Fake news dissemination (Disinformation), an excessive amount of information (Infodemia), manipulation and influence of public opinion (Propaganda) are phenomena that contribute to the concept of Cognitive Warfare: activities aimed at influencing people’s thoughts and actions. The primary threat associated with Cognitive Warfare is the discrediting and delegitimation of media and institutions. Moreover, it reduces the chances of having rational discussions on genuine, shared objective facts and non-influenced perspectives.

The Spoke will establish an excellent multidisciplinary structure that, leveraging intelligence analysis, artificial intelligence, political analysis, data science, and web intelligence capabilities, employs suitable tools and methods to deal with these information disorder issues.

Spoke 2 is coordinated by UNISA and brings together several complementary initiatives to address the thematic line in its overall complexity. It relies on the implementation of the following project scopes (i.e., Ambiti Progettuali):

- **IDA** - Information Disorder Awareness
- **FF4ALL** - Fighting Fakes: Detection of Deep Fake Media and Life-Long Media Authentication
- **DETERRENCE** - DEcision supporT SystEm foR cybeR intelligENCE
- **HUMANE** - Holistic sUpports against inforMATioN disorder

**IDA** focuses on the Monitoring of the overall phenomenon. It is aimed at profiling news sources and detecting vulnerabilities and threats related to Information Disorder. Deep learning will be applied to analyze multimedia content to evaluate the impact and identify deepfakes, adopting active authentication approaches defined by **FF4ALL, DETERRENCE** mainly focuses on the social dimension of the problem. **HUMANE** focuses on fake content and the reaction to its spreading. The synergy of these projects will enable the detection and understanding of possible radicalization path consequences and the definition of bias mitigation measures.
Spoke 2 launches Open Calls / Bando a Cascata to address certain tasks foreseen in each of the aforementioned project scopes. For each project scope, this document introduces the corresponding set of tasks that have to be managed by the participants with their proposals.

The document is organized as follows. It first outlines the overall plan of the activities and milestones of Spoke 2. Then, it provides useful details about the projects included in Spoke 2. Finally, the section "Open Calls / Bando a Cascata" details the objectives of the tasks of each project that are the subject of this notice.
Plan of activities and milestones

Extended Partnership started activities on 1st Jan 2023 (M1). The duration of the project is 36 months.

The overall plan of the activities and milestones of the Spoke is summarized in the diagram shown in Figure 1. The figure details each type of activity, separated by horizontal bars, to which the partner exposes the costs of the project. The figure also displays the checkpoints at which the Spoke leader and partners must summarize the findings obtained in the corresponding period as vertical red lines. Analogously, after selecting the most-suited proposal for the Open Call corresponding to each project, the winning candidate should provide technical reports, one at each checkpoint in the diagram (red lines). The technical report will describe the findings obtained in the corresponding period and a software implementation of the best-performing techniques and best-suited method.

In addition, the winning candidate should monthly provide an update on the activities carried out on the ongoing Open Call.

Milestones are set at the end of each period, when the results obtained in the last considered ones will be revised, and the work to be done in the subsequent period will be planned in light of the results obtained so far.

Figure 1. Gantt diagram.
Abstract

Context: Social media play a crucial role in information disorder. In this context, colluding users carry out coordinated activities (also using automation) to spread malicious content for manipulating others. The consequences of such manipulations are polarization, radicalization and misinformation, which contribute to a climate of distrust and negatively affect public opinions, other than undermining online participation.

Objective: The main objective of the DETERRENCE project is to study the Information Disorder phenomenon on social media with the aim of designing a proof of concept of decision support tools (DSS) to monitor and mitigate its impact, both for individuals and for society in general. Principal project activities will be devoted to: (i) detect and investigate, through network science methods, coordinated online behaviors especially on large-scale campaigns also identifying automated behavior; (ii) develop techniques for detecting the next generation of fake accounts and content, including malicious accounts, as well as deepfake texts and multimedia content; (iii) investigate the dynamics of communities and social networks that can be potentially exposed to cognitive bias which in turn could cause or amplify noxious phenomena such as gender discrimination, racism, cyberbullying and so on.

Work Breakdown Structure

WP1 - Network science for analyzing coordination and polarization

WP Description

Large-scale online campaigns, malicious or otherwise, require significant coordination among participants, which is needed to spread content widely and to let the campaigns obtain significant outreach, ultimately ensuring their success. For example, coordination among users emerges as part of many authentic and harmless activities, such as grassroots initiatives, online protests, and the activities of fandoms. At the same time, however, malicious campaigns use harmful coordination, such as those responsible for the spread of disinformation, influence operations, and coordinated harassment. The pervasiveness of coordination as part of online activities has sparked wide interest in the study of coordinated online behavior. This WP will design and experiment with network science methods for detecting and investigating coordinated online behaviors.
Furthermore, such networks will be leveraged to measure the degree of polarization among users and to study the complex interplay between coordination and polarization.

- Task 1.1: Strategies to detect and measure polarization and coordinated behavior (IIT-CNR: 15 PMs; ICAR-CNR: 2 PMs; RTDs: 18 PMs)

- Task 1.2: Strategies to mitigate polarization and coordinated behavior. (see the section Open Calls)

WP2 - Machine learning and Deep learning for analyzing malicious users and content

WP Description

In recent years, many organized attempts to spread deceptive content have been carried out on social media to mislead and manipulate. These malicious campaigns can take different shapes, target various individuals, online crowds, or communities, and have diverse goals. Among the strategic tools used by perpetrators are fake news, propaganda, hateful speech, colluding users (e.g., paid trolls), and automation (e.g., social bots). The ongoing endeavors to contrast such manipulations led to a vast body of work on these issues and a plethora of different solutions. However, despite the efforts, researchers debate the efficacy of the proposed solutions. Meanwhile, groundbreaking advances in artificial intelligence brought to the rise of deepfakes that allow crafting arbitrary texts resembling a target person’s writing style, or audio and video samples where a target person’s face and voice could be made to do or say anything. These techniques have already been used to create fake news and fake profiles of fraudulent accounts. This WP will develop techniques for detecting the next generation of fake accounts and content, including malicious accounts colluding and coordinating as part of large information operations, as well as deepfake texts and multimedia content.

- Task 2.1 - Tools and techniques for detecting malicious automated accounts (IIT-CNR: 3 PMs; ISTI-CNR: 3 PMs; ICAR-CNR: 2)

- Task 2.2 - Tools and techniques to detect anomalous and fake multimedia (IIT-CNR: 3 PMs; ISTI-CNR: 6 PMs; RTDs: 4 PMs)

- Task 2.3 - Detection of online propaganda (see the section Open Calls).

WP3 - Radicalization paths and mitigation measures

WP Description:

Social media platforms have substantially altered the landscape of societal debates. By delivering an extremely large amount of content to online users, they enable quick and easy access to
information and facilitate participation in public debates. This positive effect is intertwined with the growing phenomenon that online political discourses, especially on socially relevant issues, tend to fragment and polarize opinions. As a result, the propagation of information is affected by users’ propensity to select and promote claims that adhere to their beliefs and ignore or even contrast dissenting information. In this WP, the dynamics of communities and social networks that can be potentially exposed to cognitive bias will be investigated. This bias may amplify noxious phenomena such as gender discrimination, racism, cyber bullying, social inequalities, content and thinking homogeneity, spread of fake news, radicalization, by trapping users into filter bubbles, rabbit holes or echo chambers that extremize their need for ideological confirmation.

- Task 3.1 - Understanding cognitive bias (ICAR-CNR: 8 PMs; RTDs: 6PMs)
- Task 3.2 - Characterizing radicalization pathways (ICAR-CNR: 8 PMs; IIT-CNR: 2 PMs; RTDs: 8 PMs)
FF4ALL

Detection of Deep Fake Media and Life-Long Media Authentication

Abstract

The creation-diffusion of fake multimedia content, whether entirely computer-synthesized or obtained by manipulating original content, can be considered a danger to many aspects of our society. The overall phenomenon has reached unprecedented levels, thanks to the availability of artificial intelligence tools dedicated to generating fake content almost indistinguishable from real ones in a relatively simple way. Fighting so-called deepfakes, videos, images and audio clips, generated with deep learning techniques, requires the development of appropriate countermeasures. The project aims to develop theoretical and practical tools to detect multimedia fakes or counterfeited contents, trace back to their origin, and limit their diffusion, through passive analysis techniques operating when the content is used or diffused, and active protection methods to be adopted at the time of content creation, to facilitate subsequent authentication.

Work Breakdown Structure

WP1 - Deepfake Attribution and Recognition

WP Description

Detecting falsified content is often not enough to limit the diffusion of fake media and prosecute the parties who created and spread the fake contents. In many cases it is also necessary to identify the origin of the fake content and possibly trace back to its history. To this aim, identifying the tools, in particular the network architectures and the specific models, which were used to create the fake content assumes a central role. It is the goal of this WP to identify the traces, a.k.a. fingerprints, left within the deep fakes by the models which generated them, and use such fingerprints to i) distinguish deepfakes from genuine content, ii) attribute the fake content to a specific network architecture, iii) attribute the fake content to a specific model.

- Task 1.1 - Deepfake “Fingerprint” Modeling (see the section Open Calls).
- Task 1.2 - Deepfake Attribution (see the section Open Calls).
WP2 - Passive Deepfake Authentication Methods and content

WP Description

In this project, "deepfake detection" is referred to as "passive deepfake authentication" to include all three topics: deepfake fingerprinting, media authentication, and marking. The focus of this work package is on verifying the authenticity of content and preventing malicious deep learning-based processing. This is important because fake content poses a risk to security applications and the spread of fake news. However, media compression, size, and resolution present challenges to the authentication method's robustness. The work package aims to explore the problem from different angles, including biometric applications, multimodal detection, and advanced techniques using context and semantic content.

- Task 2.1 - Deepfake and Biometric Recognition (UNICA)
- Task 2.2 - Audio-Video Deepfake (see the section Open Calls).
- Task 2.3 - Advanced Methods for Deepfake Detection (see the section Open Calls).

WP3 - Deepfakes Detection Methods in Realistic Scenarios

WP Description:

The detection of deepfakes has garnered significant interest in recent years, resulting in several proposed solutions to address the growing issue of fake media content. However, most of these solutions only perform well in controlled settings, such as laboratory experiments, and fail to provide reliable results in real-world scenarios. This WP aims to go beyond the research conducted in earlier work packages and develop solutions that work effectively in highly realistic environments. The package comprises three tasks, which address specific challenges encountered in real-world applications, including dealing with limited data, open set conditions, interpretability requirements, working in social media settings, and adversarial settings where an informed adversary aims to defeat the detection tools.

- Task 3.1 - Deepfake Detection of image-videos in the wild (Sapienza, Irene Amerini)
- Task 3.2 - Deepfake and Social Media (see the section Open Calls).
- Task 3.3 - Detection of Deepfake Images and Videos in Adversarial Setting (see the section Open Calls).
WP4 - Active Authentication

WP Description:

Active media authentication techniques work in a preemptive way to ease subsequent analysis, while passive methods work after the forged content has been generated. Deepfake detection methods based on DNN watermarking and unique fingerprints inserted within the content are examples of active techniques. Blockchain can be used to trace the processing chain of images and videos. WP4 aims to study active authentication techniques as a more reliable alternative or complement to passive methods, where operating conditions permit.

- Task 4.1 - Active Fingerprinting for Deepfake Detection and Attribution (CNIT, Roberto Caldelli)
- Task 4.2 - Authentication of Devices for the Acquisition and Processing of Content (see the section Open Calls).
- Task 4.3 - Trusted Remote Media Processing on Cloud and Edge Computing Systems (see the section Open Calls).
Abstract

The problem of information pollution on social and traditional media has reached a global scale, and its impact is challenging to quantify. Information Disorder is a new term that holistically considers all the possible ways information can be manipulated (not only to harm). HUMANE aims to design, develop and test a publicly available toolkit to tackle the problem of information disorder comprehensively. HUMANE will be based on three pillars: 1) Analyze, 2) Detect, and 3) React. The "Analyze" pillar aims to understand the causes and reveal the mechanisms with which the information gets polluted. Effectively mapping the phenomena and understanding the main effects on Society will lead to success. The “Detect” pillar aims at developing, using AI technologies, a set of tools to automate the discovery of potential problems and threats. Fact-checking, verification tools, and “authenticity engines” will be built to increase awareness and trust in the information people see online. Success will be measured by analyzing the performance of AI algorithms on a set of data collected and explicitly labeled in this pillar. "React" will be responsible for finding mitigation strategies by minimizing the impact of echo chambers: feed and search custom algorithms, diversify exposure to different people and views, and allow users to consume information privately. The success of this pillar will be evaluated by measuring the effect of these mitigation strategies on controlled communities on social media platforms. HUMANE users will be able to shed light on the causes and mechanisms that lead to information pollution; they will benefit from a non-polarized information ecosystem; they will benefit from tested solutions to mitigate information disorder phenomena. The research areas involved in this project will be Social and Web Analytics, Cybersecurity, Complex Systems, Data Science, AI, IR, NLP, Human Computation, and Social and Political Sciences.

Work Breakdown Structure

WP1 - Analyze

WP Description

The "Analyze" work package of the project aims to understand the causes and mechanisms of information pollution and disorder. The process will involve a literature review, data collection, and analysis, mapping of the phenomena, understanding the effects on society, development of hypotheses, and regular reporting. The collected data will be analyzed using various tools and
techniques. The results will be used to create visual representations of the information pollution landscape and understand its effects on society. The results will guide the development of detection and mitigation tools in the later stages of the project.

- Task 1.1 - Disinformation Risk Assessment
- Task 1.2 - Sources of Information Disorder
- Task 1.3 - Cybersecurity, Forensics, and Measurement Approaches (see the section Open Calls).

WP2 - Detect

WP Description

The "Detect" Work Package focuses on the development of AI technologies to automate the discovery of potential problems and threats in the information ecosystem. The aim is to build fact-checking and verification tools, as well as "authenticity engines," to increase awareness and trust in the information people see online. Success in this WP will be measured by analyzing the performance of AI algorithms on a set of explicitly labeled data.

- Task 2.1 - AI (ML-DL) for Detection of Information Disorder (UNIROMA1, UNIVE, CNR, IMT)
- Task 2.2 - Detect Coordinated inauthentic behavior (CNR, UNIROMA1, IMT)
- Task 2.3 - Decision-theoretic, Human-in-the-loop, and Hybrid Approaches (see the section Open Calls).

WP3 - React

WP Description:

The "React" work package has the goal of finding mitigation strategies to minimize the impact of information disorder (ID) by developing and implementing tools that can help reduce the spread of misinformation and increase awareness about it. This WP aims to address the problem from a legal, social, and technological perspective and to provide the necessary tools to help users, communities, and organizations to proactively protect themselves against ID threats. The success of this WP will be evaluated by measuring the impact of these mitigation strategies on controlled communities on social media platforms.

- Task 3.1 - Legal-Social Implications (IMT, UNIMI)
- Task 3.2 - Proactive Defense Against ID Threats (CNR, UNIROMA1, IMT)
- Task 3.3 - Tools Development-Integration (see the section Open Calls).
Abstract

The excessive amount of information (Infodemia), the dissemination of fake news (disinformation), and the manipulation and influence of public opinion (Propaganda) are phenomena that contribute to defining the concept of Cognitive Warfare (i.e., the art of using technological tools to alter the cognition of human targets). In this regard, Information Disorder may lead to the delegitimation of professional, reliable media, authoritative institutions, and objective data, ending up with fewer chances for people to become aware of objective facts and talk about them. Then, there is a need for solutions to identify content affected by disinformation and perform early monitoring of social, business, and political threats and vulnerabilities. To achieve these goals, knowledge discovery techniques should be applied to check the content and source truthiness, relating multiple aspects, including source authenticity and news credibility.

The effective novelty depending on the web and social media spread is the high speed at which news is spread. As a result, there is less time for people to build a proper opinion and for fact-checking and source reliability checking. Additionally, customers seem to be not yet interested in checking news source reliability. As a consequence of this trend, news content creators do not pay attention to giving their content good quality but attract customers and leading them to share the same feelings. Moreover, any institution attempting to suppress, mitigate or control information worsens, as an effect, its reputation with people and provides support to those who exploit disagreement to their own advantage. Then, it seems more convenient to analyze social media content to detect, monitor, predict and trace discomforts that people are suffering in real-time since it provides a way to take action in order to avoid the exploitation and manipulation of vulnerabilities. Social media analysis is meant to be as successful as the analogous methodology employed in cybersecurity to fight zero-day (-hour) threats. To achieve this, the smartest strategy may be to increase the number of non-profit organizations and equip them with emerging technologies. At the same time, another successful strategy is to invest in solutions to increase content customers’ awareness about the vulnerabilities they may be exposed to, promoting information campaigns on risks of disinformation about social issues inspired by similar events occurring by the time of the early mass-media diffusion and early advertising campaigns. Opening this “digital oracle” towards external systems may lead to plug-in tools capable of acting on fake news in advance, preventing their diffusion on the Web.

The main objective of this proposal is to build a multidisciplinary center of excellence that bridges the capabilities of different areas, including intelligence analysis, political sciences and geopolitics,
data science, artificial intelligence, etc. The center should use innovative tools and methods to fight disinformation, allowing institutions, and specifically, those working on security, to fight with proper weapons the battle against the threats to social cohesion and the above-mentioned Cognitive Warfare. The add-on to this system will be focused on providing different interaction modes to communicate with the outer world and allow ad hoc-created methods to query the system for news and information validation. The center is aimed at building a hub to handle processes and services related to disinformation, and along with private and public partners, is aimed at fighting disinformation attacks that mine people’s trust in institutions and the principles of a democratic society. The idea consists of setting up an infrastructure giving experts (e.g., policymakers) tools for monitoring the Web or specific networks, identifying vulnerabilities, and making suitable decisions. The produced framework, through a human-in-the-loop approach, will give experts the possibility to identify research weaknesses and research findings useful in terms of media literacy. This way, experts, acting like intelligence analysts, will exploit the framework to investigate a target and select suitable strategies by, for example, recognizing consumers’ vulnerability to manage in terms of new research finding needs or media literacy activities. The common strategy will be implemented thanks to synergy and coordination of functional areas devoted to specific missions required to fight disinformation. They include Multidisciplinary Research, Operational Monitoring, and Policy & Education. Multidisciplinary Research will be involved in the definition and implementation of research results that could be exploited for monitoring Information Disorder online. Operational Monitoring exploits the research results for investigating Information Disorder and provides awareness through the comprehension of the phenomena and the projection of the impact in terms of interactive reports about relevant topics. Policy and Education will carry out the lesson learned through these investigations by deriving guidelines and recommendations useful to face and mitigate its impact.

Work Breakdown Structure

WP1 - Methodologies for Reliability of Sources and Contents

An important aspect to check information truthiness is news sources analysis. Social profiles, web pages, web domains, as well as e-mail addresses hide clear information, which can be detected by humans without automatic tools sometimes. Moreover, a second layer of the analysis is necessary to check content integrity. A web page, an individual post or even an image including a visible message (es. “meme”) should be checked and compared by using several parameters and digital techniques. Then, this WP is aimed at exploring techniques to check the reliability of sources and contents as a primary step to detect disinformation on the Web.

○ Task 1.1 Assessment of Reliability of Sources (UNISA)
Task 1.2 Semantic Integrity of Contents (UNISA)

Task 1.3 Digital Integrity of Multimedia Contents (UNISA)

WP2 - Methodologies for Monitoring and Mitigating Information Disorder Vulnerabilities

To build predictive analysis tools targeted at informing analysts on specific disinformation attacks, news content is not enough. Therefore, there is a need for analyzing eventual vulnerabilities related to social communities and/or groups whose users are related to the same disadvantaged condition. Vulnerability analysis alongside news spread and network analyses are necessary to identify, track and keep under monitoring eventual disorders. To prevent or mitigate information disorder events, there is a need for models and measures to quantify risks and impact according to vulnerabilities encountered.

Task 2.1 Identifying, Tracking and Monitoring Information Disorder Vulnerabilities (UNISA)

Task 2.2 Measuring Impact and Risks of Information Disorder (see the section Open Calls)

WP3 - Information Disorder Awareness Framework

IDA proposal is aimed at defining an architecture offering microservices exploiting an event-driven approach using scalable and big data-based technologies. The components prototyped in the WP1, and WP2 will be integrated and eventually orchestrated in order to implement more complex workflows of fact-checking and information disorder monitoring. The main functional goals of the overall architecture include data acquisition, processing, reliability checking and knowledge extraction bridging structured and unstructured sources. The outcome will be returned both for exploratory and demonstration purposes. The architecture goal is focused on providing expert analysts with a powerful toolkit to help them in the analysis of Information Disorder on the Web and Social Media. On the other hand, the architecture will be designed to define, build, and test ad-hoc services to satisfy the needs of heterogeneous stakeholders. For instance, it will provide policymakers and non-expert users with useful results that will be returned as specific services, including interest notification stream, brief data summary, statistics, etc.

Task 3.1 Design and Prototyping of the Framework (see the section Open Calls).

Task 3.2 Evaluation of the Information Disorder Awareness Framework (UNISA)

WP4 - Operational Monitoring, Policy and Education

This WP is targeted at building up the Italian disinformation monitoring center by exploiting the synergy of the three functional areas: Multidisciplinary Research Area, Operational Monitoring Area and Policy and Education Area. The first steps of the center will include a context analysis and
elicitation of challenging requirements corresponding to different disinformation cases and exploitation scenarios. Then, advanced tasks will exploit the results of the earlier WPs for investigating the Information Disorder concerning topics periodically identified by the experts. The outcomes of these tasks consist of interactive reports and media literacy tools to make experts and non-experts aware of vulnerabilities and risks they would take in the presence of different cases and forms of disinformation.

- Task 4.1 Italian disinformation monitoring center (UNISA)
- Task 4.2 Media literacy tool development and sharing (UNISA)
- Task 4.3 Information Disorder Policy Management and Policy Recommendation (see the section Open Calls).

![Figure 2. Work packages interaction diagram and corresponding Open Call of the project IDA.](image-url)
Open Calls / Bando a Cascata

For each project scope (i.e., Ambito Progettuale) included in Spoke 2, the sections below provide a detailed breakdown of the tasks and their corresponding main objectives. The selected proposal must successfully complete the tasks outlined below, ensuring that the requirements and objectives of the project milestones are met.

Each proposal should focus on addressing a specific project scope. In the case of more than one partner participating in the same proposal for addressing the same project scope, each of them must clearly state their role, expected outcomes, and corresponding budget. Additionally, the proposal has to plan the activities over time by producing a Gantt chart including milestones in accordance with the overall project Gantt reported in Figure 1 and respect the deadlines for documentation and software deliveries.

Project Scope: DETERRENCE

WP1 Task 1.2 - Strategies to mitigate polarization and coordinated behavior.

We will survey, evaluate, and experiment with multiple strategies (i.e., interventions) to reduce online misbehaviors and polarization. The efficacy of each intervention will be evaluated with adequate causal inference methods. Effects will be comprehensively considered at multiple scales: at the platform-, community-, and user-level. When assessing intervention effects, those characteristics will also be investigated that might influence user reactions to interventions (e.g., user’s personality, political leaning), thus providing new knowledge and tools for mitigating widespread issues in online platforms.

Main Objectives:

- Developing effective intervention strategies that can counteract online issues, promote diverse perspectives, encourage critical thinking, and foster constructive dialogue.
- Evaluating interventions and countermeasures effectiveness. This objective involves conducting rigorous experiments to assess the impact of interventions, determine their limitations, and refine them based on empirical evidence.

WP2 Task 2.3 - Detection of online propaganda.

Propaganda is conveyed through a series of rhetorical and psychological persuasion techniques. Machine learning approaches will devise to detect such techniques in online texts by combining a
learning system based on state-of-the-art language models with ad-hoc systems for the techniques that are harder to be identified by the general system. To determine how these techniques concur to influence an online user, the correlation of the use of the techniques in messages that a user has been exposed to with the online communities the user belongs to in specific time windows, as determined by dynamic analyses of online interactions and behaviors, will be studied.

**Main Objectives**

- Applying Machine learning and natural language processing algorithms and models to analyse large volumes of text, images, and multimedia content to identify propaganda techniques and propaganda networks
- Studying online interactions and behaviors aspects of propaganda consumption, exploring how users are influenced by propaganda content.

**Project Scope: FF4ALL**

**WP1 Task 1.1 - Deepfake “Fingerprint” Modeling**

Task 1.1 aims to identify unique fingerprints in deepfake images and videos generated by neural networks. The noise print extracted from them can reveal a signature that identifies the network that created it, but its robustness and universality need investigation. The selected team will develop techniques to distinguish between real and fake content using both model-based and data-driven methods. They will also consider extending the analysis to temporal signatures in video signals the goal of identifying unique markers that can aid in detecting and distinguishing between real and fake media content.

**WP1 Task 1.2 - Deepfake Attribution**

Task 1.2 aims to identify the specific network architecture or model used to generate fake content. This can be approached as either a classification or verification problem. In the classification scenario, the classifier must determine which among a predefined set of networks created the fake content. In open-set conditions, the classifier must include a rejection option to be activated when the content is generated by an unknown network. In the verification scenario, the verifier is given a fake content and a suspect network and must decide whether the content was generated by the suspect network. The selected team will consider both closed and open set verification scenarios and is expected to leverage the results provided by Task 1.1 regarding the identification of deepfake fingerprints.
WP2 Task 2.2 - Audio-Video Deepfake

To improve deepfake detection in videos, analyzing both the audio and visual tracks can expose inconsistencies between them and reveal fakes. In addition to lips and eyes motion, anomalies in emotional state conveyed by the audio and visual tracks, reverberation effects incompatible with the framed scene, lack of synchronization between video and audio and other semantic clues can be used to identify fakes. The selected team working on this task will investigate the effectiveness of audio-visual deepfake analysis, addressing challenges such as the need for suitable datasets, efficient forensic tools, and proper fusion techniques that merge audio and visual clues at various processing levels. This includes data-level, feature-level, score-level, and decision-level fusion.

WP2 Task 2.3 - Advanced Methods for Deepfake Detection

Current deepfake detection methods are vulnerable to low-level quality impairment actions. One solution is to use high-level semantic features, such as inconsistencies in individual biometric traits or facial movements, to build reliable models. Semantic-based analysis can also apply to non-facial content, using inconsistencies in motion patterns or shadows. Another solution is to analyze content within its context, which provides valuable priors to authenticate tools and disambiguate uncertain results. Contextual information can be obtained through metadata analysis or by analyzing the wider document or web page where the content is used. The selected team will use these methods to improve the interpretability of forensic analysis results.

WP3 Task 3.2 - Deepfake and Social Media

Forensic analysis of multimedia data in social media channels is complex, and a binary classification of "real" or "manipulated" may not be representative. The challenge is to develop more sophisticated authenticity indicators that capture different aspects of the object under investigation. The selected team will adapt general forensic tools to these needs and test them on data gathered from popular social media channels.

WP3 Task 3.3 - Detection of Deepfake Images and Videos in Adversarial Setting

Task 3.3 addresses the issue of adversarial attacks on deepfake authentication tools, which can be easily deceived when the adversary is informed about the details. The selected team will develop tools with enhanced security against intentional attacks, starting by defining security models that describe the framework in which the race between the forensic analyst and attacker is staged. Proposed solutions may include adversarial training, hybrid data-driven and model-based detectors, and multi-clue, multimodal analysis to robustify detectors against deliberate attacks during the training phase, considering the goals, constraints, and information available to both players.
WP4 Task 4.2 - Authentication of Devices for the Acquisition and Processing of Content

To prevent the spread of fake content, IoT device authentication is crucial for data acquisition and content processing. Traditional methods based on certificates issued by Certification Authorities (CAs) are obsolete and vulnerable to cyber-attacks, especially when IoT devices are installed in remote areas. New protocols are needed to provide additional security in communication systems between IoT devices and the Edge and Cloud systems. The selected team will develop and prototype decoupling systems and new protocols for communication between IoT devices, to create a reliable ecosystem for authenticating acquired and processed data.

WP4 Task 4.3 - Trusted Remote Media Processing on Cloud and Edge Computing Systems

Task 4.3 aims to address challenges related to media processing authentication in cloud and edge computing systems. The objective is to design secure strategies for multimedia data transmission, processing, and storage with robust authentication and verification mechanisms to prevent the spread of fake digital media. The elected team will integrate active and passive techniques into a scalable cloud/edge framework with technologies such as Blockchain and IPFS (InterPlanetary File System). They will also explore the use of AI and machine learning algorithms to improve accuracy and efficiency. Experiments and simulations will evaluate the practicality and effectiveness of the developed strategies in real-world scenarios.

Project Scope: HUMANE

WP1 Task 1.3 - Cybersecurity, Forensics, and Measurement Approaches

The final task in the "Analyze" work package, "Cybersecurity, Forensics, and Measurement Approaches," aims to develop and implement cybersecurity and forensics metrics to measure the impact of information disorder on society. This task will build on the knowledge gained from the "Sources of Information Disorder" task and the "Disinformation Risk Assessment" task and will use that information to design and implement methodologies for measuring the impact of disinformation, misinformation, and malinformation. The task will also assess the feasibility of using AI and NLP technologies to automate the measurement of information disorder. The results of this task will inform the design and development of mitigation strategies in the "React" pillar. They will provide a comprehensive understanding of the impact of information disorder on society. This task aims to provide a robust and reliable measurement framework that will help understand and mitigate the impact of information disorder effectively.
WP2 Task 2.3 - Decision-theoretic, Human-in-the-loop, and Hybrid Approaches

The Decision-theoretic, Human-in-the-loop, and Hybrid Approaches will explore the integration of human intelligence with AI algorithms to improve the accuracy and effectiveness of the tools developed in this WP. This task will focus on developing decision-theoretic, human-in-the-loop, and hybrid approaches to detect information disorders. The goal is to improve the accuracy and effectiveness of the tools by incorporating human judgment and feedback.

WP3 Task 3.3 - Tools Development-Integration

This task will focus on developing, integrating, and testing the tools that were developed in the Detect and React WPs. The goal is to provide a comprehensive and user-friendly solution that can be used to detect, mitigate, and react to information disorder threats in real-time.

Project Scope: IDA

WP2 Task 2.2 - Measuring Impact and Risks of Information Disorder

This task is aimed at investigating solutions to quantify the impact and risks of disinformation events. In this regard, metrics will be defined to measure the fake news impact in terms of vulnerabilities they may create among Internet users (Community members, news site readers). These metrics should consider various parameters over time, including the content spread in the network, its popularity among users and the emotional reactions they have on audiences. Changes in the values of these parameters are also to be monitored over time to effectively analyze the impact and risks of the shared content.

Main objectives:

More in detail, this task has to accomplish the following scientific objectives:

- Community detection by considering multiple aspects including network analysis, polarization degree assessment, echo chamber and filter bubbles detection.
- Content spreading models to analyze content spreading dynamics, contagion, and influence and information spreading on social networks by considering social interactions among social network users.
- Vulnerability and risk assessment indices assessed by means of the metrics defined for Community detection and Content spreading models to be compared with state-of-the-art indices.
WP3 Task 3.1 - Design and Prototyping of the Framework

This task focuses on the analysis of state-of-the-art about technologies and best practices to design a cutting-edge Information Disorder Awareness architecture. Its primary objective is the design and development of a full-stack platform adopting or extending the existing standards, such as interoperability languages (e.g., STIX, TAXII, etc.) and frameworks (e.g., DISARM, etc.) specifically used to align with prevailing information security (infosec) practices and tools. These standards and frameworks will assist in monitoring and countering disinformation and other forms of information harm. The result is the Information Disorder Awareness analytical platform integrating the outcomes of the previous WPs. Finally, a test plan will be devised to validate the outcomes in a representative setting.

Main objectives:

More in detail, this task has to accomplish the following scientific objectives:

- Analysis of existing reference frameworks and interoperability standards in the context of Information Disorder Awareness.
- Adoption and extension of existing frameworks in the reference context of Information Disorder Awareness.
- Implementation of the analytical platform for Information Disorder Awareness.

WP4 Task 4.3 - Information Disorder Policy Management and Policy Recommendation

In this task, the main aim is to involve different stakeholders (e.g., local and central public administrations, institutions, schools, universities, research centers, media, etc.) in arranging the outcomes produced by previous tasks for the definition of policy recommendations and activities to improve customers and public disinformation awareness. This task requires a thorough study of existing policies and the design of policy recommendations that exploit the analysis of vulnerabilities provided by the monitoring tools developed in the previous tasks to advise expert and non-expert users about the risks they are taking.

Main objectives:

More in detail, this task has to accomplish the following scientific objectives:

- Policy recommendation design: defining of regulation taking into consideration events, contingencies and domain in which disinformation incidents occur.
- Policy recommendation prototyping: designed regulation must be implemented and tested through pilot use cases.