. Overview of Motivators and Hygiene factors

Motivators - encourage satisfaction	Hygiene Factors - prevent dissatisfaction
• Achievement • Recognition • The work itself • Responsibility • Advancement • Growth	• Bureaucracy • Relationships • Work conditions • Status • Salary • Job Security

Even though Herzberg's approach is being primarily used in a business and organizational context (organizational psychology), to motivate employees and boost morale and productivity; existing research in the area indicates that it is a great tool when used to **motivate learners** (Katt and Condly, 2009; Ibrahim et al., 2023). When applying Herzberg's two-factor theory in a blended e-learning environment, Wu et. al., 2010 found that the primary determinants of student learning satisfaction (motivators) are primarily influenced by performance expectations, and learning climate as well as by system functionality, content feature, and interaction.

Vroom's Expectancy Theory (Expectancy Theory of Motivation): Developed by Victor Vroom in 1964, the Expectancy Theory focuses on the belief that motivation arises from an individual's perception of the relationship between effort, performance, and desired outcomes. According to Vroom's theory, individuals are motivated when they believe that their efforts will lead to successful performance and that this performance will result in valued rewards or outcomes (expectancy). As such the motivation of the selected behavior (over others) by individuals is determined by the desirability of the expected result. As such people will behave in a way that results in the best associated outcome; and the more they value the outcome the more motivated they will be to achieve it. The cognitive process of the motivations of decision making are central to this theory.

Vroom's theory encompasses two related models the valence model and the force model; where the former aims to capture the perceived attractiveness (valence) of an outcome by aggregating the attractiveness of the associated resultant outcomes and the latter (Geiger & Cooper, 1996). The force model of expectancy theory aims to "capture motivational force to act by associating the expectancy of resultant outcomes and their individual valences" latter (Geiger & Cooper, 1996, p:114). The applicability of this theory is primarily seen in business settings as well as in **educational settings** that has been widely demonstrated (Geiger & Cooper, 1996, Lokman et al., 2020, Motevalli, et al., 2020; among others).

As **motivation and learning processes** share a deep connection (Gopalan et al., 2017) since motivation is the force that encourages learners to succeed in educational matters. Acknowledging that the learning process is an endless lifelong process it is crucial to continuously achieving high motivation. Our examination of the key motivational theories indicated that from the one side we have theories that focus on the relationship between internal needs and the associated effort to fulfil them, such as Maslow's and Herzberg's theories. However, theories that separate effort arise from motivation, performance and outcomes like Vroom's theory.

8.3 MOTIVATION TYPES- INTRINSIC & EXTRINSIC MOTIVATION THEORIES

Existing literature indicates that motivation can be divided into two different types of motives, namely: (a) intrinsic motivation (internal motivation – intrinsic motivation theory) and (b) extrinsic motivation (external motivation – extrinsic motivation theory) (Hippel and Krogh, 2003; Lerner, and Tirole, 2002, among others), as seen in the sections that follows.

8.3.1 Intrinsic motives

Intrinsic motives (Intrinsic Motivation Theory) are internal motives that are directly connected to the act of performing a task (i.e., peer recognition, satisfaction in learning, enjoyment in performing a task, etc.,) (Hippel and Krogh, 2003) rather than external factors or rewards. Driven by the desire to engage in curiosity-driven, enjoyment-driven and/or personal-development driven behaviors, intrinsic motivation represents a natural motivational tendency that is long-lasting and self-sustaining.

- *Common intrinsic (internal) motivations:* curiosity, learning, mastery, meaning, belonging, autonomy, enjoyment, etc. (Maslow, 1943; White, 1959; Harlow, 1958).
- *Advantages:* intrinsic motivation is long-lasting and self-sustaining.
- *Disadvantages:* it requires a lengthy and well-planned process that has a slow behavioral effect as opposed to external motivation.

Intrinsic motives can have an important motivational role in TITAN:

- **Initial phases** (i.e., consideration stage): the user becomes aware of TITAN and considers using the TITAN service. Intrinsic incentives could trigger the self-desire to engage in curiosity-driven, personal-development motivation and mastery.

Indicative Examples for TITAN system:

TITAN messages should trigger curiosity and learning enjoyment associated with critical level enhancement.

- **Curiosity:** "Ever wondered why certain news items become widely accepted? Let's explore the roots of popular beliefs together!"
 - **Mastery:** "Challenge yourself! Which is the degree to which this news item is true? How can you apply your existing knowledge to discover this?"
- **Usage/adoption phases**, when the user has adopted the service in order to trigger curiosity (for finding the true facts), learning (how to advance their critical thinking, advance their skillset, etc), purpose (how the TITAN aspects could empower the user), belonging (in the TITAN ecosystem and community), and enjoyment (from personal achievement and discovery).

Indicative Examples for TITAN system:

TITAN messages to the users should be written in an intriguing way so as to drive for example:

- **Purpose:** "Understanding this concept/aspect could empower you to make more informed decisions."
- **Enjoyment:** "You can think of the TITAN questions as intellectual puzzles! So let's explore together this particular topic/information item, etc?"

8.3.2 Extrinsic motives

Extrinsic motivation (Extrinsic Motivation Theory) originates from influences outside the individual as opposed to intrinsic motivation. The incentive theory evolves around external rewards and punishments as the primary motivators of people's behaviors. In extrinsic motivation, the performance of an activity is unrelated to the nature of the task (i.e., economic motives, direct compensation, anticipated return, etc) (Lerner, and Tirole, 2002).

- **Common extrinsic (external) motivations:** rewards such as money (monetary rewards) and points (non-monetary rewards) for demonstrating the desired behaviour as well as competition and the fear of failure and punishment (Skinner, 1953; Hull, 1943).
 - *Advantages:* extrinsic motivation is easy to foster and relatively fast to implement.
 - *Disadvantages:* it has a relatively short-lasting effect that needs continuous efforts to maintain having significant cost-implications.

Extrinsic motives can have an important motivational role in TITAN, especially non-monetary incentives.

- **Initial phases** – extrinsic incentives would provide limited value in the initial phases of TITAN service consideration.
- **Usage/Adoption phases** – extrinsic, non-monetary incentives, would be value-adding during the TITAN service usage. Recognition could act as a positive incentive that can be used to motivate TITAN user and reward them with some sort of praise to the broader community, (i.e., user of the day, and long TITAN service award, etc).

Indicative Examples for TITAN system:

TITAN system could include the following:

- **Challenges:** For example, challenges with specific news items could also be included engaging with all TITAN users. The users could also define (personalisation of experience) in the beginning and during their interaction with TITAN whether they would like to be “challenged” and the frequency (i.e., weekly challenges, monthly, etc.). This could also work as a trigger for TITAN to engage with users: “Let’s embark your weekly challenge!”
- **Recognition:** “Your insights matter! Share your thoughts and ideas and let’s discuss your value-adding thoughts.”
- **Community Engagement:** “Join our community of critical thinkers and be part of our learning community!”
- **Reward:** “Unlock new levels of critical thinking!”
- **Personalisation:**
 - “You (or Name of the user) can tailor your TITAN learning experience! Do you have any preferred news themes/topics – news items, that you would like to focus on?”
 - “Choose your own critical thinking learning journey!”

8.3.3 Gamification: an emerging motivational factor

In today’s experience-based ecosystems, value is driven not only by using a product or service, but it is also attached to the experience that users have when using these products or services. This experience is frequently enhanced with the use of gamification elements. Gamification entails the use of game design elements in nongame contexts (Deterding et al., 2011). Gamification has been utilized in diverse domains, in addition to gaming, such as business, tourism (Xu et al., 2017), marketing, education (Hamari et al., 2014, Dicheva et al., 2015; Sailer & Homner, 2020; Wang et al., 2022), among others aiming to enhance users’ motivation and engagement.

- *Common gamification mechanisms:* point system, levels, leaderboards, avatars, progress bars, virtual goods/currency, badges/achievements (Dicheva et al., 2015)
 - o *Advantages:* offer better and more visible sense of progress, increase users’ participation and engagement (Dicheva et al., 2015)

Disadvantages: short term engagement, overemphasis on rewards

A key part of successfully gamifying experiences is to adopt a clear framework, for designing gamified experiences. Such game design frameworks include the following prominent ones: *the MDA model, the Elemental Tetrad model, and the 6D Framework*.

The **MDA (Mechanics - Dynamics - Aesthetics) model**, proposed by Hunicke, Leblanc, and Zubek (2004) (Table 9), describes the three components that should be taken into consideration in the elaboration of games. In this context, the (a) Mechanics: are the components of the game in relation to data representation and algorithms; (b) Dynamics: are the behavior of the mechanical executions that act in the player’s actions and the results; and (c) Aesthetics: are the desired emotional responses of the players as they interact with the gaming system.

Table 9. The MDA Model (Hunicke et al., 2004)

Game Components		
Rules	System	Fun
Design Counterparts		
Mechanics	Dynamics	Aesthetics

• Controls (timers, skill test) • Components (badges, avatars, leaderboards) • Courses (quests, levels, groups)	• Context • Constraints • Completion • Cooperation • Consequences • Choices • Competition • Chance	• Challenge • Creativity • Compliance • Confidence • Contribution
---	---	---

The MDA framework is the most widely accepted model and utilized approach to game design, despite the criticisms associated with its focus on game mechanics as opposed game design aspects, and its unsuitability for all types of games (Walk et al., 2017).

- The **Elemental Tetrad game design model** that has been developed by Schell (2008), expands the MDA model (presented above). According to Schell (2008) the four interconnected parts of this game design framework (Figure 9), include:
 - 1. *Aesthetics*: usually a players entry point, aesthetics include the appearance, sounds, and other sensations of the game;
 - 2. *Narrative*: this involves the sequence of events that will be presented to the player/user;
 - 3. *Mechanics*: that is the rules that govern the game (or the specific environment), including the game objective and how players will achieve it;
 - 4. *Technology*: that is the set of materials and interactions that make the game possible (i.e., digital or physical artefacts).

According to the Elemental Tetrad game design model (Figure 14) these four key characteristics are required, introducing essentially two new components in relation to MDA framework: story/narrative and technology, which are according to Schell are less visible to the player.

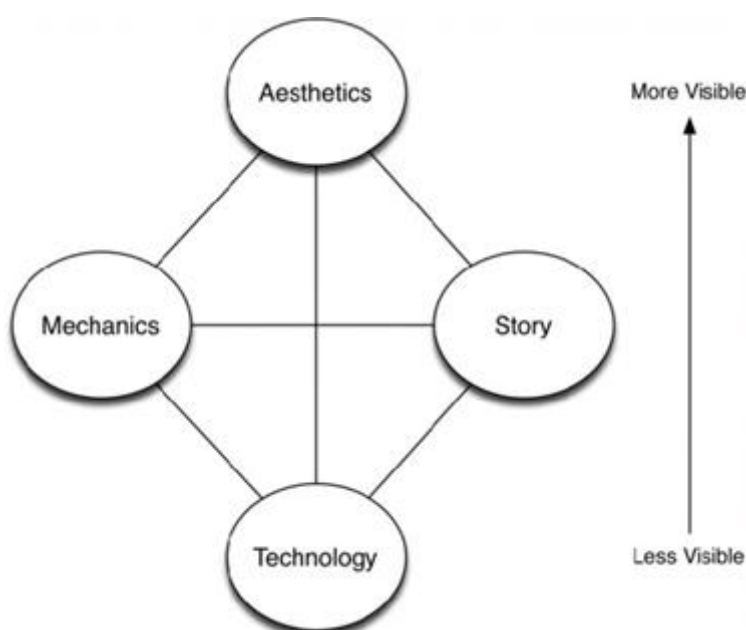


Figure 14. Schell's Elemental Tetrad

The **6D Framework**, presented by Werbach and Hunter in 2012, depicts a six-step process for developing a gamified system (Table 10), namely: (1) definition of the business objectives for the gamification project under development; (2) description of the targeted player behaviours that the process aims to encourage or

discourage; (3) description of the intended players; (4) development of appropriate short-term engagement loops and long-term progression loops; (5) integrating fun elements; and (6) deploying the appropriate tools.

Table 10. The 6D Framework

Structural Elements of 6D Framework	Applicability
Objectives, target behaviors, player's profile, activity loops, fun, deployment	Sequential iterative gamification design model based on thinking

8.3.4 Gamification Mechanisms & TITAN

The integration of gamification techniques in online applications and intelligent chatbots, such as TITAN, can have several benefits, including: **(a) Enhanced Engagement:** Gamification can enhance the interaction with the user and make it enjoyable and immersive. Gamification techniques can capture users' attention and sustaining interest for a period of time; **(b) Enhanced Motivation:** Gamification elements like rewards, challenges, and competition among users, linked with intrinsic and extrinsic motivation, can trigger higher levels of participation and engagement in the users' interaction with an intelligent chatbot like TITAN. **(c) Progress Monitoring:** Gamified intelligent chatbots can enable users to keep track of their own progress, and receive feedback about their advancement over time, acting as a positive reinforcement, that enhances the sense of user accomplishment. Taking into consideration the gamification mechanisms, presented earlier, we have tried to adapt them and explore how they could potentially be used in the context of TITAN (Table 11).

Table 11. Gamification mechanics and their potential application of TITAN

Mechanism	Description	Motive	Potential Application for TITAN
Fast feedback	Immediate feedback or response to user actions	Mastery, Progress	When the user performs a certain task, he/she will be able to view their updated points immediately via
Transparency	Being able to see where other users stand, quickly and easily	Progress, Social Interaction	The user will be able to view his/her individual points/scoring. The cumulative points will be viewed via screens.
Goals	Short- and long-term goals to achieve	Purpose, Progress, Social Interaction	Each round/level of critical thinking can be linked with a "maximum score that can be achieved" (or "maximum score that has been achieved (by other users)"). This will act as a goal for the user to reach a particular short or long-term score/point etc.
Badges	Display evidence of the accomplishments of the user.	Mastery, Purpose, Progress, Social Interaction	Points can be translated to intangible achievements (badges, etc), showing user evolution.
Levelling up	Achieving a "status" within a community/ecosystem	Mastery, Purpose, Progress, Social Interaction	Each critical thinking level has sub-levels. Once the user completes a level and/or sub-level, he/she is moved to the next level.
Competition	Being able to see where I stand against other users	Mastery, Social Interaction	The user can compete with: - his/her progress the previous day or last time he/she used TITAN. - other users in the same level

			other active users in the system on this particular point in time that the user is active
Collaboration	Collaborate with others to complete a task	Purpose, Social Interaction	The user could: - collaborate directly with other users (i.e., ask a friend or another TITAN user) - a collaborative score (or bonus points) could be offered to the users that “help” other users (this could be used for leveling up, among others)
Community	Being able to see the community and how they progress, while also the community can see me (anonymously as part of the TITAN community)	Social Interaction	The user can: collaborate indirectly (i.e., see suggestions that other users shared for the same article for example) with the TITAN community
Points	Being able to see measurable evidence of the individual accomplishments	Progress, Social interaction	The user can: acquire points (individual) for completing some specific levels, tasks, etc.

Aligned with the MDA (Mechanics - Dynamics - Aesthetics) model the three components that could be taken into consideration are presented in Figure 15.

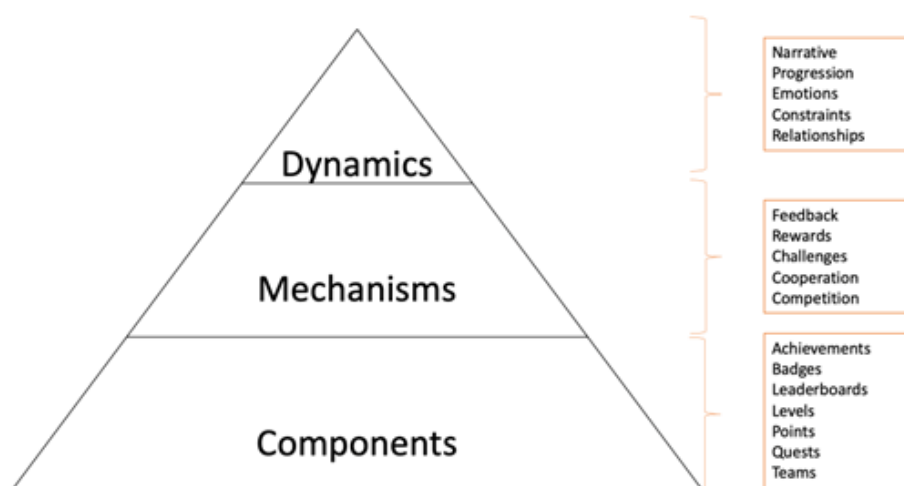


Figure 15. The MDA model in the context of TITAN

The integration of game elements and mechanics in the context of conversational agents is widely used to enhance user-engagement, create an enjoyable and interactive user experience and encourage engagement. When building a chatbot with gamification elements some aspects/challenges beyond the game-specific should be examined (Hidayatulloh et al., 2021), and which will be taken into consideration in the case of TITAN. These aspects are:

1. the reliability of the architecture – the choice of the architecture is dependent on the kind of system to be developed (i.e., Question Answer-based chatbot), and the expected usefulness of the chatbot for the user-taking into account that the user may leave the chat in the middle of the conversation and then return or not, thus not keeping track of the conversation impacts the chatbot’s usefulness. Adding gamification elements

may pose additional architectural design challenges, such as keeping the server working in real-time while maintaining the interactive gamified chatbot (which uses NLP (natural language processing) techniques that required a high level of computational resources) (Hidayatulloh et al., 2021).

2. user acceptance and motivation - the relationship that is developed between a human and a chatbot, exhibits similarities with the development of intrapersonal relationships (Skjuve et al., 2021), where trust plays a key role. User motivation plays an important role, and therefore maintaining it is important. Research in the area of learning chatbots among students indicates that emotional empathy (compassion for feelings) seems to boost the desire to learn, and the “mood” of the chatbot may impact on the users’ desire to leave the chatbot’s room (Pérez-Marín and Pascual-Nieto, 2013).

3. data security and privacy concerns – since chatbots do not rely only on the rule-based model but also collect personal conversation data (Hidayatulloh et al., 2021), increasing the risk of vulnerabilities and security threats in the future (Hasal et al., 2021). In addition, the leaderboard feature may give rise to privacy issues, which may “expose” users, making them unwilling to chat with the chatbot (Hidayatulloh et al., 2021). Additional security threats may not only exist on the server side but also on the client side (i.e., four potential attacks modules: the client module, communication module, response generation module (e.g., adversarial attacks, language model attacks and feedback engineering attacks), and the database module (Ye and Li, 2020)).

4. Gamification feature challenges – the specific gamification features (badges, scoreboards, etc) should be selected before developing a chatbot (Hidayatulloh et al., 2021) which may be a challenge. In the context of conversation-based learning environments not all game mechanics can be smoothly applied according to Hidayatulloh et al., (2021), as in the case of “points” where the challenge is how to score the activity of learners, this may be challenging for chat-bot learning environments as re-scoring for answering the same set of questions or interaction may occur as well as ensuring ‘empathetic’ behavior with nearly correct answers, mistakes, typos, etc.

The creation of a relevant incentives framework for the TITAN ecosystem is important for always fostering engaging interactions with the system users. Our examination of the various motivational theories and game-based motivational factors enables us to identify the incentive aspects that will ensure, the sustainability of the TITAN ecosystem anchored in Responsible Research and Innovation (RRI) principles (ethically acceptable and socially desirable to benefit the society at large). An overview of the different incentives, the relevancy for TITAN and the phase of the TITAN system is presented in Table 12, below.

Table 12. Overview of Incentive Frameworks and relevancy for the TITAN system

Types of Incentives	Relevancy for TITAN	TITAN System Phases	
		Initial phases	Usage/adoption phases
Intrinsic incentives (Internal motives)	Intrinsic incentives can have an important motivational role in TITAN	Intrinsic incentives could trigger the self-desire to engage in curiosity-driven, and personal-development motivation during the consideration phase.	When the user has adopted the service to trigger curiosity (for finding the true facts), learning (how to advance their critical thinking, advance their skillset, etc.), belonging (in the TITAN ecosystem and community), and enjoyment (from personal achievement and discovery).
Extrinsic incentives (external motives)	Extrinsic incentives can have an important motivational role in TITAN, especially non-monetary incentives	Extrinsic incentives would provide limited value in the initial phases of TITAN service consideration.	Extrinsic, non-monetary incentives would be value-adding during the TITAN service usage. Recognition could act as a positive incentive that can be used to motivate TITAN user and reward them with some sort of

			praise to the broader community, (i.e., user of the day, and long TITAN service award, etc.).
Gamification Elements	High impact for TITAN fostering user mastery, progress, achievement, etc.	Limited applicability – potentially for encouraging registration to the TITAN system (i.e., short quiz type of question could be posed to the user, etc.)	When the user performs a certain task, he/she will be able to experience gamification elements

Given the diversity and gamification components it is important to examine them distinctively in the context of TITAN and explore their potential implementation. Table 13 below aims to provide an overview and set an implementation horizon as well.

Table 13. Gamification mechanics, their potential application and implementation for TITAN

Mechanism	Description	Motive	Potential Application for TITAN	Implementation Aspects for TITAN
Fast feedback	Immediate feedback or response to user actions	Mastery, Progress	When the user performs a certain task, he/she will be able to view their updated points immediately	Could be implemented- Feedback to users that they successfully completed a task, activity etc
Progress Transparency	Being able to see where he/she stands as well as where other users stand, quickly and easily	Progress, Social Interaction	The user will be able to view his/her individual points/scoring. The cumulative points will be viewed via screens.	Could be implemented, so that the user can: be able to see his/her own profile information (user status, progress report statistics, user skills achievement report Receive feedback associated with his/her progress – for example "Congrats you are among the 5% of the top users of TITAN"
Goals	Short- and long-term goals to achieve	Purpose, Progress, Social Interaction	Each round/level of critical thinking can be linked with a "maximum score that can be achieved" (or "maximum score that has been achieved (by other users)"). This will act as a goal for the user to reach a particular short or long-term score/point etc.	Could be implemented: Short-term goals: for example: assess your critical level, understand the basic set of disinformation signals, debunk (to be able to debunk disinformation), successful completion of specific skill-set Long-term goals: for example enhance my critical thinking level, etc
Badges	Display evidence of the	Mastery, Purpose, Progress,	Points can be translated to intangible achievements	Could be implemented: Badges:

	accomplishments of the user.	Social Interaction	(badges, etc), showing user evolution.	Short-term/long-terms goals (i.e., skills oriented) Functional ones (i.e., successful completion of registration in the system)
Levelling up	Achieving a “status” within a community/ecosystem	Mastery, Purpose, Progress, Social Interaction	Each critical thinking level has sub-levels. Once the user completes a level and/or sub-level, he/she is moved to the next level.	This will could be considered for implementation (as a future aspect)
Competition (Project levels - ranks)	Being able to see where I stand against other users	Mastery, Social Interaction	The user can compete with: • his/her progress the previous day or last time he/she used TITAN. • other users in the same level • other active users in the system on this particular point in time that the user is active	This is beyond the scope of TITAN and thus will not be considered
Collaboration	Collaborate with others to complete a task	Purpose, Social Interaction	The user could: • collaborate directly with other users (i.e., ask a friend or another TITAN user) • a collaborative score (or bonus points) could be offered to the users that “help” other users (this could be used for leveling up, among others)	This is beyond the scope of TITAN and thus will not be considered
Community	Being able to see the community and how they progress, while also the community can see me (anonymously as part of the TITAN community)	Social Interaction	The user can: • collaborate indirectly (i.e., see suggestions that other users shared for the same article for example) with the TITAN community	This is beyond the scope of TITAN and thus will not be considered Not to be considered
Points	Being able to see measurable evidence of the individual accomplishments	Progress, Social interaction	The user can: • acquire points (individual) for completing some specific levels, tasks, etc.	This will be considered at a future level.

9 CONCLUSION

This deliverable outlines a comprehensive framework for combating disinformation, focusing on the integration of critical thinking, media literacy, and advanced technological tools. The methodologies presented are critical for the building of the TITAN complete system, these have been discussed and designed during a long time period. All of them underscore the importance of fostering citizens' abilities to critically assess information, enhancing their media literacy to navigate today's complex information environments, that is the final goal of TITAN project.

In conclusion, we can state as the application of the Socratic method within TITAN provides a classical yet innovative approach to logical assessment, encouraging deeper inquiry and critical reflection. By prompting users to examine the reliability and bias of information, this method serves as a foundational tool for fostering nuanced understanding in disinformation detection.

In the same time a fundamental cornerstone of the TITAN methodology is the Critical Thinking Assessment Methodology. TITAN's Critical Thinking Assessment (CTA) model offers a personalized approach to evaluating and enhancing cognitive skills that are crucial in identifying and countering disinformation. By focusing on both cognitive abilities and critical thinking dispositions, the project tailors interventions to individual users, reinforcing their capacity for reflective, analytical thinking.

The integration of media literacy with critical thinking methodologies allows TITAN to address both cognitive and practical challenges of misinformation. Through structured micro-lessons and dialogue-based approaches, the project equips users with the tools to critically assess media content, thus enhancing their ability to recognize manipulation and bias.

TITAN's use of state-of-the-art technologies, including few-shot and zero-shot learning, adds significant value in the detection of evolving disinformation patterns. By continuously refining these models, the system remains adaptable to the fast-paced digital environment, ensuring its relevance and effectiveness. In addition the gamification strategy employed by TITAN plays a vital role in maintaining user engagement. By providing intrinsic and extrinsic motivators, the project fosters a culture of continuous learning and active participation, encouraging users to become critical consumers of information in a sustainable ecosystem.

The results provided within this deliverable will be taken into considerations mainly for the implementation of the project, but they are integrated with the Living Labs feed-backs and results presented in an additional deliverable and as TITAN methodologies continue to evolve, the next steps is the full implementation of pilot programs across diverse settings. These pilots will test the system's effectiveness in real-world scenarios, providing essential feedback on user engagement, the adaptability of machine learning models, and the effectiveness of the critical thinking and media literacy interventions. Based on the pilot results, iterative refinements should be made to ensure the tools and methodologies are practical, scalable, and user-friendly.

Sustaining user engagement beyond the initial introduction of TITAN will be critical for long-term success. Building a community of active participants who regularly interact with TITAN's tools and share best practices in critical thinking could help create a culture of informed, skeptical media consumption. This will be conducted also in the exploitation work package of the TITAN project, this may include fostering partnerships with community organizations, implementing reward systems for active users, or hosting interactive workshops and forums. In matter of facts to maximize impact, after the definition and the complete delivery of the TITAN methodology we can consider how important can be to explore new partnerships with educational institutions, media platforms, and civic organizations. By integrating the critical thinking and media literacy tools into

broader educational curricula and public awareness campaigns, TITAN can help build a more informed, critically aware population on a larger scale.

Finally, the project's broader success will depend on collaboration with policymakers and regulators, for this reason the results achieved in the WP2 of TITAN and described in the present deliverable have been discussed and shortly presented in different policies and interdisciplinary events in the hope that by actively engaging with governments and regulatory bodies, TITAN methodology can contribute to the development of policies that address disinformation at the systemic level. Our final goal is to ensure a safer and more transparent digital information ecosystem, in this direction we believe sharing TITAN's tools and best practices methodologies in critical thinking could help create a culture of informed, skeptical media consumption.

10 REFERENCES

- Alshmemri, M., Shahwan-Akl, L., & Maude, P. (2017). Herzberg's two-factor theory. *Life Science Journal*, 14(5), 12-16.
- Bandura, A. (1982). Self-efficacy mechanism in human agency. *American psychologist*, 37(2), 122.
- Bandura, A. (1991, January). Self-regulation of motivation through anticipatory and self-reactive mechanisms. In *Perspectives on motivation: Nebraska symposium on motivation* (Vol. 38, No. 1, pp. 69-164).
- Blakeslee, S. (2004). The CRAAP test. *Loex Quarterly*, 31(3), 4.
- Bowell T. & Kemp G., *Critical Thinking A Concise Guide*, London, Routledge, 2014, p.1
- Bowell, T., & Kemp, G., *Critical Thinking: A Concise Guide* (4th ed.), London, Routledge, 2014, <https://doi.org/10.4324/9781315772035>
- Caulfield M., 2019, SIFT (The Four Moves) - Hapgood, <https://hapgood.us/2019/06/19/sift-the-four-moves/>
- Christopoulos, A., & Mystakidis, S. (2023). Gamification in Education. *Encyclopedia*, 3(4), 1223-1243.
- Computer Age, 2009 (5)
- Constitutional Rights Foundation, 2017, Understanding 'Fake News', <https://www.crf-usa.org/images/pdf/UnderstandingFakeNews.pdf> & https://www.crf-usa.org/images/pdf/fact_finding.pdf
- Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 2425-2428).
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). Bert: Pre-training of deep bidirectional Transformers for language understanding. *arXiv.org*. Retrieved September 10, 2022, from <https://doi.org/10.48550/arXiv.1810.04805>.
- Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Journal of educational technology & society*, 18(3), 75-88.
- Dueben, P. D., Schultz, M. G., Chantry, M., Gagne, D. J., Hall, D. M., & McGovern, A. (2022). Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook. *Artificial Intelligence for the Earth Systems*, 1(3), e210002. doi: <https://doi.org/10.1175/AIES-D-21-0002.1>.
- Elder, L., & Paul, R. (2020). *Fact Over Fake: A Critical Thinker's Guide to Media Bias and Political Propaganda*. Foundation for Critical Thinking.
- Elder, Linda & Paul, Richard, *Fact over Fake: A Critical Thinker's Guide to Media Bias and Political Propaganda*, The Foundation for Critical Thinking, 2020
- Facione, P. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report).
- Facione, P. A. (2011). Critical thinking: What it is and why it counts. *Insight assessment*, 1(1), 1-23.
- Flaxman S., Goel S., & Rao J.M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80, 298-320.
- Geiger, M. A., & Cooper, E. A. (1996). Using Expectancy Theory to Assess Student Motivation. *Issues in Accounting Education*, 11(1).
- Gopalan, V., Bakar, J. A. A., Zulkifli, A. N., Alwi, A., & Mat, R. C. (2017, October). A review of the motivation theories in learning. In *Aip conference proceedings* (Vol. 1891, No. 1). AIP Publishing.
- Gurzawska, A., Mäkinen, M., & Brey, P. (2017). Implementation of Responsible Research and Innovation (RRI) practices in industry: Providing the right incentives. *Sustainability*, 9(10), 1759.
- Hakulinen L., T. Auvinen, and A. Korhonen, (2013). "Empirical Study on the Effect of Achievement Badges in TRAKLA2 Online Learning Environment", In *Proceedings of Learning and Teaching in Computing and Engineering (LaTICE) conference*, March 21-24, 2013, Macau, pp. 47-54.

- Hamari, J., & Koivisto, J. (2015). Why do people use gamification services?. *International journal of information management*, 35(4), 419-431.
- Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? - A literature review of empirical studies on gamification. In *System Sciences (HICSS)*, 2014 47th Hawaii international conference on (pp. 3025e3034). IEEE.
- Harjani, T., Roozenbeek, J., Biddlestone, M., van der Linden, S., Stuart, A., Iwahara, M., Piri, B., Xu, R., Goldberg, B., & Graham, M. (2022). A Practical Guide to Prebunking Misinformation.
- Haruyama, M., Hidaka, K. (2023). What influences users to provide explicit feedback? A case of food delivery recommenders. *User Model User-Adap Inter.* <https://doi.org/10.1007/s11257-023-09385-8>.
- He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2111.09543>.
- He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., & Jiang, S. (2020). A Survey on Recent Advances in Sequence Labeling from Deep Learning Models. *ArXiv*, abs/2011.06727.
- Herzberg, F. I. (1966). *Work and the Nature of Man*. Herzberg, F. I. (1982). *The managerial choice: To be efficient and to be human* (2nd ed., Rev.). Salt Lake City, UT: Olympus.
- Herzberg, F., Mausner, B., & Snydermann B. (1959). *The motivation to work*. New York: Wiley.
- Hidayatulloh, I., Pambudi, S., Surjono, H. D., & Sukardiyono, T. (2021). Gamification on chatbot-based learning media: a review and challenges. *Elinvo (Electronics, Informatics, and Vocational Education)*, 6(1), 71-80.
- Hotjar (2023). User Feedback: What It Is and How to Collect It. <https://www.hotjar.com/user-feedback/>.
- Howard, M. C., & Hoffman, M. E. (2018). Variable-centered, person-centered, and person-specific approaches: Where theory meets the method. *Organizational Research Methods*, 21(4), 846-876.
- Ibrahim, I. W., Ghazali, I. M., Abdullah, S. A. S., Rahmat, N. H., Hamid, M. H. A., & Azhari, S. H. A. (2023). Exploring Motivation for Learning Using Hertzberg's Two Factor Theory. *International Journal of Academic Research in Business and Social Sciences*, 13(3), 1065 – 1083.
- Ilic, P. "Micro-lessons as a Response to Emergency Remote Teaching". 2022 IEEE Learning with MOOCS (LWMOOCS), 2022 (<https://ieeexplore.ieee.org/abstract/document/9927755/authors#authors>)
- Jasser, J., Garibay, I., Scheinert, S., & Mantzaris, A. V. (2022). Controversial information spreads faster and further than non-controversial information in Reddit. *Journal of Computational Social Science*, 5(1), 111-122.
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux
- Katt, J. A., & Condly, S. J. (2009). A Preliminary Study of Classroom Motivators and Demotivators from a Motivation-hygiene Perspective. *Communication Education*. 58 (2). 213 - 234.
- Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *CoRR*, abs/2112.01716. <https://arxiv.org/abs/2112.01716>.
- Kourieos, S. "Video-mediated microteaching-A stimulus for reflection and teacher growth", *Australian Journal of Teacher Education*, vol. 41, no. 1, pp. 4, 2016.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151-160.
- Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., & Kustarev, A. (2012, October). Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2335-2338).
- Labrak, Y., Rouvier, M., & Dufour, R. (2023). A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. *ArXiv*, abs/2307.12114.
- Lazer D. M. J., Baum M. A., Benkler Y., Berinsky A. J., Greenhill K. M., & Menczer F (2016). The science of fake news. *Science*, 359(6380), 1094-1096.
- Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E. Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., Vraga, E. K., Wood, T. J., Zaragoza, M. S. (2020). *The Debunking Handbook 2020*. Available at <https://sks.to/db2020>. DOI:10.17910/b7.1182

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/1907.11692>.
- Logan, F. A. (1968). Incentive theory and changes in reward. In *Psychology of Learning and Motivation* (Vol. 2, pp. 1-30). Academic Press.
- Lokman, A., Hassan, F., Ustadi, Y. A., Rahman, F. A. A., Zain, Z. M., & Rahmat, N. H. (2022). Investigating motivation for learning via Vroom's Theory. *International Journal of Academic Research in Business and Social Sciences*, 12(1), 504-530.
- Machete, P., & Turpin, M. (2020). The use of critical thinking to identify fake news: A systematic literature review. In *Responsible Design, Implementation and Use of Information and Communication Technology: 19th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2020, Skukuza, South Africa, April 6–8, 2020, Proceedings, Part II 19* (pp. 235-246). Springer International Publishing.
- Magnusson, D., & Stattin, H. (2007). The person in context: A holistic-interactionist approach. *Handbook of child psychology*, 1.
- Maslow, A. (1943). A theory of human motivation. *Psychological Review*, 50(4); pp. 370–96. Maslow, A. (1954). *Motivation and personality*. New York, NY: Harper.
- McLeod, S. (2007). Maslow's hierarchy of needs. *Simply psychology*, 1(1-18).
- Mittelman, W. (1991). Maslow's study of self-actualization: A reinterpretation. *Journal of Humanistic Psychology*, 31(1); pp. 114–135
- Motevalli, S., Perveen, A., & Michael, M. T. A. (2020). Motivating students to learn: An overview of literature in educational psychology. *International Journal of Academic Research in Progressive Education and Development*, 9(3), 63-74.
- Nosich, G. M. (2012). *Learning to think things through: A guide to critical thinking across the curriculum*.
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32, 303–330.
- Overholser, J. C., & Beale, E. (2023). The art and science behind socratic questioning and guided discovery: A research review. *Psychotherapy Research*, 33(7), 946-956.
- Parnami, A., & Lee, M. (2022). Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. ArXiv, abs/2203.04291.
- Payan-Carreira, R., Sacau-Fontenla, A., Rebelo, H., Sebastião, L., & Pnevmatikos, D. (2022). Development and Validation of a Critical Thinking Assessment-Scale Short Form. *Education Sciences*, 12(12), 938.
- Petrovic, S., Osborne, M., & Lavrenko, V. (2011). Rt to win! predicting message propagation in twitter. In *Proceedings of the international AAAI conference on web and social media* (Vol. 5, No. 1, pp. 586-589).
- Puig, B., Blanco-Anaya, P., & Pérez-Maceira, J. J. (2021, May). "Fake news" or real science? Critical thinking to assess information on COVID-19. In *Frontiers in Education* (Vol. 6, p. 646909). Frontiers Media SA.
- Qing, X., Jianghong, He. "The exploration of teaching mode under the network environment [J]".
- Ribeiro, L.A. da Silva T.L., and Mussi A.Q. (2018) Gamification: a methodology to motivate engagement and participation in a higher education environment, *Inter. J. of Educ. & Res.* 6 (4) (2018) 249-264.
- Rowan University, 2023, Debunking Misinformation, <https://libguides.rowan.edu/debunking>
- Ruthankoon, R., & Ogunlana, S. O. (2003). Testing Herzberg's two-factor theory in the Thai construction industry. *Engineering, Construction and Architectural Management*, 10(5), 333–341.
- Sahu, H., Sharma, N., Gupta, U. (2019). A New Framework for Collecting Implicit User Feedback for Movie and Video Recommender System. In: Khare, A., Tiwary, U., Sethi, I., Singh, N. (eds) *Recent Trends in Communication, Computing, and Electronics. Lecture Notes in Electrical Engineering*, vol 524. Springer, Singapore.
- Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, 32(1), 77-112.
- Schell J. (2008). *The Art of Game Design: A Book of Lenses*. Burlington, USA: Morgan Kaufmann Publishers.
- Sinnott-Armstrong, Walter, *Understanding arguments: an introduction to informal logic.*, Australia: Cengage Learning, Edt. by Robert J. Fogelin, 2015

- Spector, J. M., & Ma, S. (2019). Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence. *Smart Learning Environments*, 6(1), 1-11.
- Steinkuehler, C.A.; Williams, D. Where Everybody Knows Your (Screen) Name: Online Games as “Third Places”. *J. Comput.-Mediat. Commun.* 2006, 11, 885–909.
- Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In 2010 IEEE second international conference on social computing (pp. 177-184). IEEE.
- Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, 11(1), 99-113.
- Tsay, C.H.; Kofinas, A.K.; Trivedi, S.K.; Yang, Y. Overcoming the novelty effect in online gamified learning systems: An empirical evaluation of student engagement and performance. *J. Comput. Assist. Learn.* 2020, 36, 128–146.
- VALAMIS. Personalized Learning: <https://www.valamis.com/hub/personalized-learning#what-is-personalized-learning>
- Van Eemeren, F. H., & Henkemans, A. F. S. (2016). *Argumentation: Analysis and evaluation*. Taylor & Francis.
- van Laar, E. (2019). What are Essential skills?: A multimethod approach to 21st-century digital skills within the creative industries.
- Visser, J., and Lawrence, J. (2022) "The skeptic web service: Utilising argument technologies for reason-checking". In *Proceedings of the Ninth International Conference on Computational Models of Argument (COMMA 2022)* pp 375-376, IOS Press, 2022.
- Vlachos A. 2006. Active Annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.
- Von Schomberg, R. A vision of responsible research and innovation. In *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*; John Wiley & Sons: Hoboken, NJ, USA, 2013; pp. 51–74.
- Walk, W., Görlich, D., & Barrett, M. (2017). Design, dynamics, experience (DDE): an advancement of the MDA framework for game design. *Game dynamics: Best practices in procedural and dynamic game content generation*, 27-45.
- Wang, Y. F., Hsu, Y. F., & Fang, K. (2022). The key elements of gamification in corporate training–The Delphi method. *Entertainment Computing*, 40, 100463.
- Wu, J. H., Tennyson, R. D., & Hsia, T. L. (2010). A study of student satisfaction in a blended elearning system environment. *Computers and Education*, 55(1), 155-164.
- Xu, F., Buhalis, D., & Weber, J. (2017). Serious games and the gamification of tourism. *Tourism Management*, 60, p. 244–256. <https://doi.org/10.1016/j.tourman.2016.11.020>
- Yinheng Li. 2023. A Practical Survey on Zero-Shot Prompt Design for In-Context Learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 641–647, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Zaman, T., Fox, E. B., & Bradlow, E. T. (2014). A Bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, 8(3). [Original source: <https://studycrumb.com/alphabetizer>]

ANNEXES

In the following some annexes are included, in order to demonstrate better how all methodological considerations have been adapted and integrated in the technical implementation of the TITAN system, these are:

ANNEX 1: Types of information disorder

ANNEX 2: Tactics of disinformation

ANNEX 3: Item-focused dialogues

ANNEX 4: Media Literacy Protocol

ANNEX 5: Dialogues examples for CT dimensions

ANNEX 6: Dialogues examples for Self Regulation dimension

ANNEX 1: TYPES OF INFORMATION DISORDER

Table 14. Types of Information Disorder

Type of Mis/Disinformation	Description (Wardle, 2019)	Examples (USAID, 2021)
Fabricated content	New content that is 100% false, made to deceive and do harm	Indian politician uses deepfake to show himself giving a speech in a different language. https://www.vice.com/en_in/article/igedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp
False connection	When headlines, visuals or captions don't support the content	The title of the video plays on anti-Chinese sentiment that is prevalent in Latin America to get people to click on the video and share. A textbook example of False Connection. https://www.youtube.com/watch?v=QillhGzhW7o
False content	When genuine content is shared with false contextual information	Video allegedly from Wuhan province where coronavirus originated is really from Indonesia. https://twitter.com/AFPFactCheck/status/1221732075885101061
Imposter content	When genuine sources are impersonated	Fake screenshot that shows Newsweek article about Iran air strikes. Shows side-by-side comparisons of real/fake screens. https://twitter.com/elisethoma5/status/1215604086780743680?s=20
Satire or parody	No intention to cause harm but has potential to fool.	The "news" about an Australian couple on cruise ship ordering wine via drone was debunked. The couple admitted that it was a joke post on Facebook for their friends. https://factcheck.afp.com/australian-couple-quarantined-onboard-diamond-princess-cruise-reveal-wine-drone-delivery-story-was

Manipulated content	When genuine information or imagery is manipulated to deceive	Daily Mail edited a photo to make two people in a garden to appear closer than they really are. https://twitter.com/jamescracknell/status/1254395457033379843?s=21
Misleading content	Misleading use of information to frame an issue or individual	BBC Trending investigates cases of disinformation on Australian bushfires maps on social media. https://www.bbc.com/news/blogs-trending-51020564

ANNEX 2: TACTICS OF DISINFORMATION

Table 15. Tactics of Disinformation

Technique/Tactic	Description	Examples (e.g. Harjani et al., 2022)
Discredit	The aim is to undermine trust in reliable sources through ad hominem attacks, spreading rumors about individuals, or discrediting institutions	“Barbara has an uncontrollable temper and apparently a personality disorder too! We can’t have someone crazy in power.” This example attacks characteristics of the leader, instead of discussing their policies or leadership decisions.
Emotional Manipulation	Manipulating emotions like fear, anger, or sympathy through sensationalism, emotive language, and evocative imagery to influence perceptions and actions.	“What this airline did for its passengers will make you tear up — SO heartwarming.” This example shows how information can be presented to deliberately spark an emotional reaction to promote clicking and sharing and reduce critical evaluation.
Polarization	Dividing public opinion and creating social conflict by amplifying existing societal divides, promoting extreme views, and demonizing opposing sides.	“People’s Party: Don’t believe the Worker Party liars. They said they would abolish student debt yet more people today are in debt than ever.” This example uses hostile “othering” language by describing another party as liars.
Impersonation, Fake or Misleading Personas	Mimicking the appearance of a real person or organization by creating fake profiles, websites, or news outlets that look legitimate. Networks of fake personas and websites.	“NASA admitted that climate change occurs naturally as a result of changes in Earth’s solar orbit and not anthropogenic factors.” This example uses NASA as a way to increase the credibility of the statement, even though NASA has never made such a claim.

Digital Manipulation Tactics	Coordinated Inauthentic Behavior, Sock Puppets, Troll Farms, Manufactured Amplification, Computational Propaganda, Clickbait, Content Farm, Flooding.	Coordinated Inauthentic Behavior was exemplified by the 2016 Russian interference in the U.S. election, where fake accounts were used to spread misleading information and amplify divisive content, creating a false sense of widespread support or opposition.
Conspiracy Theory	Creating complex false narratives that attribute events to covert, malevolent actions by powerful groups, with no evidence available	“Vaccines are just a way for billionaires to track us with their microchip vaccines! Who's really in control of our bodies here?” This example encourages conspiratorial ideation by claiming people are not in control, referring to a mysterious group who is, in this case billionaires, and making unsubstantiated claims,
Trolling	Disrupting meaningful conversations and provoking anger through inflammatory comments, fake personas, and coordinated attacks	Posting inflammatory comments on a forum to derail a productive discussion or deepen societal divides.
Pseudoscience	Using the appearance of scientific credibility to promote ideas or products without empirical support, often to advance specific agendas or discredit legitimate findings.	Promoting a miracle cure for diseases without any scientific evidence
Science Denial	Disregarding or denying established scientific evidence and consensus to create doubt and support ideological, political, or commercial agendas.	Denying climate change despite overwhelming scientific consensus
False Dichotomy/Dilemma	Presenting two opposing options as the only possible choices, ignoring or excluding other viable alternatives to oversimplify complex issues.	“Either you support the energy protests or you don’t believe in justice.” This example positions two ideas as opposite sides of a spectrum — making “supporting energy protests” and “believing in justice” as opposites — when it is possible to support both or neither at the same time, as

		well as many other positions someone may take.
Manipulation of Evidence and Sources e.g. Cherry Picking	Selectively presenting only the evidence that supports a particular viewpoint while ignoring or dismissing contradictory evidence to distort the overall picture.	Highlighting a single study that supports a harmful health claim while ignoring the bulk of scientific research that contradicts it.

ANNEX 3: ITEM-FOCUSED DIALOGUES

THEORETICAL BACKGROUND OF ITEM-FOCUSED DIALOGUES

What is a Socratic dialogue?

Dialogue is a conversation between two or more characters. The purpose of a dialogue is to exchange information, share ideas, solve a problem, discuss or debate and reconcile differing viewpoints to reach an agreement/mutual understanding. TITAN's dialogues differ from other forms of conversation in purpose, method, and tone. In TITAN we developed a dialogical format based on modern adaptations of the Socratic method of questioning. In particular, the modern Socratic pedagogy in critical thinking involves probing questions for reflective examination of ideas and information. Participants learn to rigorously analyze and question assumptions, evaluate evidence and reasoning, explore alternative perspectives, and develop well-supported conclusions through a structured, interactive dialogue. In effect, participants learn to think independently, reason logically, and articulate their ideas clearly, ultimately fostering a deeper understanding of complex issues. TITAN's Socratic dialogues adopt the Q/A format to enhance users' critical thinking against disinformation and bad journalism. In TITAN's dialogues, the chatbot is the critical thinking coach (questioner) who guides the users (the coachees or respondents) to critically examine media items and themselves (their beliefs, biases and media use habits).

Purpose of TITAN's dialogues

We have designed three different forms of Socratic dialogues. In the Item-Focused dialogues, the AI chatbot is a critical thinking coach who guides users with the right critical thinking questions to perform the following tasks: (a) examine the trustworthiness of the author/source of a media item and (b) recognize specific disinformation tactics in the content of the item. In effect, the users reflect on the chatbot's questions to perform these tasks using media-specific critical thinking (see #1 below). In contrast, the Person-Focused dialogues are designed to help the users examine and become aware of their own psychological barriers (e.g. media use habits and personal bias) to critical thinking, thus becoming vulnerable to dis/misinformation (see #1 below; Metacognition). Finally, in the Logic-based dialogues the chatbot guides users with open critical thinking questions to evaluate the reasoning and rhetoric found in a media item without focusing on the detection of a specific disinformation tactic (see #1 below; e.g. Analytic thinking & logic).

Item-Focused dialogues

#1. We use Media Critical Thinking (MCT)

MCT is a blend of generic critical thinking (CT) and Media Literacy. Generic CT includes skills in interpretation, argument analysis, inference, identifying assumptions, evaluation of evidence and metacognition. These are the core skills required to (1) recognize, analyze and evaluate media messages, evidence and arguments, in order (2) to detect the most common logical fallacies and rhetorical ploys used in dis/misinformation. However, the critical media consumer also different set of critical thinking skills required (3) to judge the reliability of experts, sources, and evidence, (4) to analyze narrative framing and disinformation narratives, (5) to detect deceptive persuasion tactics, including digital manipulation tactics, and (6) to recognize media bias, agendas, and political economic influences.

#2: We keep the dialogues Socratic

In the Item-focused dialogues we created a mix of open-ended and closed questions addressing the desiderata of Socratic critical thinking. The open-ended questions are Socratic in that they encourage reflection and self-discovery. A number of closed questions are tailored for media critical thinking (use media literacy protocols) prompting users to check specific, verifiable information in a media item. However, the closed questions remain Socratic: they retain the substance of Socratic questions to spark critical thinking on specific aspects of media information and show how to refute fallacious arguments, rhetorical tricks and disinformation narratives (Table 16).

Table 16. MCT aligns with Socratic questions and Facione's CT skills

Socratic questions & Facione's CT skills	MCT skills
Clarification / INTERPRETATION	Interpreting/making sense of media messages by categorizing and comprehending the meaning of information, concepts and ideas within contexts.
Probing Assumptions / ANALYSIS	Breaking down media information into its component parts, such as reasons, claims, arguments, data, and assumptions.
Probing Rationale, Reasons, Evidence / EVALUATION	Assessing the credibility of statements and the quality of reasoning/arguments, determining if claims are supported and arguments are logically valid.
Questioning Viewpoints and Perspectives / INFERENCE	Drawing reasonable conclusions based on evidence, recognizing what follows logically from the available data.
Questioning the Question / SELF-REGULATION	Reflecting on and adjusting one's thinking and biases when evaluating media content.
Media-specific CT skills and dispositions	Identify deceptive persuasion tactics (mainly rhetorical tactics)
	Evaluate the credentials of authors and outlets
	Understand media agendas/media bias, political and economic influences
	Identify digital manipulation tactics
	Constructive Skepticism

#3: We keep the dialogues coherent and short

Open Socratic questions invite a wide range of user responses, which can be challenging for a chatbot to follow up accurately, especially because multiple conversational paths are possible. In particular, open questions pose highly challenging demands to chatbots to quickly come up with a wide range of responses, understand the conversation's context, make sense of complex language, and handle messy or varied input. Hence, open questions may lead to errors and confusion, miscommunication, loss of coherence, difficulty in interpretation, and response delays. These factors may result in user frustration, reduced trust, decreased engagement, misguided responses, lack of depth, increased user effort and user impatience. In contrast, we opted for closed questions because they promote more structured, coherent, quick and contextually relevant dialogues. Our closed question provide specific instructions that prompt for clear, verifiable answers, typically inviting only a limited set of users/responses:

USER 1. Yes; No,

USER 2. I am not sure

All else; varied input

This ensures that user responses are clear and directly relevant. In sum, we use a mix of open and closed questions to simplify processing and information management, reduce miscommunication, improve the user experience and allow better control over the dialogue.

New Structure of Item-Focused Dialogues (2nd Edition)

In the first edition, we used a set of WHO questions for evaluating both news outlets/websites and social media. However, websites and social media require different methods of evaluation. Second, it has escaped our attention that emotional manipulation, false dichotomies, cherry-picking and questionable sources/experts are shared by discredit, polarization, conspiracism, trolling, pseudoscience and science denial. Hence, these generic signs were missing from the lists of signs of the unique tactics. Third, the first edition felt less interactive, less exploratory and less Socratic. The second edition was created to amend these shortcomings while retaining and enhancing the benefits of the previous dialogues, i.e. coherence, structured interaction, shortness, clarity.

The new WHO part

We created new WHO questions that will appear in all tactics. The WHO questions are designed to help users identify disinformation by evaluating the credibility, bias, and reliability of the author and source (news outlet, social media). To amend the problem mentioned above, we have developed two separate lists of WHO questions: one for evaluating articles from news outlets/websites and another for assessing posts from social media. This distinction is necessary because news outlets and websites are evaluated based on adherence to editorial and journalistic standards, while social media posts are assessed according to verification badges, account activity, and other relevant indicators of credibility.

The new WHAT Part

#1. We divided the tactics in Generic and Unique

We noticed that emotional manipulation, false dichotomies and cherry-picking are shared constituents of discredit, polarization, conspiracism, trolling, pseudoscience and science denial. However, the opposite is not true. For example, conspiracism is an overarching strategy that may leverage cherry-picking to selectively highlight/ignore information. However, cherry-picking does not use conspiracism to go through. We called the former group generic tactics and the latter group unique tactics. The signs of the generic tactics are very important for the detection of the unique tactics but were missing from the lists of signs. Hence, in the new edition we appended the signs of the generic tactics to the signs of the unique tactics. In particular, we created signs of four generic tactics:

GENERIC TACTIC DESCRIPTION SIGNS

Emotional manipulation Tactics that exploit feelings rather than facts/logic. These appeals target emotions to persuade or influence. Emotional appeals, sensationalism, fear-mongering, provoking outrage, etc.

Denial of credibility and diversion tactics Scrutinizes the integrity of arguments and critiques used to discredit opponents by trolls, conspiracists, smear campaigns, science denialists, etc. Ad Hominem (attacking the person), Straw Man (misrepresenting an argument), Red Herring (diversion), etc.

Polarizing language and dualistic thinking Highlight rhetorical ploys and potential biases in framing issues, often creating a "us vs. them" mentality. False dilemmas, black-and-white thinking, us vs. them narratives, overgeneralizations, scapegoating, etc.

Questionable Evidence and Sources Addresses concerns about the validity of the evidence presented and the reliability of the sources used Cherry-picking data, misleading statistics, anecdotal evidence, pseudoscience, fake experts, etc.

The signs of the four generic tactics resulted as follows. First, we coupled emotional manipulation with sensationalism because they exploit feelings rather than facts/logic and are virtually shared by all unique tactics. Second, starting from the false dichotomy's tactic, we created the broader category of dualistic thinking and polarizing language to encompass a wide range of fallacies and ploys to exclude alternatives, manipulate emotions and reinforce divisions. Third, starting from the cherry-picking tactic that skews evidence, we created the broader category of Questionable Evidence and Sources to include unreliable, unverifiable, or biased sources of information or evidence. We created the category Denial of Credibility and Diversion Tactics because it is the most widespread of the generic tactics, thus shared by the unique tactics to undermine opposition, maintain control of the narrative, deflect accountability, protect interests. By appending the signs of the generic tactics to the signs of the unique tactics, the lists of signs of all tactics (unique and generic) are

now complete and more accurate. Hence, in any example dialogue, the user goes through all signs, generic and unique signs, and learns to detect more instances of disinformation than with the previous dialogues.

#2. The new structure of the WHAT part

New structure

1. One (1) question about argument or message interpretation analysis
2. Four (4) questions about generic tactics (2-5)
3. *Questions for unique tactics (6+)
4. One (1) question about agenda/purpose

*the number of questions depends on the number of tactic's signs

The first question prompts users to identify the claims, if they rest on an argument or is just sensationalism, and what the argument is. Having understood the claim or argument they use critical thinking to answer four questions designed to detect the generic signs. Then follow the questions about the unique signs of the tactic at hand which we have updated and improved. With the final question, the user critically reflects on the agenda behind the message/author.

#3. We added descriptions and rationale to all the WHO and WHAT questions

Every question is coupled with a description and a rationale. The descriptions contain more information/detail than that contained in the question. They are designed to aid users understand/answer the question. They also offer an explanation why the question is important to spark the user's reflection on the question (its rationale) and preempt those who might challenge the question. Adding descriptions and rationale to all chatbot questions enhances clarity, ensures structured dialogue, provides preemptive clarification, and enhances learning and reflection, ensuring better response quality, reducing misinterpretation. The preemptive clarification reduces the need for follow-up questions and answers, thus streamlining the conversation.

#4. We amended Branch 2 to be more responsive to users

In the first edition, the Q/A with the closed questions limited the types of users' responses/answers in two branches:

1. User 1: "Yes/No" (needs little or no guidance)
2. User 2: "I am not sure" (needs guidance)

However, this format might have felt that it is not responsive to user needs, especially those with lower media literacy who might ask for guidance, clarification and tips. So, we amended Branch 2 to anticipate users who ask questions such as "Can you show me how to do that?" or "Can you clarify this term/idea?". The amended branch goes under User 2 to aid them answer the chatbot's question by offering definitions, tips on what/how to check with examples from actual events (e.g. slogans, phrases, narratives).

#5. We created a third branch (User response) to increase interaction

The first edition allowed little/no interaction and 'debate' with the chatbot, e.g. it might have felt that there was no interaction and the dialogue was not a real conversation. So, we added a third branch to allow more interaction with users who might debate the questions or detract from the Q/A flow:

1. Neither 1 nor 2

In particular, the third branch (User 3) is designed to encompass numerous types of users, such as those who are overly skeptical, have lower media literacy, or are prone to disinformation. This includes individuals, like cynics, overly suspicious, contrarians, hyper partisans, echo chamber participants, 'red-pilled' users, devil's advocates, and similar profiles.

Branch 3 (U3): Lower Media Literacy/ Challenges the questions.

#6: We added Warnings and Refutations under Branch 3

We added critical thinking material under Branch 3 in each WHO question (Warnings) and in each WHAT question (Refutation) to anticipate users who might challenge the questions. This material is designed to encourage deeper analysis and judgment, and to prepare the general user to respond critically to specific signs of disinformation (e.g. fallacious arguments, rhetorical ploys and disinformation narratives). This is also to preempt these actual instances of disinformation.

Benefits of the new WHAT structure

#1. Detects more instances of disinformation

Now the lists of signs of the unique tactics include the generic signs that were missing from the first edition. We also updated and improved on the unique signs. The lists of signs of all tactics are now complete and more accurate. Hence, in any example dialogue, the user goes through all signs, generic and unique signs, and learns to detect more instances of disinformation than with the previous dialogues.

#2. Easier detection

Moreover, this sequence enables easier detection of the tactic at hand. Previously the WHAT part started with the question: “Do you recognize signs of polarization, conspiracism, trolling, pseudoscience, etc...?”. While this made the dialogue short and quick, the users with lower media literacy might not recognize the tactic straightaway or feel lost. Now they can choose to do signal detection step-by-step, moving from the general (e.g. emotional manipulation, questionable evidence, or attacks on credibility) to the particular (discredit or smear campaign, conspiracism, pseudoscience, etc.).

#3. The new WHAT part is more exploratory and reflective

In the first edition of the Item-Focused dialogues, we opted for quick and short dialogues, using a closed Q/A format for signal detection. This might have made some users feel like running a checklist of signs, thus leaving less space for reflective exploration and autonomy. In the second edition, the task remains the same, i.e. signal detection, but the five new WHAT questions are open enough to allow more exploration and autonomous engagement with the content/message.

#4: The new WHAT part is more ‘Socratic’

The new WHAT part resonates with TITAN’s Socratic approach. First, the five WHAT questions align with the Socratic type of questions: questioning Assumptions, Evidence/Reasons, Reasoning and Perspectives. In particular, the questions prompt the user to analyze and interpret a media message (i.e. arguments, rhetoric, context and narrative framing), question its assumptions and reasoning, evaluate the evidence and author/source, identify its perspective (e.g. angle, purpose/agenda) and how it affects the accuracy of media information. Second, the new WHAT part taps on generic critical thinking, i.e. analysis, evaluation, interpretation, inference, explanation (e.g. Facione, 2015). Third, the reader learns not only to detect the unique signs of the disinformation tactics but also considers objections to the fallacies, ploys, narratives leveraged by the tactic. This is an essential critical thinking task, inspired by Socratic Elenchus, that was missing from the first edition.

MANUAL: PART 1

INSTRUCTIONS FOR WRITERS

#1: Keep answers consistent with the three branches (user profiles)

Your answers should be aligned with three user responses:

- U1: Medium Media Literacy & Co-operative
- U2: Low Media Literacy & Co-operative
- U3: Low Media Literacy & Uncooperative/Challenger

#2: Keep answers natural with colloquial language

In order to attract the user, it is equally important to keep dialogues natural-sounding using colloquial language as it would be the communication with another person and not an AI agent.

#3: Use language of answers consistently

Be consistent in the language and way of expression we use in order to create a sense of security and familiarity for the user. We can't attract users with poorly formatted and punctually incorrect dialogue.

BEFORE WRITING AN EXAMPLE DIALOGUE

1st step: Select the tactic.

Remember we create example dialogues for each tactic. The more the example dialogues the better the chatbot. Select the tactic that your example dialogue will be about.

2nd step: Select a tactic-specific article.

Suppose you want to create an example dialogue about conspiracism. We suggest you read several articles with conspiratorial content and select one based on the following steps. Use the WHO questions to determine if the source and author are trustworthy or not. Then use the WHAT questions to determine if the article shows signs of disinformation content. If the article or post is not credible it qualifies as an example dialogue for the system.

WRITING THE ANSWERS OF EXAMPLE DIALOGUE: INSTRUCTIONS

1. Write the three answers based on the text and according to the User profiles:
 - a. User 1 has medium media literacy and has no difficulty answering the question because the question is straightforward and the Explanation/Description provides both a rationale and further hints.
 - b. User 2 has lower media literacy, thus having difficulty answering the question. Hence
 - c. User 3 has lower media literacy and detracts from the questions for various reasons. To sincerely challenge the question or because they want debate. They might want to play smart, play devil's advocate, be suspicious of the chatbot, or think they are 'red-pilled', etc. In all cases, there are two options:
 - i. They might check the text and answer the question but at the same time challenge the question or might want to debate the chatbot.
 - ii. They might totally disregard what the text says and just challenge the question or change the subject.
 - iii. Hence, your answers to U3 might either be based on the text and challenge the question or might be irrelevant with what the text says and write a challenge to the question.
2. Further interaction with users
 - a. User 1. None
 - b. User 2: Offer help/guidance for each question using the TIPS/EXAMPLES from Manual 2, Tables 2a, 2b.
 - c. User 3: Offer WARNING and REFUTATIONS for each question using the WARNINGS (Manual 2, Tables 2a, 2b) and REFUTATIONS from (Manual 2, Tables 6, 8, 10)

EXAMPLE TABLE

Question Explanation/Description	with	Answers: 3 branches
-------------------------------------	------	---------------------

1. Can you tell me if the author uses his/her real name or if he/she uses a pseudonym? Are his/her credentials verifiable? We need to know if the author is a real identifiable person with a track record, or if they are an impostor / troll with a fake identity hiding behind a pseudonym or anonymity.	USER 1: “Yes, the author seems to use his real name. ” or “No, the author seems to use a pseudonym” USER 2: “I am not sure/ I need more help. How can I check if he uses a pseudonym” CB TIP/EXAMPLE: To verify if an online author is real, check their bio for detailed background info, look for professional social media presence, email contacts, confirm their institutional affiliation via official websites and search for their name in reputable publications or third-party mentions. Do you need further help? USER 3: “Names don't matter. Some insiders or whistleblowers authors must remain anonymous for self-protection” (Debates the question) CB WARNING: Authors might sometimes use pseudonyms or remain anonymous to comply with editorial policy (e.g. the article is collaborative) or to protect their privacy and safety. But be cautious: trolls and imposters (impersonation) hide behind fake identities, anonymity and/or fake expertise to spread disinformation, avoid accountability or harass others. Beware of placing trust in authors without a proven track record or transparency. Without verifiable credentials or openness about their work, the reliability of their information and intentions may be uncertain.
--	---

MANUAL: PART 2

Manuals for writing example dialogues:

1. CONSPIRACISM
2. TROLLING
3. DISCREDIT
4. Then manual for writing example dialogue for POLARIZATION, SCIENCE DENIAL & PSEUDOSCIENCE

MANUAL FOR WRITING EXAMPLE DIALOGUES ON CONSPIRACISM

1. Begin with GENERAL INTRO: Hi there! I am CB, your personal AI coach! I can help you use critical thinking to evaluate the credibility of media items, i.e. the source and content of articles and text posts from websites and social media. I do not fact-check or authenticate: text, videos, photographs, images. How can I help you?
2. Continue with the WHO INTRO: In this section we will use critical thinking to find out WHO is behind the article, i.e. the author and the outlet, and evaluate their reputation, credentials, agendas and bias. Ready?
3. Ask and answer the WHO questions
 - a. Ask and answer each question of Table 1a (below; websites/articles) or Table 1b (social media) in arithmetic order, in the column called ‘Question with Explanation/Description’.
 - b. You will create three answers for each question based on the item, in accord with the Example Table.
 - c. Then you will add the corresponding TIP/EXAMPLE under USER 2’s question using Tables 2a or 2b, in accord with the Example Table.

d. Then you will add the corresponding WARNING under USER 3 using Tables 2a or 2b, in accordance with the Example Table.

TABLE 1A. THE WHO QUESTIONS FOR WEBSITES & ARTICLES

Blocks, Goals, Questions with Explanations/Descriptions

Blocks	Goal	Question with Explanation/Description
Author's credentials	Evaluating the credibility and qualifications of the author.	1. Can you tell me if the author uses his/her real name or if he/she uses a pseudonym? Are his/her credentials verifiable? We need to know if the author is a real identifiable person with a track record, or if they are an impostor / troll with a fake identity hiding behind a pseudonym or anonymity.
Need for expert opinion	Assessing whether an expert perspective is essential and evaluating the expertise of the author.	2. Do you think that an expert's opinion is needed to address this particular topic? If so, can you tell me if the author is himself/herself an expert on the subject or if he/she relies on other experts' opinions? Some journalists or authors may lack a track record or formal qualifications in specific fields, such as health, politics, climate science, finance, or technology, potentially affecting the credibility and depth of their reporting in those areas.
Website appearance	Evaluating the professionalism and design of a website.	3. Does the website/account have a professional or a messy appearance? There is a strong possibility that sensationalist, clickbait and fringe websites look messy, unprofessional, and cluttered with ads.
URL purpose and bias	Analyzing the intent and potential bias of a source based on its URL.	4. Can you tell me what the URL reveals about the possible purpose, angle, and bias of the source/media outlet? Here are some common endings (domains): .com (commercial), .edu (educational), .gov (government), .org (nonprofit organization) or .net (network), etc.
Funding disclosure	Examining the financial backing of a website and its potential influences.	5. Can you see in the 'About Us' section or in the footer who funds this website? We need to know who funds and advertises in an outlet because sponsors, donors and advertisers might influence what is reported.
Policy statements	Evaluating the alignment of a website's policies with standards.	6. Does the website include policy and mission statements? Good websites clarify what they publish, their fact-checking practices, how they handle users' info and how they make sure they keep up with journalistic standards.
Editorial team details	Assessing transparency and editorial standards of a website.	7. Have you noticed if the website lists its editorial team and contact details? Providing names and contact details of the editorial team ensures accountability and allows readers to reach out for further information, clarification or feedback. Knowing the editorial team also gives assurance about the publication's standards and integrity, ensuring the information is well-researched, balanced and ethical.

Timeliness of information	Considering the relevance and currency of information on a website.	8. Is timeliness important for your topic? If yes, is the information current or updated and revised? Ensuring that information is up-to-date helps maintain accuracy, relevance, and reliability. Updated/revised information ensures readers are getting the latest and most relevant data.
---------------------------	---	---

TABLE 1B. THE WHO QUESTIONS FOR SOCIAL MEDIA ACCOUNTS/POSTS

Blocks, Goals, Questions with Explanations/Descriptions

Blocks	Description	Questions with Explanations/Descriptions
Author's credentials	Evaluating whether the user/account is real or fake or impersonation	1. Is the name of the user/account the name of a person, organization/group or is it a pseudonym? Understanding the nature of the name is a key first step in figuring out whether the user/account is real—whether it stands for a real person, organization—or just hide their true identity and malicious intentions behind a pseudonym.
Verification badge		2. Does the account have a verification badge or any other indicator of authenticity from the platform? To verify if an account/user is real with a real name, look for details such as a full name and a verification badge. This badge serves as a trust signal, indicating that an account has been verified by the platform as authentic.
Profile, photo and bio of user/account		3. What does the profile suggest about the identity of the author/account? To further distinguish real accounts from impersonators or fake profiles, try to verify if the photos are clear and relevant, and check if the profile bio details are complete and consistent.
Account age and current status		4. Is the account new or has it been around for a while? Is it still active? Knowing the age of an account and the duration of its activity helps spot fake accounts. You can check the account's profile for its creation date and see whether it has a long posting activity.
Account activity and character (frequency and nature of posting)		5. What is the frequency and nature of the posts? Are there sudden changes in content or tone? The frequency of posts refers to how often a social media account publishes content. It's important because consistent posting can indicate active engagement and reliability.
Account engagement with followers		6. Does the account have a significant number of followers and legitimate engagement, or does it seem artificially inflated? The profile page shows the follower count. Genuine engagement means meaningful interactions with followers in recent posts, like comments, likes, and shares.

TABLE 2A. THE WHO QUESTIONS FOR WEBSITES & ARTICLES

Questions with Users' Answers, Tips & Warnings

QUESTION	USER 1 “Yes/No”	USER 2 “I am not sure/ I need more help”	USER 3 “WT “
1	USER 1’S ANSWER: e.g. “Yes, the author seems to use his real name.” or “No, the author seems to use a pseudonym”	CB TIP/EXAMPLE: To verify if an online author is real, check their bio for detailed background info, look for professional social media presence, email contacts, confirm their institutional affiliation via official websites and search for their name in reputable publications or third-party mentions. Do you need further help?	CB WARNING: Authors might sometimes use pseudonyms or remain anonymous to comply with editorial policy (e.g. the article is collaborative) or to protect their privacy and safety. But be cautious: trolls and imposters (impersonation) hide behind fake identities, anonymity and/or fake expertise to spread disinformation, avoid accountability or harass others. Beware of placing trust in authors without a proven track record or transparency. Without verifiable credentials or openness about their work, the reliability of their information and intentions may be uncertain.
2	USER 1’S ANSWER: “Yes/No. He/She is/isn’t an expert”	CB TIP: To decide whether an expert’s opinion is needed, first consider if the topic is complex (like medical or legal issues), involves big decisions (like health or finance/money) or involves controversial issues of science or public policy. Those qualified to speak on a topic are usually either experts with degrees in the field (such as doctors, biologists or political scientists) or specialized journalists reporting on the topic (e.g. medical reporting). Do you need more help?	CB WARNING: Independent and alternative opinions are important as far as they are expert opinions relevant to the topic. Always be careful of authors using fancy-sounding titles such as ‘finance wizard’ or ‘health guru’, ‘independent researchers’, ‘truth warriors’, ‘fighting the system’, ‘exposing Big Pharma Lies’, or ‘battling deception’. Also beware of authors who rely on personal stories instead of expert opinion, who are more famous in media than in their field, who try to sell their products, or make big claims without proof.
3	USER 1’S ANSWER: e.g. “It is/isn’t professional”	CB TIP: Check for spelling errors, broken links or inconsistent formatting, sensationalist headlines, excessive/ intrusive pop-ups and ads. Do you need further help and/or see some examples?	CB WARNING: Poorly designed websites show that nobody's really checking what gets put online. They don't clearly say where their content comes from or who's paying for it.

4	USER 1'S ANSWER: e.g. "It is commercial"	CB TIP: The purpose might reveal a possible outlet's agenda, angle and/or bias. For example, a website with the domain extension ".gov" indicates that it is affiliated with the government and likely provides official information or services. Do you need further help?	CB WARNING: Domain endings aren't everything, but they might indicate a website's angle and commitment to factual reporting. For example, a .gov or .edu website usually focuses on objective reporting and educational credibility, while a .com site might prioritize commercial content/interests, like news sites versus online shops.
5	USER 1'S ANSWER: e.g. "The 'About us' says the website is funded from.... / doesn't mention this information."	CB TIP: A good website discloses who its sponsors and donors are, it clearly marks ads and separates sponsored from general content. To find out who funds a media outlet (sponsors, donors, etc), you can try checking their 'About Us' section. You can also research ownership and consult media watchdog organizations. Do you need further help?	CB WARNING: If an outlet does not disclose financial information, you should consider this a red flag, because it does not allow readers to track its possible biases, agenda and level of independence.
6	USER 1'S ANSWER: "Yes/No."	CB TIP: A policy statement includes a commitment to honesty, fairness and accuracy, assurance of unbiased reporting, clear practices for verifying information, procedures for correcting errors, disclosure of funding sources and affiliations and policies on protecting user data. Also, the mission of reputable websites is factual reporting over sensationalism. Fringe or extremist websites focus on criticizing mainstream narratives, pushing one viewpoint or sensationalizing news to attract readers. Do you need further help?	CB WARNING: Fringe websites commit to exposing the truth and to unbiased reporting but at the same time they consistently publish articles which push one viewpoint hence, do not align with their mission. Their statements are vague, loaded with buzzwords, unrealistic promises and inconsistent with the actual content and practices of the outlet.
7	USER 1'S ANSWER: "Yes it does/No it doesn't."	CB TIP: Look for the editorial team and contact details in specific sections of the website, such as "About Us" "Contact Us" or "Team/Staff". Do you need further help?	CB WARNING: Propaganda and conspiracy theory websites, fringe websites with pseudo-science, clickbait and fake news websites have no editorial team or editor-in-chief listed and do not provide contact details
8	USER 1'S ANSWER: e.g. "Yes, and it's	CB TIP: Fields where timeliness is crucial include news and	CB WARNING: Timeliness can be a strong indicator whether

	updated/No it is not updated."	journalism, finance and economics, technology, healthcare and medicine, academic research, consumer information, sports, travel, etc. Reliable sites usually indicate whether the content of an article is updated or revised, at the beginning or the end of the article. If not, you can cross-reference the article with other reliable sources. Do you need further help?	the website is up to date and still functional. Be careful because a revised or updated article might mean that the content has been altered and misinterpreted. Outdated information can be misleading.
--	--------------------------------	---	--

TABLE 2B. THE WHO QUESTIONS FOR SOCIAL MEDIA ACCOUNTS/POSTS

Questions with Users' Answers, CB Tips & Warnings

QUESTION	USER 1 "Yes/No"	USER 2 "I am not sure/ I need more help"	USER 3 "WTF"
1	USER 1'S ANSWER: e.g. "It is pseudonym"	CB TIP: Real persons or organizations on social media use their real, full names. Pseudonyms are fake names sometimes chosen for privacy or creativity. Names and pseudonyms hinting fake accounts are often nonsensical (e.g., XafTH42), antagonistic (e.g. ClimateChangeHoax), overly common/generic (e.g., John Smith or ConcernedCitizen), offensive/antagonizing (e.g. FeminismSucks) overly patriotic or radical (e.g. TruePatriot) or fake expertise (e.g. DoctorTruth). Do you need further help?	CB WARNING: Be cautious when you encounter social media accounts who use pseudonyms. Trolls often use pseudonyms to hide their identity while posting provocative content, fraudsters and scammers to deceive, spammers to distribute unsolicited content, bots to inflate metrics and spread misinformation, marketers for deceptive promotion, impersonators to deceive fans, and business competitors to defame rivals.
2	USER 1'S ANSWER: "Yes it does/No it doesn't"	CB TIP: The verification badge is often a blue checkmark typically displayed next to the account's name or handle. This badge is found on profiles of well-known individuals, brands, or organizations and signifies that the account has been authenticated by the platform. If they do not have a verification badge, look for external verification and social proof by searching the name or account to see if it appears in trustworthy sources like news articles, professional profiles, or official sites. Do you need more help?	CB WARNING: Lacking a verification badge on a social media account isn't necessarily a bad sign. Legit accounts lack a verification badge because they have not applied for it, or do not meet the platform's eligibility criteria (like having a large following or being a notable figure), or prefer privacy. However, the badge helps check whether we are interacting with real people or reputable organizations, thus helping us distinguish real accounts from impersonators or fake profiles.

3	USER 1'S ANSWER: e.g. "The profile includes poor information about the user. The profile picture is a stock photo sourced from the internet."	CB TIP: If someone's profile pic is a generic, blurry stock photo (e.g. a celebrity or a sunset), it might be used by a fake account. Look further for detailed personal info, such as a real job, location, or hobbies—if their bio is just "Dream big, live bigger!" with no specifics or interactions, that could be a red flag. Also, watch for vague bios or mismatched details, such as someone claiming to be a teacher but having no connections related to education or a business owner with no contact info. Do you need further help?	CB WARNING: Trolls have vague or provocative bios and share very little personal info. Fraudsters and scammers create convincing profiles with tempting offers to trick people. Spammers' profiles have generic or irrelevant info without much detail, while bots have generic and repetitive bios with no personal info. Marketing accounts act like fan pages with enthusiastic promotions, while impersonators mimic real celebrities or companies, using similar names, photos, and branding to make followers think they are real.
4	USER 1'S ANSWER: e.g. "The account seems new and very active."	CB TIP: Always check the account's creation date and review its activity history. For instance, an account created last month with only a few posts or an account with large gaps in posting activity can be a red flag. Do you need further help?	CB WARNING: Older accounts with a longer, consistent history are generally more trustworthy, having built a reputation over time. New or irregularly active accounts with sudden high volume activity may raise authenticity concerns. Fake accounts are usually new, lack a consistent posting history, or have suspicious gaps in activity. However, while account age is a useful indicator, it should be combined with other factors like profile content, interaction patterns, and overall behavior.
5	USER 1'S ANSWER: e.g. "Very frequent posting, no changes in the tone."	CB TIP: Watch for accounts that post excessively, as this could indicate spam or bot behavior. If you see multiple accounts repeatedly sharing, liking, or commenting on the same posts in a suspiciously coordinated way, it might be a sign of manipulation. Accounts that post around the clock or flood your feed with memes, clickbait, or low-quality content should raise red flags. Pay attention also to discrepancies in the quality or type of content, to poorly written posts with spelling errors or awkward language, as these are again signs of automated content or bots. Do you need further help?	CB WARNING: Real users post regularly and consistently, showing genuine engagement with their followers through updates, photos, and comments. In contrast, fake accounts often have irregular patterns, like sudden spikes or repetitive posts. Watch out for these signs, as scammers, bots, and trolls often adjust their posting frequency to trick people, such as increasing activity during scams (e.g. fake offers) or spamming with irrelevant content.

6	USER 1'S ANSWER: "Yes/No."	CB TIP: See if followers engage in conversations with the account, and if the interactions appear authentic and relevant, rather than generic or spam-like. Consider using Twitter analytics for quick engagement metrics. Also, evaluate the profiles of those interacting to gauge their legitimacy. Do you need further help?	CB WARNING: Real accounts usually have lots of genuine followers who engage with their posts, while fake accounts might have bought followers or low-quality ones. Legit engagement means real people commenting and sharing because they're interested. If an account has tons of followers but few real interactions, like comments that don't make sense, it could be suspicious. Accounts with inflated follower counts might not have real engagement or could be using tactics to appear more influential than they are.
---	----------------------------	---	---

5. Continue with the reflective question: "You did a great job, well done! Keeping these in mind, do you think the author and the source can be considered trustworthy?"

Give the three answers of User 1 (e.g. I think it is/isn't reliable.), User 2 (e.g. I'm not sure yet.) and User 3 (e.g. I still can't find reasons not to trust this website.).

6. Conclude the WHO part with the question:

Answer to USER 1: "Nice work! Shall we proceed to analyze the content?"

Answer to USER 2: "I see what you mean. Checking the content of the article might help you more. Shall we proceed to analyze the content?"

Answer to USER 3: "I understand your point. It's good to have an open mind and critical evaluation always helps to ensure accuracy. Shall we proceed to analyze the content?"

7. Continue with the WHAT part

Use the following WHAT INTRO: "In this section we will use critical thinking to evaluate the credibility of the information in light of the TOP tactics of disinformation: polarization, conspiracism, discredit and trolling."

8. Ask and answer the Bridge Question:

a. Ask the following question:

"Have you spotted any signs of these tactics so far? If not, do you need help?"

b. Answer the question

i. USER 1. "Yes/No."

ii. USER 2. "I need help."

Add the TIP: "People use false information to DISCREDIT critics/opponents, i.e. to damage the reputation or undermine the ideas of persons, groups or organizations. CONSPIRACY THEORIES seed doubt about official narratives, suggesting that events or situations are the result of secret, often malicious, plots by powerful groups or individuals. Many agents of disinformation use tricks and false information to POLARIZE, i.e. to create division and stir up conflicts in society. TROLLS also use disinformation to provoke outrage, mislead, and disrupt online conversations.

iii. USER 3. Debates the question/ doesn't answer/detracts from the discussion

Use the WARNING:

“It’s good to have an open and critical mind with the content. People use false information to DISCREDIT critics/opponents, i.e. to damage the reputation or undermine the ideas of persons, groups or organizations. CONSPIRACY THEORIES seed doubt about official narratives, suggesting that events or situations are the result of secret, often malicious, plots by powerful groups or individuals. Many agents of disinformation use tricks and false information to POLARIZE, i.e. to create division and stir up conflicts in society. TROLLS also use disinformation to provoke outrage, mislead, and disrupt online conversations. Do you recognize any sign of these?”

9. Use this INTRO: “First, we need to clarify the thesis of the article/post and argument and try to find out if it is factual or just speculation. Ready?”.

10. Ask and answer the WHAT Questions about General Signs of Disinformation

- a. Ask and answer each question of Table 3 in arithmetic order (see column called ‘Question with Explanation/Description’).
- b. Create three answers for each question based on the item, in accordance with the Example Table.
- c. Finally add the corresponding TIP/EXAMPLE under USER 2’s question, using Table 4, in accord with the Example Table.

TABLE 3: THE FIVE GENERAL WHAT QUESTIONS ABOUT ALL TACTICS

Blocks, Goals & Questions with Explanations/Descriptions

Blocks	Goal	Question with Explanation/Description
Clarification and Argument Evaluation		9. What is the main claim or message of the author/user? Does he/she use an argument and facts to make his/her case or is it his/her personal speculation/opinion? <i>Understanding the main claim clarifies the author’s message and context. Also, distinguishing facts (provable truths) from opinions (subjective beliefs) helps reveal if they are presenting opinions as ‘facts’ without argument and evidence.</i>
Emotional Manipulation and Sensationalism		10. Does the author/user rely on sensationalist content/headlines and emotional manipulation? Some authors use exaggerated and loaded language in the content or in the headlines, to sway readers by appealing to their emotions or provoke strong emotions like outrage, fear, and suspicion.
Denial of Credibility & Diversion Tactics		11. Does the author/user attack the credibility of critics, opponents, experts, mainstream media, institutions? Smear campaigns, conspiracy theories and trolls, use ad hominem attacks, smear, straw man arguments, selective quoting, whataboutism to damage the reputation/credibility of their targets and distort their views.
Polarizing language and Binary Thinking		12. Does the author/user rely on black and white thinking and/or polarizing language? Trolls, propagandists and conspiracy theorists frame everything as either completely good or completely evil or "us" vs. "them", demonize opponents, and use false dichotomies to oversimplify complex issues or deepen societal divisions.
Questionable Evidence & Experts		13. Does the author/user provide credible evidence for claims that rely on facts? Conspiracy theories, trolls, smear campaigns provide no evidence or seeming evidence for their factual claims. Also, they cherry-pick data, fabricate content, create misleading narratives, or use pseudoscience.

TABLE 4. ANSWERS TO QUESTIONS THREE BRANCHES

General WHAT Questions	TIPS (Answer to User 2)
9. What is the main claim or message of the author/user? Does he/she use an argument and facts to make his/her case or is it his/her personal speculation/opinion?	You need to track the key message the author wants readers to take away. Knowing the difference between fact (provable truth) and opinion (personal /subjective belief) is key to spotting arguments and understanding what a strong argument is. Arguments are the reasons and evidence that one gives to support or oppose a claim, idea, or viewpoint. The better the reasons one provides, the stronger the argument. Do you need further help with spotting and evaluating arguments?
10. Does the author/user rely on sensationalist content/headlines and emotional manipulation?	Sensationalist content evokes strong emotional reactions by using exaggerated claims or conspiracy claims, fear mongering and scare tactics, overgeneralizations, false dichotomies, celebrity endorsements, urgency and pressure, sensational visuals. Emotional manipulation involves exploiting people's feelings—such as fear, anger, or sympathy—to influence their opinions or actions, often by presenting misleading, sensational or false information. Do you need further help?
11. Does the author/user attack the credibility of critics, opponents, experts, mainstream media, institutions?	Ad hominem is the fallacy of attacking the person's character, motive, or other attribute, instead of discussing the substance of the person's argument itself. Another tactic is Straw Man and Selective Quoting, i.e. taking statements out of context or selectively quoting to misrepresent someone's views or arguments. Whataboutism, i.e. responding to criticism or discussion points by deflecting to unrelated issues or past events. Do you need further help?
12. Does the author/user rely on black and white thinking and/or polarizing language?	False Dichotomy frames issues as having only two extreme options, ignoring other possibilities. For instance, labeling vaccine opponents as "anti-vaxxers" versus "pro-science" oversimplifies the debate. Polarizing Language uses divisive terms like "medical tyranny" to stir conflict and deepen divides, especially in discussions about pandemic measures. Demonization casts groups or individuals as evil; for example, referring to climate activists as "alarmists" or "extremists" to discredit them. Is that clear or do you want to read more?
13. Does the author/user provide credible evidence for claims that rely on facts?	Factual claims (e.g. "CO ₂ contributes to global warming") and fact-based proposals (e.g. "we must reduce carbon emissions") are credible when they rely on reliable factual evidence (e.g. data on emissions) from expert testimony. Some authors cite questionable experts/sources, e.g. unnamed "insiders", anecdotes, or sensational claims such as "Top expert uncovers conspiracy". Other cherry-pick data, i.e. select only the evidence that supports their argument while ignoring evidence against it, or use pseudoscience, i.e. theories and methods that claim to be scientific but lack the rigor and evidence of true science. Do you need further help?

11. Ask and answer the Bridge INTRO & Question

a. Use the following text to go to the tactic's unique questions: "Let us now dive deeper into the narratives, psychological tricks and agendas of the top-four tactics, trolling, discredit, conspiracism and polarization. Also,

explore objections to these tactics and learn how to dismantle them. Which tactics do you recognize now and want to dive deeper?

b. Answer the question

i. USER 1. "Conspiracism."

ii. USER 2. "I need help."

lii. USER 3. "...." (debating the CB)

12. Use the following CONSPIRACY INTRO text:

"Let's figure out if the author's assumptions (beliefs/narratives) about the supposed conspirators match the usual signs of conspiracy theories. You will also see if these assumptions hold up under close analysis."

13. Ask and answer the WHAT Questions about the Unique Signs of Conspiracism:

a. Ask and answer each question of Table 5 in arithmetic order, in the column called 'Question with Explanation/Description'.

b. Create three answers for each question based on the item, in accord with the Example Table.

c. Then add the corresponding TIP/EXAMPLE under USER 2's question, using Table 6, in accord with the Example Table.

d. Finally, add the corresponding WARNING under USER's 3 challenges, using Table 6, in accord with the Example Table.

TABLE 5. QUESTIONS & DESCRIPTIONS ABOUT CONSPIRACY THEORY

CONSPIRACY THEORIES SIGN(S)	QUESTION & DESCRIPTION
a. There is a secret plot by powerful groups b. Narrative of "suppression of truth" by the alleged conspirators.	1. Does the author assume a secret plot by powerful groups who can hide the truth? Conspiracy theories usually attribute random events and coincidences to a deliberate secret plot orchestrated by a powerful group of conspirators, who are extremely efficient in suppressing the truth.
c. This secret group has nefarious intent/agenda against the public	2. Does this secret group have a nefarious intent/agenda? What is the agenda and motives of the alleged conspirators? Conspiracy theorists support that the hidden agenda of secret plots have evil or harmful goals (e.g. to control the government to impose authoritarian rule). But often lack a coherent explanation for why the alleged conspirators would engage in harmful actions. The motives they assign to these plotters are usually overly simplistic, absurdly grandiose, or unrealistically evil.
d. Narrative of persecuted victims or heroes	3. Does the author invoke themes of persecution, victimhood and heroism vs. dark forces? Conspiracy narratives frame a battle of good versus evil, where the "people" are portrayed as oppressed and victimized by powerful, dark forces acting against the interests of society.

e. "Something must be wrong" with the official accounts	4. Does the author assume that "something must be wrong" (i.e. deception, lies or propaganda) with the official accounts? Conspiracy theories are based on persistent suspicion/doubt of official narratives, thus dismissing any information coming from governments, scientific bodies, experts and mainstream media.
f. Unfalsifiable/self-sealed claims, i.e. no evidence can prove them wrong because any evidence against them is dismissed as part of the conspiracy or as being <u>deliberately hidden</u> . g. Narrative of "suppression of truth" by the alleged conspirators.	5. Is the claim/theory of the author falsifiable? Conspiracy theories cannot be proven wrong because they are "unfalsifiable" or "self-sealed", i.e. no evidence can prove them wrong because any evidence against them is dismissed as part of the conspiracy or as being deliberately hidden.
h. Author has inside knowledge of the secret plot or relies on anonymous insiders	6. Does the author claim to have inside knowledge or uses evidence from anonymous insiders, anecdotal accounts or rumors? Do you trust this evidence? Conspiracy theories often say they have inside knowledge or 'evidence' from anonymous insiders/whistleblowers who have access to the secret plots.

14. Continue with the PURPOSE INTRO:

"Let us recap the author's and the outlet's agenda. This critical approach provides insight into the motives and bias influencing the presentation and interpretation of information.

15. Ask and answer the PURPOSE question/explanation:

c. Purpose Question: "What do you think about the agenda or motives of the author and the outlet now?" following with the question's explanation:

"Disinformation efforts include agendas such as political manipulation, geopolitical influence, social division, personal motives such as attention-grabbing, economic gain, ideological warfare, and undermining trust in institutions and scientific consensus."

d. Give the User 1, User 2, User 3 answers:

b. e.g. User 1: "Maybe it is both ideological and financial reasons."

c. e.g. User 2: "I don't know. Maybe it is for financial reasons."

d. e.g. User 3: "Maybe it is for financial reasons but it might also tell us the truth about the real dangers of vaccines."

e. CB answer as follows:

e. to USER 1: "Do you have any reservations about this article's conspiracy content? Or Do you want to check any of the signs or tactics in more detail? "

f. to USER 2: "To identify political/geopolitical motives, ask questions like the following: "Does the information promote or criticize specific political figures, parties, policies or countries?". To find out if the motive is social disruption, ask "Does the content question the legitimacy of institutions such as the government, scientists, media, or healthcare systems?" To find out if the motive is financial, you may ask: "Is there a call to action that involves spending money, such as buying a product, subscribing to a service, or donating to a cause?" To find out if the motive is Ideological Propagation, try questions like, "Does the information promote a specific set of beliefs or values, such as religious, cultural, or social ideologies?". Do you want a micro lesson about media agendas and media bias? "

g. to USER 3: “I understand the reasons why you might feel distrustful or skeptical, especially with so much information and misinformation out there. However, it can be helpful to look at the evidence and evaluate it critically. Do you want to check any of the signs or tactics in more detail? “

14. Conclude the dialogue thanking the User for his/her participation in the discussion.

MANUAL FOR WRITING EXAMPLE DIALOGUES ON TROLLING

Included all the above 14 steps from the Manual for writing dialogues on Conspiracism.

Here are the Unique Signs, Questions with Descriptions, Tips to User 2 and Refutations to User 3 for Trolling.

TABLE 7. UNIQUE SIGNS/QUESTIONS & DESCRIPTIONS ABOUT TROLLING

TROLLING SIGN(S)	QUESTION & DESCRIPTION
a.Exaggerated or Outrageous Claims: The user makes exaggerated or outrageous claims about sensitive issues to provoke emotional reactions or spread misinformation. Off-Topic Comments: The user persistently posts off-topic comments or attempts to redirect the discussion to irrelevant and inflammatory topics, disrupting the conversation.	1. Does the user make exaggerated or outrageous claims about sensitive issues? Trolls bait people by exploiting sensitive issues like race, gender, or politics with exaggerated claims, outrageous proposals, hoaxes, or misleading stories to provoke outrage, start heated debates, and deepen divides. 2. Do the user post persistent off-topic comments in the discussion or try to direct the discussion to irrelevant inflammatory topics? Trolls derail threads by bringing up unrelated points or shifting the conversation to irrelevant but inflammatory topics, making it hard for participants to stay focused on the main topic, leading to confusion and conflict.
b. Flooding with Posts: The user floods the conversation with an excessive number of rapid, repetitive, or contradictory posts/messages, overwhelming and confusing other participants.	3. Does the user flood the conversation with an excessive number of rapid, repetitive or contradictory posts/messages? A single troll or a small group may spam repetitive messages, coordinate to flood the thread with disruptive content (brigading), or use fire-hosing tactics with high-volume
c. Provocation through Insults: The user focuses solely on provocation by using insults, personal attacks, and inflammatory language to incite conflict and derail discussions.	4. Does the user focus solely on provocation by using insults, personal attacks, and inflammatory language? Trolls use derogatory comments and personal insults, highly charged, exaggerated language to upset, outrage and create disagreement, ultimately disrupting the conversation without contributing to the discussion.
d. Fake Polite Requests for Proof: The user repeatedly asks politely for proof or more details in a way that seems insincere or annoying, even after the topic has been thoroughly covered, to stall or divert the conversation.	5. Does the user keep asking politely for proof or more details in a way that feels fake or annoying, even though the topic has already been covered? Sea-lioning is a favored tactic of trolls: they ask persistent, repetitive questions under the guise of politeness and curiosity or confusion to derail

	discussions, exhaust participants, and avoid engaging with the topic meaningfully.
e. Hashtag Campaign Participation: The user's posts are part of a hashtag campaign.	6. Is the post part of a hashtag campaign? If it is, what do we know about this campaign/thread (i.e. can you identify who started it and for what purpose?)? Trolls, hoaxers and/or impersonators sometimes do "hashtag hijacking/spam", i.e. appropriate hashtag campaigns (i.e. #NationalFriedChickenDay (KFC), etc.) to achieve their malicious purposes by undermining its original intent.
f. Inflammatory Jokes and Memes: The user posts mocking or inflammatory jokes and memes, or uses satire inappropriately, to belittle others or spread divisive content.	7. Does the user post mocking or inflammatory jokes and memes, or use satire inappropriately? Trolls use inappropriate humor, sarcasm, and satire—often flooding discussions with mocking memes and dark humor—to trivialize serious issues, disrupt conversations, and provoke conflict.
G. Pretending to Oppose Group: The user may push inflammatory or polarizing content by pretending to be part of a group they oppose, aiming to discredit or sow discord within that group.	8. Is there a chance that the user is pushing inflammatory or polarizing content by pretending to be part of a group they actually oppose? So-called 'False-flag' trolls fake being part of a group or pretend to support an opinion they actually oppose. They often show exaggerated support for shocking ideas or push far-fetched, offensive measures to make the group look bad and provoke a negative reaction.

TABLE 8. ANSWERS TO TROLLING (THREE BRANCHES)

TROLLING SIGNS	TIPS (ANSWER TO USER 2)	REFUTATIONS (ANSWER TO USER 3)
a. Exaggerated or Outrageous Claims: The user makes exaggerated or outrageous claims about sensitive issues to provoke emotional reactions or spread misinformation	Reflect on whether the post addresses genuine issues or exploits sensitive topics like gender rights, racial justice, immigration, freedom of speech, privacy, economic policy, climate change, multiculturalism, or religious symbols, that may cause heated debates. Watch for outrageous claims about individuals and groups, sweeping generalizations, and extreme proposals such as "End Islamization—#BanIslam Now!", "Calexit Now—California deserves its own future!", and "Solve overpopulation—implement	Trolls exploit sensitive issues by inflating threats and blaming specific groups (fear-mongering and scapegoating), by presenting only extreme options and ignoring the middle ground (false dichotomies), by cherry-picking or twisting data to fit their agenda (misleading statistics and data manipulation), by exaggerating the severity of situations to justify drastic measures (moral panic and crisis amplification) and by using

	mandatory euthanasia for the elderly!". Do you need further help?	personal or group identities to provoke reactions and deepen divisions (identity politics).
b. Off-Topic Comments: The user persistently posts off-topic comments or attempts to redirect the discussion to irrelevant and inflammatory topics, disrupting the conversation	Check if the user constantly makes unrelated comments in a thread, for instance, in a thread about renewable energy, a troll might suddenly discuss immigration, gun control or government control conspiracies. Here is another example: Imagine how a discussion about falling crime rates and the importance of the efforts of a community and/or institution, could be diverted by mentioning an isolated incident such as a recent, terrible, crime (or vice-versa). Would you like to discuss more and see some more examples of trolling attempts to redirect the discussion?	To derail a thread, trolls use Red Herring to divert attention by introducing irrelevant information, Gish Gallop to overwhelm with excessive, misleading information, Appeal to Emotion or Moral Panic to create alarm about something that seems to threaten a society's values, and Straw Man Arguments to misrepresent and attack distorted versions of arguments.
c. Flooding with Posts: The user floods the conversation with an excessive number of rapid, repetitive, or contradictory posts/messages, overwhelming and confusing other participants.	Having checked the posting history of an account/user, as well as its credentials, might help you determine whether a person (or automated bot) is aiming at flooding a conversation/discussion thread. Check for over-posting, repetitive messages and comments (i.e. spamming). Here's an example: From the early 2000's the Chinese government was suspected of hiring many people (a.k.a. the '50 Cent Party') to flood the web in China with an overwhelming amount of positive news and pro-regime messages. Would you like to see some more examples and learn more details about 'flooding' techniques?	Disinformation tactics like 'flooding', 'astroturfing' and/or 'fire-hosing', usually employed by trolls (persons and/or automated bots), involve spamming posts and/or comments with the intention of shaping a narrative or drowning opposing viewpoints. Moreover, by disrupting legitimate debate trolls might aim to irritate and vex internet users, therefore discouraging people from participating in online spaces.
d. Provocation through Insults: The user focuses solely on provocation by using insults, personal attacks, and inflammatory language to incite conflict and derail discussions	Trolls use inflammatory and abusive tactics to push emotional buttons, thus stirring up arguments rather than sharing useful info or offering reasoned discussion. Always be mindful of excessive sarcasm or irony (e.g., "Oh, look at Mr. Expert"), personal attacks with belittling remarks (e.g., "Only an idiot would believe in climate change!", "Typical liberal snowflake—can't handle the truth!"), personal insults (e.g., "You're just too stupid to understand this!"), extreme, inflammatory statements (e.g. "Only a "Nazi would say that"). Do you need further help?	Trolls leverage emotional provocation to disrupt discussions and escalate conflicts by using Ad Hominem to attack individuals' character instead of their arguments, Appeal to Ridicule to mock and dismiss viewpoints as absurd, Appeal to Force to intimidate and silence through threats, Appeal to Fear of possible threats or catastrophes, False Dilemma to present issues in extreme terms and create divisions, and Straw Man to misrepresent and simplify arguments for easier attack.

e. Fake Polite Requests for Proof: The user repeatedly asks politely for proof or more details in a way that seems insincere or annoying, even after the topic has been thoroughly covered, to stall or divert the conversation.	In a discussion about climate change, sea-lioning can involve repeatedly asking repeatedly for clarification (e.g. Can you explain it again?"), maintaining a facade of politeness while being intrusive (e.g., "I'm just asking for a simple explanation of the greenhouse effect, why won't you help me understand?"), ignoring or dismissing evidence and continuing to ask the same questions (e.g., "That source isn't credible, can you find another one?"), and aiming to exhaust and frustrate participants (e.g., "Why are you avoiding my questions about climate models?"). <i>Do you need further help?</i>	Trolls might use sea-lioning to pass the burden of proof by constantly asking for more details, making it seem like if you can't provide everything, your point is wrong (Shifting the Burden of Proof). They often distract from the main topic with endless, irrelevant questions (Red Herring). Just when you answer, they change their demands, making it impossible to move the discussion forward (Moving the Goalposts). They might twist your argument to attack a weaker version (Straw Man), or create a false choice where you either give endless details or give up (False Dilemma).
f. Hashtag Campaign Participation: The user's posts are part of a hashtag campaign	Sometimes, trolls might infiltrate a hashtag thread/campaign launched by famous brands, activists' organizations, influencers, political campaigners, etc., to promote a different agenda, make false associations, hijack or disrupt the thread. Here's an example: K-pop (Korean pop music) fans tagged pictures and videos of a Korean pop band with #whitelivesmatter, a hashtag commonly used by American white supremacists (i.e. false association). Do you want to learn more about hashtag hijacking and see some more examples?	Trolls/hoaxers might monitor trending hashtags or create their own to spread disinformation and/or harass targeted individuals/groups. This enables them to inject their messages into popular conversations and gain more exposure. By posing as someone else, hijacking or distorting an existing campaign they might try to deliberately spread disinformation.
g. Inflammatory Jokes and Memes: The user posts mocking or inflammatory jokes and memes, or uses satire inappropriately, to belittle others or spread divisive content.	Appropriate humor or satire aims to entertain or critique thoughtfully, respecting the subject and promoting constructive dialogue. On the other hand, trolls weaponize humor or satire inappropriately. For example, during the protests following George Floyd's death in 2020, trolls created and shared offensive memes and jokes to mock the Black Lives Matter movement and the protests. In defense, trolls might say, "Just a joke, bro—if you can't take it, you're too soft", "I guess some people can't handle a little satire". Do you need further help?	When trolls use inappropriate humor or mocking, they often appeal to ridicule and minimize the importance of a serious or sensitive issue by treating it lightly as a joke. They might use jokes or satire to attack people personally rather than their arguments, or as red herrings to distract from the main issue. Often, they set up a straw man by distorting arguments into something absurd and then mock that distorted version.
h. Pretending to Oppose Group: The user may push inflammatory or polarizing	Try fact-checking extreme claims with credible sources as radical claims often signal trolling. Assess their understanding and knowledge of the group since false flag trolls usually have shallow understanding of the group's actual views.	False-flag trolls distort or exaggerate the values of the group they claim to support to make it appear extreme or threatening for the values of society, thereby causing moral

content by pretending to be part of a group they oppose, aiming to discredit or sow discord within that group.	Watch for exaggerations as trolls aim to mislead and provoke reactions. Here is an actual example of over-the-top advocacy: trolls posed as Bernie Sanders' supporters, claiming to abolish the Federal Reserve, to discredit him. Do you wish to discuss further?	panic (i.e. unnecessary fear) or outrage within society. This form of trolling is very difficult to detect since the author might sincerely advocate the cause or be part of coordinated and/or sponsored efforts to discredit or sow discord (e.g. political or state-sponsored propaganda troll factories/farms, etc.)
--	--	--

MANUAL FOR WRITING EXAMPLE DIALOGUES ON DISCREDIT

Included all the above 14 steps from the Manual for writing dialogues on Conspiracy.

Here are the Unique Signs, Questions with Descriptions, Tips to User 2 and Refutations to User 3 for Discredit.

TABLE 9. UNIQUE SIGNS/ QUESTIONS & DESCRIPTIONS ABOUT DISCREDIT

DISCREDIT SIGN(S)	QUESTION & DESCRIPTION
Personal Attacks and Offensive Language: The author uses personal attacks instead of reasoned arguments, employing abusive/offensive language, insults, name-calling, and irony to attack the opponent's moral character, motives, or ideas.	1. Let's see in more detail if the author uses personal attacks instead of reasoned arguments. Does he/she use abusive/offensive language, insults, name calling and/or irony to attack the opponent's moral character, motives or cause/ideas? These rhetorical attacks, by appealing to our feelings and prejudices, aim at turning our attention away from an opponent's arguments (and what they represent) highlighting irrelevant aspects (or deficiencies) of said person, attacking their credentials, ridiculing them, commenting on the manner they express their arguments, or connecting them with something that might be unrelated and/or has negative connotations.
Misleading Audiovisual Material: The article is accompanied by audiovisual material that may reinforce or create misleading narratives, attempting to erode trust or discredit opponents.	2. To this end, audiovisual material is commonly used as part of these attacks, reinforcing or creating misleading narratives in an attempt to further erode trust or discredit opponents. Can you tell me if the article is accompanied by any (potentially harmful) audiovisual material? It is scientifically established that audiovisual material can be highly effective in inducing powerful emotions and feelings thus affecting our reactions, attitudes and responses. 'This tendency is likely due to visual information coming with an implicit guarantee of being closer to the truth than other forms of communication'.
Targeting Public Figures in Organized Campaigns: The target of the attacks is a well-known public figure or group, suggesting that the article might be part of an	3. Personal attacks, like those we discussed, are also used in organized campaigns (i.e. by political opponents, newspapers, etc.) that aim at smearing and discrediting. Common targets of such attacks are usually important public figures

organized/deliberate effort to smear or discredit.	or groups (e.g. politicians, NGOs, scientists, activists, investigative journalists, etc.) and not the average person. Can you tell me if the target of these attacks is a well-known, public figure? Do you think that this article might constitute an organized/deliberate effort to smear/discredit? Moreover, evidence exists to suggest that smear campaigns, like other forms of disinformation, rely on how (i.e. based on social dominant norms, background, standing, etc.) we tend to perceive people who might be different. Our tendency to prefer people/groups/institutions more familiar can trigger an 'us versus them' mentality that can be exploited in organized efforts to discredit someone or something.
Seeding Doubts and Negative Labels: The article aims to seed doubts about the trustworthiness of the target by attaching short labels with strong negative connotations, which may not accurately represent the target's work, motives, or affiliations.	4. By seeding doubts that someone/something 'can't be trusted' the credibility of a potential target might suffer severely. For this purpose, smear campaigns seek to attach short labels to their targets that might not represent their actual work, motives and/or affiliations. After reading this article can we assume that the target of these attacks 'can't be trusted'? Is the target being labeled and associated with anything that has strong negative connotations? Theory suggests that labeling might have a powerful impact on our own behavior or on the way we perceive others. For example, because labels can be 'sticky', they might lead to severe stigmatization or even complete social rejection (negative effect).

TABLE 10. ANSWERS TO DISCREDIT (THREE BRANCHES)

DISCREDIT SIGNS	TIPS (ANSWER TO USER 2)	REFUTATIONS (ANSWER TO USER 3)
a. Personal Attacks and Offensive Language: The author uses personal attacks instead of reasoned arguments, employing abusive/offensive language, insults, name-calling, and irony to attack the opponent's moral character, motives, or ideas.	Look for words and phrases used to describe the opponent's personal characteristics/traits usually with negative connotations and/or ironic intent (i.e. they are immoral, incompetent, stupid/ignorant, are being manipulated and/or their bodily traits/habits are unfavorably portrayed, etc.). Try to determine whether the author is being dramatic on his/her description of said	Personal attacks usually aim at discrediting opponents, damaging one's reputation or deflect criticism. They can be the product of (sometimes organized) propaganda aiming to create suspicions and delegitimize a person by eroding trust. A common result is the manipulation and confusion of the general public thus fostering

	person and avoids reasoned debate with structured arguments. Here are some examples: “We shouldn't read this book by Freud. Everyone knows he used cocaine.” “We cannot approve of this recycling idea. It was thought of by a bunch of hippie communist weirdos”, “I don't have any confidence in his expertise, since he has obvious trouble controlling his own appetite”, etc. Do you need any further help determining a personal, ad hominem attack?	distrust for certain persons, groups and institutions.
b. Misleading Audiovisual Material: The article is accompanied by audiovisual material that may reinforce or create misleading narratives, attempting to erode trust or discredit opponents.	In case the article is accompanied by pictures you find disturbing or otherwise make you feel that they are deliberately used; you might want to check their authenticity using image authentication apps. In case the images/audiovisual material used appears to be trying to discredit/delegitimize/ridicule someone or something (i.e. funny/satirical representations, negative associations, alignment of irrelevant or questionable pictures, etc.) try to take a step back and reexamine its potential purpose and context it's used in. Would you like to see some examples of discrediting audiovisual material?	As we already discussed, refuting what we see can be difficult. Therefore, we must always be mindful of any underlying/implicit (or false) associations that may accompany an image/video. Think: is there any other evidence to suggest that a claim is valid aside from what I am presented with? Do the actual claims/opinions/habits of the person(s) in question match with the way they are presented? Is the audiovisual material accompanying the article accurate (i.e. is the time and place accurate/relevant? Does the author bother to reference a source? etc.

c. Targeting Public Figures in Organized Campaigns: The target of the attacks is a well-known public figure or group, suggesting that the article might be part of an organized/deliberate effort to smear or discredit.	Try checking what else has been published by your source regarding this person/group. If the website has a habit of constantly attacking, making a one-sided, unfavorable presentation without hosting any alternative viewpoints, it might be a strong indication its purpose is to push a specific viewpoint or agenda. Here are some examples: “In the 1950s, Senator Joseph McCarthy used a smear campaign to accuse people of being communists. He would make baseless accusations and use fear tactics to ruin reputations and careers”. “In the 1990s, McDonald's launched a smear campaign against Burger King by portraying their burgers as inferior and unhealthy. This led to a decline in Burger King's sales and a boost for McDonald's”. Would you like to see some more examples and/or information on how smear campaigns work?	Organized efforts to discredit, such as smear campaigns, are usually based on rumors and/or false/fabricated accusations. Like personal (ad hominem) attacks, smear campaigns tend to make use of exaggerations, distortions or even outright lies, etc. in order to avoid discussing real substance, deflect criticism, undermine the credibility of their target, delegitimize their ideas/cause, or even silence them. If you feel that after reading this article your trust to a person/group/institution is irredeemably damaged, try to maintain an open mind towards other interpretations, evidence and opinions.
d. Seeding Doubts and Negative Labels: The article aims to seed doubts about the trustworthiness of the target by attaching short labels with strong negative connotations, which may not accurately represent the target's work, motives, or affiliations.	Attaching short labels to targets might be as profound as attributing a nickname with negative connotations and/or ironic intent. Try looking for words that are directly linked to the target's name and/or are repeatedly used to describe the target's personality/identity/traits resulting in a potentially direct association of said person with what the author claims to be true about them. Here is an example: “Sleepy Joe” (among other unflattering nicknames/labels) is a nickname describing Joe Biden, the 46th president of the United States, coined and used by those who oppose his presidency, which has	Personal attacks (especially when orchestrated) can spread potentially disinformative content faster and more extensively than the truth. The implications of attacking the credibility of a person (i.e. a politician) or group (i.e. the scientific community) can extend far beyond the individual/group. The resulting erosion of trust dampens the public's ability to differentiate between real and fake, legitimate/non-legitimate, who to trust, etc. Moreover, these confrontational/heated situations suppress alternative narratives/viewpoints and

	become an Internet meme. The nickname was originally coined in 2019 by Biden's predecessor, Donald Trump. It appeared after a series of videos where Biden appears to occasionally fall asleep during important meetings and television interviews. Would you like to see some more examples and further discuss how labeling can affect someone's reputation?	discourage people from engaging in productive debate.
--	---	---

ANNEX 4: MEDIA LITERACY PROTOCOL

A. The CRAAP TEST ([CRAAP Test - Evaluating Resources and Misinformation - Library Guides at UChicago](#))

C	Currency: The timeliness of the info
R	Relevance: How the info fits your needs
A	Authority: The source of the info
A	Accuracy: Reliability and correctness of the info
P	Purpose: The reason the info exists

Use questions from each dimension you need, when you evaluate the author/source (WHO) or the content of the item (WHAT).

Currency: the timeliness of the information

- When was the information published or posted?
- Has the information been revised or updated?
- Is the information current or out-of-date for your topic?
- Are the links functional?

Relevance: the importance of the information for your needs

- Does the information relate to your topic or answer your question?
- Who is the intended audience?
- Is the information at an appropriate level (i.e. not too elementary or advanced for your needs)?
- Have you looked at a variety of sources before determining this is one you will use?
- Would you be comfortable citing this source in your research paper?

Authority: the source of the information (WHO)

- Who is the author/publisher/source/sponsor?
- Are the author's credentials or organizational affiliations given?
- What are the author's credentials or organizational affiliations given?
- What are the author's qualifications to write on the topic?
- Is there contact information, such as a publisher or e-mail address?
- Does the URL reveal anything about the author or source?
 - examples:
 - .com (commercial), .edu (educational), .gov (U.S. government)
 - .org (nonprofit organization), or
 - .net (network)

Accuracy: The reliability, truthfulness, and correctness of the content (WHAT)

- Where does the information come from?

- Is the information supported by evidence?
- Has the information been reviewed or refereed?
- Can you verify any of the information in another source or from personal knowledge?
- Does the language or tone seem unbiased and free of emotion?
- Are there spelling, grammar or typographical errors?

Purpose: the reason the information exists

- What is the purpose of the information? to inform? teach? sell? entertain? persuade?
- Do the authors/sponsors make their intentions or purpose clear?
- Is the information fact? opinion? propaganda?
- Does the point of view appear objective and impartial?
- Are there political, ideological, cultural, religious, institutional, or personal biases?

B. The **SIFT METHOD** ([The SIFT Method - Evaluating Resources and Misinformation - Library Guides at UChicago](#))



The SIFT method is an evaluation strategy developed by digital literacy expert, Mike Caulfield, to help determine whether online content can be trusted for credible or reliable sources of information.

STOP

What you already know about the topic.

What you know about the source. Do you know its reputation?

INVESTIGATE THE SOURCE

What can you find about the author/website creators?

What is their mission? Do they have vested interests? Would their assessment be biased?

Do they have authority in the area?

Use lateral reading. Go beyond the 'About Us' section on the organization's website and see what other, trusted sources say about the source. You can use Google or Wikipedia to investigate the source.

Hovering is another technique to learn more about who is sharing information, especially on social media platforms such as Twitter.

FIND BETTER COVERAGE

What coverage is available on the topic?

Keep track of trusted news sources.

Many times, fact checkers have already looked into the claims. These fact-checkers are often nonpartisan, nonprofit websites that try to increase public knowledge and understanding by fact checking claims to see if they are based on fact or if they are biased/not supported by evidence.

- [FactCheck.org](#)
- [Snopes.com](#)
- [Washington Post Fact Checker](#)
- [PolitiFact](#)

TRACE CLAIMS, QUOTES AND MEDIA TO THEIR ORIGINAL CONTEXT

Was the claim, quote, or media fairly represented?

Does the extracted information support the original claims in the research?

Is information being cherry-picked to support an agenda or a bias?

Is information being taken out of context?

Remember, headlines, blog posts, or tweets may sensationalize facts to get more attention or clicks. Re-reporting may omit, misinterpret, or select certain facts to support biased claims. If the claim is taken from a source who took it from another source, important facts and contextual information can be left out. Make sure to read the claims in the original context in which they were presented.

C. The SMART CHECK ([The SMART Check - Evaluating Resources and Misinformation - Library Guides at UChicago](#))

The SMART Check is particularly helpful when evaluating news stories. Determine if your news source is SMART before believing what is reported.

[Lone Star College Library:](#)

S	Scope	What is the scope of the article? Does it cover the subject with enough depth and detail?
M	Motivation	What is the purpose of the article? To sell you an idea or to inform?
A	Authority Accuracy	Who created the article, and what are the author's credentials? Can the information be verified?
R	Relevancy Reliable	Is the content appropriate for your needs? Do you have to have scholarly sources?
T	Timely	Is the content current?

D. The A.S.P.E.C.T. TEST ([A.S.P.E.C.T. - Evaluating Information - Research Guides at Clark College \(libguides.com\)](#))

A: Authority

- Is there an author's name?
- Can you locate the author's credentials?
- Can you find evidence of author expertise in the subject?
- Have you located similar works by this author?
- Do you have personal recommendations for this author?
- Do you know the publisher's credentials and reputation?

- Are there similar works from this publisher?

S: Sources

- Is information presented as fact? If yes...
 - Does the author provide documentation? (Bibliography, footnotes, links, etc.)
- If documentation/sources are included, are they from credible sources?

P: Purpose

- Was this source written to inform and educate?
- Does the source argue a perspective or specific opinion?
- Is the source intended to entertain or sell?
- Is the content aimed at a general audience, or is it written for readers with expertise in the subject?
- Is the source too basic, too technical, too advanced?
- Is the source just right for your research needs?

E: Evenness

- Does the author recognize other points of view?
- Is the information presented objective?
- If the source is biased, does the author acknowledge the bias?

C: Coverage

- Is the information new? Does it support what you have found in other sources?
- Is the source comprehensive or inclusive enough for your needs?
- Does this source provide information that is relevant to your needs?

T: Timeliness

- When was the source published?
- Is the date appropriate for your topic?

E. [RESIST Counter-disinformation Toolkit \(civilservice.gov.uk\)](https://civilservice.gov.uk)

Table 1. Pool of PURPOSE Questions
(from Hameleers' article)

Michael Hameleers, Disinformation as a context-bound phenomenon: toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination, Communication Theory, Volume 33, Issue 1, February 2023, Pages 1–10, <https://doi.org/10.1093/ct/qtac021>

Table 1.

Distinguishing between Intentions of Disinformation Campaigns and Suggestions on How to Detect Them Embedded in Their Communicative Context

Intentions	Detection
De-legitimization (sow discord, increase polarized divides, fuel cynicism, harm foreign countries, mal-information warfare)	- Look at broader influence campaigns: Who is harmed for what reason? - Identify which actors aim to sow discord - Reveal political background of source and compare this with targeted recipients - Focus on key moments of influence and peaks in disseminator's behaviors - Reveal (micro-)targeting strategies - Dissect techniques of manipulation
Mobilization (increase support or legitimize for (national) political agendas, consolidate power imbalances/silence the opposition, manage (inter)national image)	- Look at broader influence campaigns: Who seeks profit for what reason? - Identify which actors are seeking support - Reveal political background of source and match this with targeted recipients - Focus on key moments of influence and peaks in disseminator's behaviors - Reveal (micro-)targeting strategies - Dissect techniques of manipulation
Ideological motives (create support for alternative ideologies/share personal perspective on truth)	- Distill the ideological message from influence campaigns - Analyze the ways in which the ideological narrative is congruent with senders and targeted receivers - Relate construction of truth claims to conventional truth claims
Financial gain (increase advertising revenues, maximize profits, 'greenwashing' corporate image)	- Reveal the advertising strategies and sponsoring of influence campaigns - Assess the potential financial profits of deceptive information: What can the actor gain by creating or disseminating falsehoods? - Reveal the profile of the actor disseminating deceptive information and their overall corporate interests

POOL of STANDARDS and WARNINGS

GOOD SIGNS

- Credible media outlets have defined standards: a news piece must have more than one source and therefore, a comment offering an alternative view should be included.
- Credible media keep one-sided opinions where they belong: in the op-eds.
- A well-respected journalist always has a track record and is signed under a news piece
- An expert has professional and/or academic credentials in the field or topic they speak.

WARNING SIGNS (use them as WARNINGS or you can convert them into QUESTIONS)

- Random social media accounts do not count as a reliable source, they might be trolls or impostors.
- An "expert", specializing for instance in the Deep State and ancient aliens, and foreign policy should make you cautious.
- Think tanks that promise to spread progressive ideas and critical thought might be just good old pro-Kremlin proxies.
- If a story uses only anonymous sources or no sources at all, it should make you cautious.
- If the author has language or syntax troubles and problems engaging in a real conversation on social media, you should be cautious.
- Beware of bots! If you see posts from very active profiles on social media (who posts 200 times a day on Twitter?), you should be suspicious. According to a recent study, bots created 46% of Russian-language messaging about the NATO presence in the Baltics and Poland. The main topic of the tweets were mishaps that happened during NATO exercises.

ANNEX 5: DIALOGUES EXAMPLES FOR CT DIMENSIONS

In the following to facilitate the production of new dialogues some examples have been provided, one for each CT dimensions, in addition each dimension will take into consideration one of the dimension- sub skill, as indicated in the previous paragraph and figure.

Dimension 1: Interpretation

Sub-skill: Categorization

TITAN chatbot: Hi there! Let's explore how you categorize information. How do you usually classify the news articles you read?

Citizen: I usually categorize them based on the topic, like politics, health, or entertainment.

TITAN chatbot: That's a good start. Can you think of a time when categorizing information helped you understand it better?

Citizen: Yes, when I read about health, I always look for scientific studies to categorize it as reliable or not.

TITAN chatbot: Great! Categorizing helps in comprehension. Do you think this method can be applied to other areas of your life?

Citizen: Probably. It makes things clearer and easier to understand.

TITAN chatbot: Exactly. Developing strong categorization skills can enhance your ability to interpret and understand complex information.

Dimension 2: Analysis

Sub-skill: Examining Ideas

TITAN chatbot: Hello! Today, let's examine how we analyze ideas. When you encounter a new idea, what steps do you take to understand it?

Citizen: I usually read about it from different sources and see what experts say.

TITAN chatbot: That's a good approach. How do you compare or contrast these ideas to form your own opinion?

Citizen: I look for similarities and differences in the arguments and see which one makes more sense to me.

TITAN chatbot: Excellent! Comparing and contrasting ideas is crucial for deep analysis. Can you recall a situation where this helped you understand a complex issue?

Citizen: Yes, it helped me understand different perspectives on climate change.

TITAN chatbot: Perfect. Keep practicing this skill to enhance your analytical abilities.

Dimension 3: Evaluation

Sub-skill: Assessing Claims

TITAN chatbot: Hi! Let's talk about evaluating claims. When you hear a claim, what factors do you consider to assess its credibility?

Citizen: I consider the source of the claim and whether there is evidence to support it.

TITAN chatbot: That's a strong start. What makes a source credible to you?

Citizen: A credible source is usually an expert in the field or a reputable organization.

TITAN chatbot: Exactly. Assessing the source is vital. Can you think of a recent claim you evaluated and how you determined its credibility?

Citizen: I recently read about a new diet trend, and I checked if the claims were backed by scientific research.

TITAN chatbot: That's a great example. Evaluating claims thoroughly helps in making informed decisions.

Dimension 4: Inference

Sub-skill: Drawing Conclusions

TITAN chatbot: Greetings! Let's focus on drawing conclusions. When presented with multiple pieces of information, how do you decide what conclusion to draw?

Citizen: I look at the evidence and see which conclusion is most supported by the data.

TITAN chatbot: Good strategy. What if the evidence is conflicting? How do you handle that?

Citizen: I try to understand why the evidence conflicts and which sources are more reliable.

TITAN chatbot: Excellent approach. Can you provide an example of a conclusion you drew from conflicting evidence?

Citizen: I read different studies on the effects of coffee and concluded that moderate consumption is generally safe based on the majority of credible sources.

TITAN chatbot: Great job! Drawing well-supported conclusions is a key inference skill.

Dimension 5: Explanation

Sub-skill: Presenting Arguments

TITAN chatbot: Hello! Today, let's discuss how we present arguments. How do you usually present your reasoning when explaining something to others?

Citizen: I start with my main point and then provide the evidence that supports it.

TITAN chatbot: That's a solid approach. How do you handle objections or counterarguments?

Citizen: I try to address them by showing why my evidence is stronger or more relevant.

TITAN chatbot: Excellent. Addressing counterarguments strengthens your position. Can you recall a time when you successfully presented an argument?

Citizen: Yes, I explained to a friend why renewable energy is important, using data on its environmental benefits and cost savings.

TITAN chatbot: That's a great example. Clear and well-supported arguments are crucial for effective communication.

ANNEX 6: DIALOGUES EXAMPLES FOR SELF REGULATION DIMENSION

The utilization of the self-regulation grid for managing online misinformation presents a transformative approach to understanding and enhancing how individuals interact with information in the digital era. This grid, with its nuanced axes of self-confidence in dealing with misinformation (rasm) and the importance given to truthfulness (value_truth), provides a detailed framework that captures the complexities of individual behaviors and attitudes towards information. As we delve into the practical application of this grid, we can see its potential to revolutionize our strategies for education, intervention, and the promotion of a more informed and discerning public. Imagine a vast digital landscape where countless pieces of information compete for attention. Within this landscape, individuals navigate based on their confidence and their commitment to truth. The grid helps us map out these navigational tendencies, identifying where individuals fall within four distinct types: passive indifferent, confident indifferent, passive committed, and confident committed. Each quadrant represents a unique combination of self-confidence and value placed on truth, offering a lens through which we can better understand and influence user behavior.

Consider the journey of John, a representative of the **Type 1 (passive indifferent)** quadrant. John scrolls through social media with little scrutiny, sharing content that resonates with his interests without verifying its accuracy. He lacks both the confidence and the motivation to challenge the information he encounters. By utilizing the grid, we can identify John's specific needs and tailor interventions that aim to build his self-confidence and underscore the importance of truthfulness. Educational campaigns that highlight the impact of misinformation on society, coupled with tools that simplify fact-checking, can gradually transform John's passive indifference into a more engaged and conscientious approach.

In contrast, Maria, who falls into the **Type 2 (confident indifferent)** category, exhibits high confidence in her ability to spot misinformation but remains largely indifferent to the value of truthfulness. Maria's behavior might involve selectively verifying information that aligns with her biases, disregarding the broader implications of spreading falsehoods. The grid allows us to design targeted strategies for Maria that emphasize the ethical and societal benefits of truthfulness. Engaging Maria in discussions about the consequences of misinformation and highlighting stories where truth made a significant difference can nurture a more holistic appreciation of accuracy, complementing her existing confidence.

For individuals like Alex, who occupy the **Type 3 (passive committed)** quadrant, the challenge lies in their lack of confidence despite a high commitment to truth. Alex values accurate information but feels overwhelmed and unsure about how to verify it. Here, the grid directs us to provide Alex with empowerment through practical tools and resources. Workshops on digital literacy, step-by-step guides on fact-checking, and easy-to-use verification tools can bolster Alex's confidence, enabling him to act on his commitment to truth more effectively.

Finally, we have Emma, a quintessential **Type 4 (confident committed)** individual. Emma's high confidence and strong commitment to truthfulness make her a potent ally in the fight against misinformation. The grid's utilization in Emma's case focuses on sustaining and amplifying her impact. Encouraging Emma to take on mentorship roles, participate in community-driven fact-checking initiatives, or even lead educational programs can harness her abilities for the greater good. Emma's journey exemplifies how the grid not only helps identify areas for improvement but also recognizes and leverages existing strengths.

The true power of this grid lies in its dynamic application. As individuals progress through their interactions with misinformation, the grid facilitates continuous reassessment and adaptation of strategies. Feedback loops ensure that interventions remain relevant and effective, fostering an environment of ongoing learning

and improvement. This iterative approach transforms the grid from a static tool into a living framework that evolves with the digital landscape.

Moreover, the grid's narrative extends beyond individual interventions to encompass community and societal impacts. By collectively raising awareness and enhancing skills across different quadrants, we can cultivate a culture that values truth and integrity. This cultural shift, driven by the targeted and personalized strategies derived from the grid, has the potential to significantly mitigate the spread of misinformation and strengthen public trust in information.

The utilization of the self-regulation grid for managing online misinformation is a profound advancement in our approach to digital literacy and critical thinking. It allows us to understand the diverse ways individuals engage with information and provides tailored pathways for enhancing their skills and attitudes. Through targeted interventions, continuous adaptation, and a focus on both individual and community growth, the grid offers a holistic solution to the challenges posed by misinformation, paving the way for a more informed, critical, and truth-oriented society.

EXAMPLES OF PARAMETRIC DIALOGUES

The following presents first, a dialogue dedicated to **Type 2** and then a dialogue dedicated to **Type 3**.

For **Type 1**, which refers to those who have a low score on both axes of self-regulation, the two dialogues should be presented together in sequence and/or in a combined manner.

For **Type 4**, there is no need for such dialogues because the person already possesses the self-regulation characteristics necessary to handle misinformation. This, of course, does not mean that they do not need dialogues at all; they might need them, for example, on dialogues focused on specific items or on another dimension of Critical Thinking (namely, on one of the first 5 dimensions described above).

MODULE OF DIALOGUE MODULE DEDICATED TO BLOCK 4 - TYPO 2 OF THE GRID

Titan chatbot: Now, please, let me start with a simple question: How would it feel to pause and think for a moment more, every time you encounter a news story before deciding whether to share it or like it?

Citizen: I guess it would make me a bit slower in responding.

Titan chatbot: That's a fair point. Try to think: what might change in your interactions online if you took that extra moment?

Citizen: Probably, I'd make fewer mistakes in what I share; maybe fewer arguments with friends.

Titan chatbot: Interesting. Can you recall a time when you wished you had thought more before sharing something online?

Citizen: Yes, one time I shared a news that turned out to be false; it was embarrassing.

Titan chatbot: That sounds frustrating. What would you do differently now with the right strategies?

Citizen: I'd probably check the facts first, maybe see if reliable sites are also reporting it.

Titan chatbot: That's a great strategy. If you felt more in control of how you respond to news stories, how might this influence your credibility and the trust others place in you?

Citizen: I think people would take my words more seriously; they might even look up to me for reliable information.

Titan chatbot: I really do believe you can stop, think, and check for yourself now. Feeling more capable and in control can truly transform how we interact online. Let's think about some steps you might take to verify information before sharing it. What are some actions you could start doing today?

Citizen: Maybe I could start by checking a few trusted news sources or use fact-checking websites?

Titan chatbot: Starting even with small steps can make a big difference!

MODULE OF DIALOGUE DEDICATED TO BLOCK 4 – TYPO 3 OF THE GRID

Titan chatbot: Now, let's shift our focus slightly. What do you think are the consequences of spreading unverified news?

Citizen: It could cause unnecessary panic or misinformation.

Titan chatbot: ...and...how might it affect individuals or communities?

Citizen: Misinformation could lead to wrong decisions, harm people's reputations, or even affect public safety.

Titan chatbot: That's interesting. In which ways do you feel responsible for the accuracy of the information you share online?

Citizen: I guess since my sharing can influence others, I have some responsibility to make sure it's correct.

Titan chatbot: And speaking about yourself again, how does ensuring the accuracy of what you share reflect on you?

Citizen: It shows I'm reliable and trustworthy. People might take what I say more seriously.

Titan chatbot: Let me help you focus: can you think of a time when misleading information had real-world impacts?

Citizen: There was a rumour about a health scare that created a lot of chaos until it was debunked.

Titan chatbot: Thank you for sharing your thoughts. What might be the long-term effects of such incidents on public trust?

Citizen: People might become more skeptical or lose trust in certain sources or platforms.

Titan chatbot: Again, how important is truthfulness in maintaining meaningful conversations and trust in a community? What do you believe?

Citizen: It's very important. Without trust, real, meaningful communication breaks down.

Titan chatbot: And what role does each of us play in this, in your opinion?

Citizen: We all need to verify information before sharing and encourage others to do the same.

Titan chatbot: If I asked you what could be done to enhance the role of online truth in the information we come into contact with, what could you tell me? What would you think?

Citizen: Perhaps learning how to quickly verify facts and educating friends and family about the importance of doing so.

Titan chatbot: Thank you for your thoughtful responses; they're crucial for building a truthful online community.