Funded by the European Union
Horizon Europe
(HORIZON-CL4-2021-HUMAN-01-27
AI to fight disinformation)

AI4TRUST

D2.1

DESIGN OF THE METHODOLOGICAL TOOLBOX

PARTNERS

Document information

Title	D2.1 - Design of the methodological toolbox	
Editor	FBK	
Contributors	UNITRENTO, CNRS, MALDITA, DEMAGOG, ELLENIKA, EURACTIV, SKYTG24, ADB, EMD, UCAM, CERTH, NCSR, UPB, SAHER & GDI	
Dissemination Level	☐ CO: Confidential, only for members of the consortium (including the Commission Services) ☐ RE: Restricted to a group specified by the consortium (including the Commission Services) ☐ PP: Restricted to other programme participants (including the Commission Services) ☐ **PU: Public**	
Reviewers	☐ FBK ☐ CERTH ☐ UNITN ☐ NCSR ☐ CNRS ☐ UPB ☐ SAHER ☐ GDI	☐ DEMAGOG ☐ MALDITA ☐ ELLENIKA ☐ EURACTIV ☐ SKYTG24 ☐ ADB ☐ EMD ☐ **UCAM**
Status	☐ Draft ☐ WP Manager accepted ☐ **Coordinator accepted**	
Action requested	☐ To be revised by Partners involved in the preparation of the deliverable ☐ To be reviewed by applicable AI4TRUST Partners ☐ For approval of the WP Manager ☐ For approval of the Project Coordinator ☐ **Ready for submission**	
Requested deadline for Action	Due date: 31 October 2023 Date of submission: 1 November 2023	

Summary of modifications

VERSION	DATE	AUTHOR(S)	SUMMARY OF MAIN CHANGES
0.0	01/09/2023	Riccardo Gallotti (FBK)	Table of contents
0.1	22/09/2023	ALL PARTNERS	First round of contributions
0.2	01/10/2023	Riccardo Gallotti (FBK)	Homogenisation by WP leader and project coordinator
0.3	06/10/2023	ALL PARTNERS	Second round of contributions
0.4	25/10/2023	ALL PARTNERS	Final round of contributions
0.5	26/10/2023	Serena Bressan & Maria Vittoria Zucca (FBK)	First review by the project coordinator
0.6	27/10/2023	Hugo Leal & Stefanie Felsberger (UCAM)	Review by the designated Quality Assurance Manager, i.e. UCAM
1.0	01/11/2023	Serena Bressan (FBK)	Final review by the project coordinator for report submission

Table of contents

List of abbreviations

ABBREVIATION	MEANING
ACL	Association for Computational Linguistics
AI	Artificial Intelligence
AVMSD	Audiovisual Media Services Directive
CULT	Committee on Culture and Education
DGs	Directorates-General
DMA	Digital Markets Act
DSA	Digital Services Act
ECAT	European Centre for Algorithmic Transparency
ECE	Expected Calibration Error

EDAP	European Democracy Action Plan
EDPS	European Data Protection Supervisor
EEAS	European External Action Service
EER	Equal Error Rate
EFJ	European Federation of Journalists
EU	European Union
EFMA	European Media Freedom Act
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
HLEG	High Level Expert Group
LLM	Large Language Models
MEPs	Members of the European Parliament
NeRF	Neural Radiance Fields
NLP	Natural Language Processing
PbD	Privacy by Design

PLD	Product Liability Directive
TFEU	Treaty on the Functioning of the European Union
VLOPs	Very Large Online Platforms
VLOSEs	Very Large Online Search Engines
WP	Work Package

List of figures

List of tables

Executive summary

Deliverable D2.1 "Design of the methodological toolbox," lead by the project coordinator Fondazione Bruno Kessler (FBK), is a public deliverable of “AI4TRUST - AI-based-technologies for trustworthy solutions against disinformation” and part of Work Package 2 (WP2) entitled “Methodological design, data gathering and pre-processing”. This report aims to identify and define together with the consortium the exact design of the set of data-driven, model-driven, and artificial intelligence-driven (AI-driven) tools that will be tailored and integrated to specifically serve the needs of the AI4TRUST platform and ecosystem.

This deliverable lays the groundwork for the intricate development process of the methodological toolbox within the AI4TRUST project. The toolbox will be progressively developed throughout the course of the project based on a comprehensive analysis of the current state-of-the-art approaches addressing the classification of online disinformation and misinformation. Specifically, we provide an overview of our conceptual framework to map misinformation, disinformation, and malinformation, along with the challenges and opportunities associated with linguistic diversity. We also highlight AI4TRUST's use of cutting-edge AI methodologies for textual, visual, and audio classification of misinformation and disinformation, including the development of tailored counter-narratives for real-time response mechanisms, that will be further illustrated in Work Package 3 (WP3) – “AI-driven data analysis methods”. Keeping the needs of the chosen AI methods in mind, we then outline the design of a data pre-processing pipeline to meet the requirements of our diverse technical partners. This collaborative effort aims to streamline the tasks outlined in WP3, ensuring the efficient realisation of the project's objectives.

Furthermore, the deliverable delves into the project's primary focus areas, providing a detailed account of the selection process for critical topics such as “Climate Change”, “Public Health”, and “Migrants”. It offers insights into the methodologies used for actively mining relevant data from various media platforms, reflecting the comprehensive analysis conducted in collaboration with the AI4TRUST consortium.

Lastly, we define the ethical and security boundaries inherent in the methodological options, emphasising the significance of upholding legal and ethical standards within the Work Package 5 (WP5) framework. WP5 is titled “Technical implementation of the platform & Security Framework”. This ensures a strong legal and ethical foundation throughout the execution of the project. In conclusion, this document illustrates the AI4TRUST project's commitment to integrating cutting-edge technologies while prioritising ethical considerations, contributing to the establishment of a more secure and trustworthy information environment within the European Union (EU).

1. State of the art on multimodal classification and countering of mis/disinformation

In its effort to address the significant challenges posed by misinformation and disinformation, the AI4TRUST project conducts a thorough investigation into the intricate domain of multimodal classification, employing a comprehensive set of effective strategies to combat the proliferation of unreliable information.

This section offers a comprehensive overview of the multifaceted strategies utilised to tackle the nuanced phenomena of misinformation, disinformation, and malinformation. To establish a coherent conceptual understanding, the section initially dissects the fundamental differences between these categories, establishing a robust framework for precise identification and differentiation, building what can be seen as the "AI4TRUST perspective" on this complex topic.

Additionally, we examine the intricacies the project encounters concerning linguistic diversity in the digital landscape, uncovering the inherent challenges and potential opportunities for managing information across diverse languages.

Finally, in its comprehensive analysis of the current landscape, AI4TRUST builds upon state-of-the-art methodologies, working towards the development of cutting-edge AI solutions. This section presents an overview of this state-of-the-art technology for the tasks of textual, visual, and audio classification of misinformation and disinformation, along with our efforts associated with the automated generation of counter-narratives tailored to specific social contexts that can play a pivotal role in real-time response mechanisms.

1.1 Conceptual Framework: Defining (and distinguishing) misinformation, disinformation, and malinformation

Before describing the AI methods integrated in our efforts to combat misinformation and disinformation, it is essential to establish a comprehensive conceptual framework delineating this intricate subject. The following conceptual framework builds on the scholarly consensus on the information distortions (Wardle & Derakhshan, 2017, p. 4), the best practices of our fact-checking partners[1], institutional definitions laid out in official EU documents (e.g., European Democracy Action Plan[2]) and our own adaptations:

[1] https://eufactcheckingproject.com/

[2]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

- **Misinformation:** Incorrect information produced or reproduced without either knowledge about its accuracy nor harmful intent.
- **Disinformation:** Incorrect, fabricated or misleading information that is intentionally shared by actors with the aim of disseminating it through social networks. The goals could go from deception to political or economic gains and may result in both individual and social harm.
- **Malinformation:** typically, factual information deliberately used with malicious and harmful intent (e.g., doxing, image-based sexual abuse (IBSA).

The focus is not only on the authenticity of the message but also on the authenticity and intentionality of the messengers and their interconnections. Intent appears as the defining feature and differentiating factor between online misinformation and disinformation. This distinction is crucially important for technical and analytical reasons as it entails the need for tools and methods with the capacity to elicit complex networked contexts behind and beyond the discrete units of content (individual posts, videos, images). Not all misinformation is disinformation, but all disinformation contains misinformation.

Nonetheless, there are various challenges in the definition of disinformation, and counter-disinformation practitioners have developed various terms of reference. While the definition outlined above is adopted by various actors in this space, other definitions have emerged given the growing complexity and the evolving nature of the online threat landscape. For instance, within the AI4TRUST consortium, the Global Disinformation Index (GDI) has developed a new conceptual framework to define disinformation.[3] In fact, current disinformation campaigns tend to combine seeds of factual information mixed with fabricated elements, leveraging overall adversarial narratives.[4] To respond to this challenge, GDI's definition of disinformation moves beyond the true and false dichotomy, and focuses on adversarial narratives.

These adversarial narratives weaponize social tensions, by exploiting and amplifying perceived grievances of individuals or groups and institutions, aiming to foster long-term conflict while undermining human rights.[5] Adversarial narratives are deployed through a combination of manipulated and factual information that crescendos into larger disinformation campaigns deployed across various spheres of the internet. Most importantly, GDI does not focus on intentionality given the practical challenges to establish the intention of an online actor sharing misleading information but rather the potential risk of harm/or actual harm. Within the AI4TRUST project, we have in particular selected three topics where adversarial narratives are expected to be present: migration, public health and climate change.

This conceptual framework leads to the central question at the heart of the AI4TRUST platform: What do we know about the truth claims presented in the information, such as text, video, image, and more? If the answer is simply that the information is incorrect, it should be classified as misinformation. If the information is accurate, but there is a reasonable assumption (made by a

[3] https://www.disinformationindex.org/mission

[4] https://www.disinformationindex.org/research/2019-4-1-adversarial-narratives-a-new-model-for-disinformation/

[5] https://www.disinformationindex.org/blog/2023-07-13-how-disinformation-is-undermining-our-human-rights/

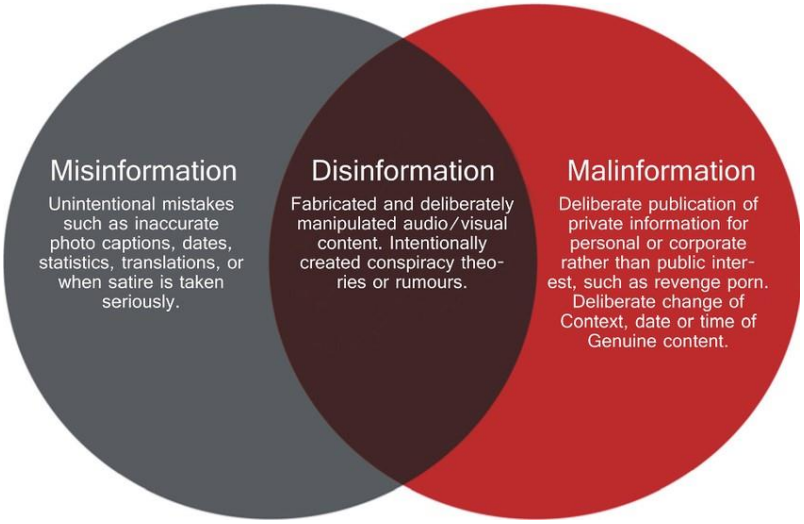
human) that the information is being disseminated with harmful intent, it should be classified as malinformation. If the information is found to be incorrect and both an indication from the automated network tool and human judgement point to deliberate distortion, it can be classified as disinformation. Therefore, the AI4TRUST initiative to identify intent must extend beyond natural language (NL) models such those developed in WP3 and involve network analytical tools that will be developed in Task T2.5 of WP2. Only the latter can meet the burden of proof regarding intentionality and coordination, enabling us to make analytical claims not just about misinformation, but also about disinformation.

1.2 Linguistic challenges and opportunities

While the term disinformation (and its linguistic variants) has been widely adopted, the same has not happened with the terms misinformation and malinformation. Both terms are well established in English but in most other languages we observe: 1) misuses of the word disinformation (and "fake news" used in its English form) as an all-encompassing description of every type of information distortion; and/or 2) adoption of borrowed words from English (e.g. fake news, misinformation); and/or 3) utilisation of an open compound word (closed compound in German and French) that describes the phenomenon accurately (e.g., *Cattiva informazione*, *Fehlinformationen*). To address this challenge we will strive to maintain the academic integrity and validity of the conceptual framework while valuing the cultural and linguistic diversity of the project. The terminology across the languages can be based on a curriculum model published under UNESCO auspices (Ireton & Posetti, 2018), which was translated in a series of European languages by academic teams.

For instance, the key terms from the Venn diagram (Fig. 1), *misinformation, disinformation* and *malinformation* are translated in Romanian *informarea greșită, dezinformare cu scop strategic*, and *informarea rău-voitoare*, respectively, while in Spanish the correspondent teams are *información errónea, desinformación e información maliciosa*. The project might use syntagma indicated in the translated version of the UNESCO Handbook for Journalism Education and Training (Ireton & Posetti, 2018) or equivalent terms as indicated by the project experts.

Fig. 1. Venn diagram for misinformation, disinformation, and mal-information (Wardle et al., 2018)



For that purpose, the commitment for our project is to: 1) employ correct terminology that follows the best professional practices and academic consensus 2) respect the linguistic diversity of the project by avoiding borrowed words and 3) promote the use of existing (if not yet popular) compound words and/or expressions (see Table 1). Since the users of the AI4TRUST platform are not the general public but rather experts (Fact-checkers, Journalists, Policy Makers, and Researchers) accustomed to dealing with various facets of mis/disinformation, our tool does not encounter the same challenges faced by organisations operating in public spaces within countries that use different terminologies. Our goal is to provide precise information to experts, facilitating their efforts to communicate effectively in their respective contexts.

Table 1. Information distortions and linguistic diversity in the 8 AI4TRUST languages (plus portuguese)

Language	Incorrect information	Deliberately incorrect information	Accurate information shared with harmful purposes	Other words/ expressions with social currency
English	Misinformation	Disinformation	Malinformation	“Fake News”
French	Mésinformation	Désinformation	Malinformation	“Fake News” ‘’Infox’’
German	Fehlinformation, Falschinformation	Desinformation	Malinformation	‘’Fake News’’

Greek	Παραπληροφόρηση	Σκόπιμη Παραπληροφόρηση	Κακόβουλη Πληροφόρηση	"Fake News" (Ψευδείς Ειδήσεις)
Italian	Misinformazione	Disinformazione	Malinformazione	Bufala, Macchina Del Fango, "Fake News", "Notizie False", "Un Fake", "Propaganda"
Polish	Mylne Informacje	Dezinformacja	(Malinformacja) Informacja Prawdziwa Udostępniona Z Intencją Wyrządzenia Krzywdy	Fake, Fałszywa Informacja, Propaganda, "Fake News"
Portuguese	Informação Falsa, Informação Incorreta	Desinformação	Ma Informacao	Noticias Falsas, "Fake News", Propaganda
Romanian	Informațiile Eronate	Dezinformarea	Informare Rău-Intenționată	"Fake News"
Spanish	Información Errónea / Desinformación No Intencional	Desinformación	Información Maliciosa	"Fake News", Información Falsa, Noticias Falsas, Bulo

1.2.1. A note on "Fake News" (in academic texts)

As many researchers, journalists and fact-checkers have stressed, the term "fake news" is wholly inadequate to describe, let alone identify and counter, the current climate of information disorder. Even though a falsehood can be newsworthy, a story is either news or fake. It cannot be both. In that sense, the expression "fake news" is an oxymoron that fuses and confuses two mutually exclusive words. Furthermore, it is important to acknowledge that, in recent years, the term "fake news" itself has been appropriated for political and economic purposes turning it into a scientifically inadequate, semantically inaccurate, and politically loaded expression. The term "fake news" should be avoided in any form of scientific communication. We can seize the current prevalence of the term as an opportunity to explain its inadequacy, particularly in public-facing communication settings. Nevertheless, the term "fake news" is widely employed in various contexts.

1.2.2 Terminology used in different mediatic contexts

To illustrate the potential variation in the usage of this terminology across different professional circles, the following section outlines how the AI4TRUST media experts and fact-checkers navigate the intricate terminological distinction between common usage and academic discourse.

1.2.2.1 Terminology used by fact-checkers

MALDITA (Spain)

Maldita.es only uses “fake news” or “false information” as an example of a wrong way of calling this phenomenon, since it was never information in the first place. Maldita understands that information is a piece of content that is truthful, and we should not give disinformation/misinformation the category of “information”. Though in English the terms “disinformation”/“malinformation” are just one word, in the Spanish translations the “information” word would be separated and Maldita believes that it attributes those contents an “information” status that they do not have.

Moreover, "fake news," although widely used, is an entirely abstract term that has become the preferred label for those in authority to dismiss genuine information they find unfavourable (e.g., governments classifying legitimate anti-corruption investigations as fake news). This term often finds itself at the heart of political conflicts. As the paper “Una Reflexión Sobre la Epistemología del Fact Checking Journalism: Retos y Dilemas” (Rodríguez Pérez, 2019) states, *“language is a reflection of our thoughts and the association between falseness with news, this understood as real facts, it’s a spear to the heart of journalism.”*

Nevertheless, it is still a term commonly used by citizens and sometimes in the media. “Noticias falsas” was the term used by the Centro de Investigaciones Sociológicas (Centre of Sociological Investigations, a Spanish public research institute - Government Agency) to ask questions about public perception of this issue in the last years, referring for example to the pandemic or the Ukraine war.

DEMAGOG (Poland)

DEMAGOG uses the term fake news on its own website (as a category of analysis debunking false information on social media) and in some article headlines, so that the reader clearly knows that it is an article debunking false information. This is mainly because in Poland the term fake news is the most popular. It is followed in order of popularity by the Polish term *dezinformacja* (disinformation) and the term *fałszywe informacje* (false information). The term “fake news” is often used to refer to information that is true, but does not fit a person's narrative or belief. The 2019 DEMAGOG report titled "Critical Mind (Krytyczny Umysł)" revealed that, according to the survey, 77% of respondents perceived 'fake news' as false information, 14% as manipulation, 6% as

humour, 1% as advertising, 1% had no specific opinion, and 1% offered an alternative definition[6] . During the DEMAGOG educational workshops, which are conducted as part of the Fact-Checking Academy, it is emphasised that the term "fake news" in Poland is highly politicised and the term "false information" should be used on a regular basis.

In Poland, the term "fake news" is frequently used by those spreading false information to reduce its importance and desensitise users to false information appearing on their sites. After the war in Ukraine, more and more people started to use the term fake news, also to refer to Russian propaganda. Mainstream media often treated the term “fake news” interchangeably with the term “disinformation”.

ELLINIKA HOAXES (Greece)

Ellinika Hoaxes continues to use the term “fake news” both as a category and occasionally in the titles of articles. Ellenika Hoaxes chooses to do this because the term is still prevalent in Greek society and is familiar to the general public. Primarily for this reason, Ellenika Hoaxes’s editors employ the term in specific instances, usually when addressing entirely fabricated claims.

However, Ellenika Hoaxes recognizes that the term is problematic. It is abstract and has been highly politicised, often used by politicians and online trolls to undermine the credibility of legitimate information. This issue is not unique to the term “fake news”; it also applies to the Greek term for “misinformation” (παραπληροφόρηση).

To address both the abstract nature and the politicisation of these terms, Ellenika Hoaxes has introduced more specific categories for debunking and has also implemented what Ellenika Hoaxes call “stamps”. Stamps are images displayed at the beginning of each article, as well as in each article's main image. Stamps inform readers about the type of debunking that the article pertains to. For example, Ellenika Hoaxes has stamps such as 'Misinformation,' 'Pseudoscience,' 'Conspiracy Theories,' 'Mix of truth and falsehoods,' 'Modified image,' 'Modified video,' and 'AI-generated image,' among others.

As previously mentioned, in Greece, the term “fake news” is often used interchangeably with “disinformation” by both the public and mainstream media. According to Google Trends, the use of the term “fake news” peaked in March 2020, likely because of the outbreak of the COVID-19 pandemic and the ensuing infodemic. Google Trends also indicates that both “fake news” and “misinformation” far outpace the Greek terms for “disinformation” (σκόπιμη παραπληροφόρηση) and “malinformation” (κακόβουλη πληροφόρηση), which are scarcely used and hardly popular at all.

[6] Stowarzyszenie Demagog. (2019). Krytyczny umysł. Problem fake news w Polsce. https://krytycznyumysl.pl/

1.2.2.2 Terminology used by media professionals

EURACTIV (European Union)

The term ''fake news'' initially referred to intentionally false information presented as news, aiming to deceive and/or manipulate the audience. However, the term later also started to be used interchangeably with other terms, misinformation, disinformation, and malinformation. The term has evolved and does not make a distinction between the others. Therefore ''fake news'' has become an all-encompassing term for problematic information. However understanding the nuances between the types of wrong information is crucial, as the intent with which certain information is being spread is different. Today, the term ''fake news'' is still used interchangeably with misinformation and disinformation. The term malinformation seems to be quite unknown. Within Euractiv the terms "fake news", misinformation, and disinformation are being used; however, the use of "fake news" is decreasing.

The nature of conflict reporting and disinformation across Europe has changed drastically over the last decade. The speed at which disinformation travels has increased and now information reaches citizens via all types of media and channels, from many different and possibly unreliable sources. It is not only media professionals who have the skills to create videos, but the general public is also equipped with the tools needed to produce content that can go immediately viral on TikTok or Instagram. For example, Renate Schroeder, European Federation of Journalists Director at the EFJ Conference said in 2021 that *"[p]olitical interference has always existed, but all our monitoring shows it is a growing phenomenon. We experience an 'illiberal turn' in Europe, where in some countries populist voices are given more space than facts. Journalism as a public good must be protected by all stakeholders including the public's right to know."* While media concentration is currently on the rise, media literacy remains low. A major threat to independent journalism is interference and pressures on public service media from both changes in the political environment and from increase of "news" from unreliable sources.

ADB (Romania / Eastern Europe)

Although the academic environment largely disagrees with the use of the term, it is frequently employed by Romanian journalists and professionals as a simplified and popular synonym for disinformation. It emerged in public discourse, particularly in language, as an opposing (popular) term to accurate information. Notably, in the title of the UNESCO Handbook for Journalism Education and Training (Ireton & Posetti, 1998), "fake-news" is graphically struck through, implying the need to counter it, but also raising a question about the term itself. In specialised academic circles, terms like misinformation, disinformation, and malformation (Wardle, 2020) are preferred.

However, there are instances where its usage facilitates generalisation. Recently, the term "disinformation" (without nuances) is more commonly chosen in academic discourse.

A recent study (The Newsreel Project Consortium, 2021, pp. 125-127) reviewed the literature on so-called "fake-news" and interviews with journalists in four countries were conducted to determine how they perceived and defined the term. Their main points are summarised below: Adriana Turea (Romanian journalist) believes that "fake news is built on a well-established pattern: it is based on a small part of the truth and the rest are lies mingled with facts". Jan Tvrdoň, Editor, Denik N in the Czech Republic pointed to the "false context" as the most significant distinction of so-called "fake-news." He also said that "in the Czech Republic, fake news is/was spread by the highest politicians". Tvrdoň referred to the social media ecosystem, the time and the ad-pressures creating "a system" that "often leads to publishing rubbish news and fake news as well". Adriana Turea (Radio Romania) also points to the pressure to broadcast information quickly, making thorough verification challenging. Propagating disinformation from lack-of-time and resources to check the information properly has become a challenge for quality journalists. Kathrin Wesolowski, freelance journalist in Germany and fact-checker for Deutsche Welle, pointed to "*false claims and false information being spread via media as well as via speech*", and named among the motivations political aims, or money (ex. click-bait). *"Fake news is not basically news, it's information, which is false, fabricated or misleading. It could be spread on purpose but it's not necessary*", as stated the Portuguese journalist Paulo Pena in 2021.

The Newsreel Project Consortium (2021) reviewed academic definitions, from Wardle (2017, 2020, respectively) to Allcott & Gentzkow (2017) which defined "fake news" as *"news articles that are intentionally and verifiably false, and could mislead readers"*. The research project aimed to provide the educational environment new tools in approaching the disinformation field. It had a module focusing on debunking, including definitions, reviewing literature, and publishing opinions of the professionals of the field.

Concerning the use of "fake news" among citizens, the majority of people, including journalists in television, radio, and print or online media, do not realise the oxymoronic nature of the phrase, the impossibility of associating news with falsehood. However, there are media experts, both in specialised NGOs and in the academic realm, who are acquainted with the inherently false nature of the expression. The extensive use of the term occurs quite naturally, mostly in the context of contrasting efforts to combat disinformation. "Fake news" has become a widely used term and is employed frequently in day-to-day language, especially when someone wants to refer to a deluge of false information, either as an item or overall. Terms such as false information or infodemic are rarely used as substitutes. The main Fact-Checking organisation of Romania, Funky Citizens, made use of the term "Fake News" in their call-to-action to citizens to report on "fake-news" and in the title of their newsletter called "Bulletin of fake-news"[7]. "It is time to fight the fake-news pandemic"

[7] Funky Citizens publishes Factual.ro, the first political fact checking site in Romania, launched in 2014. Apart from it, the organisation fact-check other themes like climate changes and since May 2023 is part of the Romanian-Bulgarian Observatory of Digital Media, https://brodhub.eu/ro/

is their slogan. However, on their website, factual.ro, they rather use the term “disinformation” when dismantling the fake from the fact.

EMS (Poland)

In Poland, there is no official dictionary definition for the term “fake news”. Commonly, “fake news”, refers to the dissemination of false or misrepresented information, often characterised by its sensational nature, intended to evoke strong emotions, and typically propagated for political or economic motives. This type of information is often disseminated without reliable verification. The above definition was formulated for the “2017 Youth Word of the Year”[8] poll in Poland, addressing the absence of a precise dictionary definition prior to this initiative.

EMS can recognizes three basic forms of fake news:

1. The complete untruth - the given information is completely fabricated.
2. The truth is disputed - facts are presented selectively or in context with the result that the recipient is misled.
3. The quotation manipulation - a statement is placed in context or sentences or parts of sentences are removed, changing the meaning of the statement and, as a result, supporting a particular thesis.

Fake news is often reflected in a form of disinformation, which can be any textual or audio-visual content that is disseminated consciously or unconsciously and has a negative impact on its audience. Harmful actions can influence a change of opinion, decision, or the assertion of a particular worldview. Disinformation also affects the sphere of knowledge, attitudes of individuals or entire social groups. “Disinformation content can lead to specific actions or inaction", according to the *Code of Good Practices* of the Polish Research Institute "NASK"[9].
Moreover, EMS can distinguish between:

1. Disinformation - a deliberate action aimed at fabricating or distorting an information message in order to achieve one's own political, social, financial, military, etc. gains. The effect of such a narrative is to mislead the other side (person/group/population), who is the recipient of the disinformation message.
2. Misinformation - reproduction of false and unverified information by a user who indiscriminately passes on manipulated content. These types of phenomena arise as a consequence of a lack of knowledge and contextual familiarity, and they are shared without any intention to cause harm.
3. Malinformation - truthful information shared with the intent to cause harm.

Within the Polish media ecosystem, three fundamental forms of fake news are identifiable:

[8] https://sjp.pwn.pl/mlodziezowe-slowo-roku/haslo/fake-news;6368870.html
[9] https://www.nask.pl/pl/wlaczweryfikacje/kodeks-dobrych-praktyk/4991,Kodeks-Dobrych-Praktyk.html

- Complete fabrication: Information that is entirely fabricated without any basis in truth;
- Disputed truth: Facts presented selectively or out of context, leading to the misleading of recipients;
- Quotation manipulation: Deliberate contextual framing of statements or selective removal of sentences or parts of sentences, altering the meaning of the statement and supporting a particular narrative.

Moreover, there are instances where troll farms are capable of generating and disseminating false information to benefit a specific organisation or a hostile state.

SKYTG24 (Italy)

Disinformation, misinformation and malinformation are terms that have no equivalent words in Italian, although, in the case of "disinformazione", now it is often used in the sense of English disinformation. Disinformation, in other respects, can also indicate (adjective) an "uninformed" person, a person who has had or obtained approximate, inaccurate, or wrong information[10]. From this point of view, the sense is similar to that of Italian adjective "male informato" (very different meaning from the English malinformation).

In general the “fake news” semantic field is difficult to catalogue, mostly because the expression “fake news” is widely used among Italian media, official sources, and general public to identify generic information, partly or completely false, spread both to cause harm but also out of lack of competences and of domain knowledge[11]. However, usage of “fake news” has been criticised due to the lack of a clear definition[12].

Instead, the media ecosystem constantly refers to a “disinformation campaign” to describe the organised dissemination of false or manipulated information with the purpose of causing harm. There is, however, a widespread use of “fake news” also within the media ecosystem. This same expression is also widely used by official sources, as e.g. by the Italian Healthcare Ministry[13].

In this context, at Sky TG24 we use both terms, “fake news” and disinformation; rarely misinformation. Moreover, it is widely used by the general public – and sometimes in media too – the generic term “bufala” or its plural “bufale”, which is almost synonymous with “fake news”. This expression carries the same problem of “fake news”, lacking a clear definition of it and making it hard to understand at first sight if it is indicating a generic misinformation or a wider disinformation campaign.

The term "fake news" has been widely used in Italy to indistinctly identify any kind of information disorder, especially in the 2017-2020 period. According to Google Trends, the peak of "fake news"

[10] (source: https://accademiadellacrusca.it/it/consulenza/misinformation-e-debunking-abbiamo-i-mezzi-per-tradurli/2997)
[11] (source: https://www.valigiablu.it/disinformazione-fake-news-propaganda/)
[12] (source: https://www.treccani.it/enciclopedia/fake-news/)
[13] (source: https://www.salute.gov.it/portale/nuovocoronavirus/archivioFakeNewsNuovoCoronavirus.jsp)

use in Italy was in 2020. In the following years, use of the expression has decreased. Regardless, it continues to be more widely used than the more correct terms such as "misinformation" and “disinformation." In the aftermath of the Covid-19 pandemic, the term "conspiracy" has received strong popularity in public discourse, with a usage similar to that of "fake news." In particular, it has been used to refer to disinformation campaigns against the vaccine campaign.

The term misinformation is very little used in Italy: it is a neologism borrowed from English, used mostly in research and educational settings and sometimes in the media ecosystem. Lastly, it is useful to underline that the term “malinformation” has no widely used translation in Italian or any dissemination neither in media outlets nor general public. Despite the sporadic appearance of the translation “malinformazione” in official sources[14], this term still has to make a breakthrough in Italian public debate. There is also another concept in Italian – mala/cattiva-informazione - which sounds almost identical but carries a slightly different meaning (deliberate false information, for attacking enemies, especially in the political discourse).

The problem with the categorization disinformation/misinformation/malinformation is that we often do not know what the real intentions are of those who produce content that falls under information disorder. Doing an analysis of intentions is not at all easy, especially in the case of "misinformation": how do you determine that a piece of content was published or shared without intentions to do harm? Intentionality is a highly subjective factor that cannot be defined objectively.

1.2.3 Contextualising terminology used in the European Law

Defining dis-information, mis-information, and mal-information from a legal standpoint is problematic. The EU’s definitions and those of the member governments are beginning to converge. Some member governments have laws covering disinformation or making the dissemination of disinformation a punishable, even a criminal, offence. Disinformation may not just cover potentially harmful content but content that some member governments define as illegal. The UN Rapporteur on Freedom of Expression has stressed how the concept of disinformation is an *“extraordinarily elusive concept to define in law”*, and open to arbitrary interpretation by providing executive authorities *with “excessive discretion to determine what is disinformation, what is a mistake, what is truth”*. As a result, penalties can be disproportionate and arbitrary, as has been found by the European Court of Human Rights (Fathaigh et al., 2021).

There is currently no legal framework at the EU level specifically addressing disinformation, except for the provision stated in Article 11 of the Charter of Fundamental Rights, which guarantees the freedom of expression and information. This provision ensures that everyone has the right to

[14] (source: https://www.esteri.it/it/sala_stampa/archivionotizie/approfondimenti/2023/05/voci-dalla-farnesina-giornata-mondiale-della-liberta-di-stampa-le-minacce-della-disinformazione/)

express their opinions and receive or share information without any interference from public authorities, and upholds the importance of media freedom and pluralism.[15]

Series of policy initiatives and action plans have been developed by the European Union (EU) to tackle the issue of disinformation. These initiatives are voluntary and do not have immediate enforcement until regulatory measures are introduced. The objective of these initiatives is to safeguard democratic systems, protect public opinion, and combat false or misleading information. Here are some notable examples: (a) European Democracy Action Plan: This plan, which was announced in December 2020, seeks to address disinformation and foreign interference, promote media freedom and diversity, and enhance the integrity of elections. It includes measures to bolster the resilience of democratic systems, such as implementing transparency regulations for digital political advertisements and providing support to fact-checking organisations; (b) Code of Practice on Disinformation: In 2018, the EU established a voluntary Code of Practice that was endorsed by major online platforms with the aim of combating the spread of disinformation. This code requires signatories to increase transparency regarding political advertising, minimise financial incentives for disinformation, enforce stricter ad policies, and enhance collaboration between platforms and fact-checkers.

There is no EU legal framework governing disinformation apart from Article 11 of the Charter on Fundamental Rights on the Freedom of Expression and Information, and a series of policy initiatives, actions, and action plans that rely on voluntary compliance: until regulatory measures are introduced and implemented, they do not take direct effect nor are they binding or immediately enforceable. This has implications regarding the legal or voluntary responsibilities of member governments and or the EU and service providers regarding how to define and respond to the misuse of information to create harm and undermine trust in public authorities.

The EU's commitment to tackling disinformation and safeguarding public opinion is evident through these initiatives and plans. They aim to promote trustworthy information, increase transparency in political advertising, and foster collaboration among platforms, fact-checkers, and Member States. Consequently, this also raises questions about the legal obligations of member governments, the EU, and service providers when it comes to defining and addressing the misuse of information with the objective of causing harm and eroding trust in public authorities.

The Digital Services Act (DSA) is a proposed legislation by the European Union designed to update the regulatory framework for digital services and platforms, focusing on liability, competition, transparency, and user protection. Its goal is to shape a safer, fairer, and more accountable digital

[15] Paragraph 2 of this Article spells out the consequences of paragraph 1 regarding freedom of the media. It is based in particular on Court of Justice case law regarding television, particularly in case C-288/89 (judgement of 25 July 1991, Stichting Collectieve Antennevoorziening Gouda and others [1991] ECR I-4007), and on the Protocol on the system of public broadcasting in the Member States annexed to the EC Treaty, and on Council Directive 89/552/EC (particularly its seventeenth recital).

environment for businesses and users. Robust enforcement of horizontal content moderation rules in the DSA is essential and the most important legal instrument against disinformation. There were initial concerns that big platforms, including state media, would be exempted from content moderation obligations under the European Media Freedom Act, and conflict with DSA provisions[16]. At this stage, the DSA remains the most important legal basis for action, and derives its legal power from being based on the internal market.

EU policy assumes that disinformation is not illegal per se, but is potentially harmful depending on intent to deceive. The EU Commission differentiates illegal content (such as child abuse or hate speech) from harmful content. In practice, there could be overlap, contradictions and ambiguities in definitions, scope, and legal competence within the EU and across the Member States (Betzel et al., 2020).

EU action is derived from the responsibilities related to the European treaties. EU Commission's Directorate-General (DG) or EU services acts in this field are continuously evolving and this complicated a uniform definition of the problem, terminology and action. Therefore, member governments are responsible for combating disinformation. The EU's legal role is to support that with a common vision and actions to strengthen coordination, communication, and the adoption of good practice according to the articles 2-6 of the Treaty on the functioning of the European Union (TFEU).

The EU Court of Auditors recommended actions to support the member governments, stressing the need for augmenting coordination and effectiveness, showing proportionality to the type and scale of threat, and building on the success of the EUvsDisinfo[17] which it regarded as "instrumental in raising awareness about disinformation". The location of the action inside the European External Action Service (EEAS), was problematic (# 114).

Ongoing legislative initiatives in 2023 are, therefore, important. Media and digital services law, along with cyber security, are evolving fast. While the European Parliament was able to adopt its negotiating position on the European Media Freedom Act (EFMA) on 4 October 2023, there are concerns among the Members of the European Parliament (MEPs) and professionals that the absence of an absolute prohibition on the use of spyware (like Pegasus and Predator) against journalists will compromise their role as watchdogs, impede accurate reporting and induce self-censorship[18]. The provision to exempt news media from content moderation on VLOPs, as prescribed in the Digital Services Act, means that platforms are prohibited from removing content published by media service providers for 24 hours that could potentially allow rogue actors to disseminate disinformation for 24 hours before platforms are allowed to take it down, according to

[16] European Media Freedom Act: No to any media exemption, 15 May 2023 https://www.euractiv.com/section/digital/opinion/european-media-freedom-act-no-to-any-media-exemption/

[17] euvsdisinfo.eu

[18] https://www.mappingmediafreedom.org See EUI Centre for media Pluralism and Media Freedom, The Media Pluralism Monitor (2022), https://cmpf.eui.eu

reports in Euractiv citing the Computer and Communications Industry association after the vote. This constitutes a major problem for combating disinformation and will be raised during the upcoming trialogue negotiations starting on 18 October 2023[19].

The problems of diverse legal competence and policy responsibility for disinformation related actions must be acknowledged. How terminology is used and its strategic or policy purpose affects legal effect.

1.2.3.1 Cross-cutting legal context

The wider legal context for what happens next includes the recently agreed Digital Services Act; the 2021 Commission recommendation on the protection, safety and empowerment of journalists, and its April 2022 proposal for a directive to protect journalists and rights defenders from strategic lawsuits against public participation (SLAPPs), to protect independent media outlets against litigation aimed at intimidating or silencing them.

To assist understanding of the problems that arise because policy competence for disinformation is diffused across different EU DGs in the Commission and EU services, brief pointers to important EU actions are provided below, along with the most recent common definitions now being used in the EU.

There are discrepancies in how the terms fake news, malinformation, misinformation and disinformation are used or conflated by different institutions, including UNESCO[20], by third state governments and companies, and especially in public discourse in different states. The distinctions overlap and in public discourse may be used interchangeably. Politico-cultural context, traditions and expectations remain influential[21]. However, the EU and US are attempting to align and create a common understanding and taxonomy for artificial intelligence with a view to avoid inconsistencies and unhelpful divergences, and so enhance the basis for trustworthy AI.[22] The politico-cultural context remains influential.

1.2.3.2 Defining the terms

The academic definition (Wardle & Derakhshan, 2017) useful for the project distinguishes between three types of false or harmful information as discussed above:

[19] https://www.euractiv.com/section/media/news/eu-media-law-enters-home-stretch-but-spyware-disinformation-concrns-persist/ accessed 4 October 2023

[20] https://unesdoc.unesco.org/ark:/48223/pf0000381137

[21] EU Joint Research Centre, (2022), Glossary of human-centric artificial intelligence, Seville. See too CNIL, Artificial Intelligence:the CNIL opens a consultation on the creation of datasets for AI, 16 October 2023. https://cbil.fr

[22] EU-US Terminology and Taxonomy for artificial intelligence (2023) TTC Joint Roadmap for Trustworthy AI and Risk Management, 2 December 2022, updated 4 February 2023 . https://digital-strategy.ec.europa.eu,

- dis-information : intent to cause harm by deliberately sharing false information
- mis-information : false information shared inadvertently, without intent to cause harm
- mal-information : genuine information or opinion shared to cause harm, such as harassment, hate speech

To assist understanding of the problems that arise because policy competence for disinformation is diffused across different EU Directorates-General (DGs) in the Commission and EU services, brief pointers to important EU actions are provided below, along with the most recent common definitions now being used in the EU.

The most widely agreed definition is that of the EU High Level Expert Group: "disinformation includes all forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or profit" (Buning et al., 2018).

a) "Fake news"

The term "fake news" has been rejected by the High-Level Expert Group (HLEG) appointed by the EU Commission to advise on fake news and online disinformation. Member governments tend to agree that the term is too broad, vague and ambiguous. Instead, information manipulation has been proposed as more accurate, and academic studies refer to information disorder.

Whereas "fake news" is typically used as a catch-all to cover all the above terms, a more precise definition might describe fake news as fabricated news: content that may be a lie, a distortion of the truth, a fantasy or an idea disseminated with mal-intent designed to deceive. That differentiates it from misinformation where mal-intent is not to be inferred automatically, although the content may include false information which the disseminator nevertheless believes to be true. Mal-information is based on reality but is used with the intention of inflicting harm.

b) Disinformation and misinformation

The most widely agreed definition is that of the EU High Level Expert Group: *"disinformation includes all forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or profit."* (Buning et al., 2018). Disinformation is different as the context of its use may be designed to manipulate, influence and skew decision making, with or without mal-intent. As will become clear below, the EU's definitions imply that content falling into these categories can be incidentally harmful or intentionally harmful by design.

It is important to note that there is a risk that different automated decision-making outcomes may arise as a result of divergent classifications. There are difficulties in defining disinformation in EU legislation, including that on online platforms and removal of illegal content, as they may arise in wide national laws criminalising false news and false information.

Where the EU is concerned, legal definitions are beginning to converge (Fathaigh et al., 2021). It is important to recognise that definitions are contextually contingent. In 2020 the EU Commission published its European Democracy Action Plan (EDAP). The EDAP is part of the EU wider strategy

to promote free and fair elections, strengthening media freedom and counter disinformation. It was an important milestone in the development of the "Strengthened Code of Practice on Disinformation" in 2022. Some 34 signatories - platforms, tech companies and civil society - followed the 2021 Commission Guidance and took into account the lessons learnt from the COVID19 crisis and Russia's war of aggression in Ukraine.[23] It defines:

- misinformation as "false or misleading content shared without harmful intent though the effects can be still harmful", and
- disinformation, as "false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm"(Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European Democracy Action Plan, 2020, p. 18)

Accepting such definitions and implementing policy measures accordingly is not easy. EU law takes precedence over national law with which it conflicts but the Codes of Practice and Communications such as the 2018 "Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Action Plan against Disinformation" (Brussels, 5.12.2018 JOIN(2018) 36 final)[24] are not binding and rely on goodwill and voluntary compliance. This means that as important as the Code is (as described below), it is insufficient by itself. Therefore, new laws are essential. Even so, the Commission saw this as positive.

In this regard, Věra Jourová - Vice-President of the European Commissionefor Values and Transparency - said:

"This new anti-disinformation Code comes at a time when Russia is weaponizing disinformation as part of its military aggression against Ukraine, but also when we see attacks on democracy more broadly. We now have very significant commitments to reduce the impact of disinformation online and much more robust tools to measure how these are implemented across the EU in all countries and in all its languages. Users will also have better tools to flag disinformation and understand what they are seeing. The new Code will also reduce financial incentives for disseminating disinformation and allow researchers to access platforms' data more easily."

Secondly, Thierry Breton – European Commissioner for Internal Market - said:

[23] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423 and https://digital-strategy.ec.europa.eu/en/news/disinformation-commission-welcomes-new-stronger-and-more-comprehensive-code-practice-disinformation

[24] Communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European Democracy Action Plan, COM/2020/790 final (2020).

"Disinformation is a form of invasion of our digital space, with tangible impact on our daily lives. Online platforms need to act much more strongly, especially on the issue of funding. Spreading disinformation should not bring a single euro to anyone. To be credible, the new Code of Practice will be backed up by the DSA - including for heavy dissuasive sanctions. Very large platforms that repeatedly break the Code and do not carry out risk mitigation measures properly risk fines of up to 6% of their global turnover." [25]

He issued an open letter to X[26] and a letter to TikTok requiring immediate responses in the wake of disinformation about Israel and Hamma. X prevaricated but TikTok acted to combat it over the weekend of October 14, 2023.[27]

Together with the recently agreed Digital Services Act[28] and the upcoming legislation on transparency and targeting of political advertising[29], the strengthened Code of Practice is an essential part of the Commission's toolbox for fighting the spread of disinformation in the EU. The 34 signatories include major online platforms, notably Meta, Google, Twitter, TikTok, and Microsoft, as well as a variety of other players like smaller or specialised platforms, the online ad industry, ad-tech companies, fact-checkers, civil society or that offer specific expertise and solutions to fight disinformation.

The Code aims to become recognised as a Code of Conduct under the Digital Services Act to mitigate the risks stemming from disinformation for Very Large Online Platforms. It is at the core of the EU strategy against disinformation. It is significant that whereas the original communication in 2018 originated in the EU's External Action Service[30], combating disinformation and misinformation designed to impair EU democracy and undermine public trust in political and legal authorities has been mainstreamed. However, this has been a slow process (Lodge, 2010).

For the EU, actions to combat disinformation, misinformation and malinformation are embedded in digital initiatives in different policy fields and sectors, notably that on the internal market (art 114 TFEU, see below). However, the emerging legal definitions are informed by their sectoral location. Where media policy and digital policies are concerned, they are to serve the purposes of core EU goals: the protection of EU values and rights against anti-democratic forces, and to combat foreign interference and communication designed to undermine democracy.

[25] *https://ec.europa.eu/commission/presscorner/detail/en/ip_22_3664*
[26] pic.twitter.com/J1tpVzXaYR
[27] https://www.euractiv.com/section-global-europe/new/eur-breton-urges-musk-to-tackle-spread-of-disinformation-on-x-after-hamas-attack;
https://www.theguardian.com/technology/2023/oct/15/tiktoksays-it-has-acted-to-curb-disinformation=amid-israel-hamas-war;
[28] *https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545*
[29] https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6118
[30] www.eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf

1.2.3.3 Intersecting policies relevant to creating trust

The EU is at the forefront of crafting policies aimed at fostering trust in a rapidly evolving digital landscape. As technological advancements continue to shape our societies, several intersecting policies have emerged that address key concerns such as media freedom, foreign information manipulation, media literacy, responsible artificial intelligence (AI), and digital services and markets. These policies, including the Media Freedom Policy, Foreign Information Manipulation and Interference measures, Media Literacy & Responsible AI initiatives, and the Digital Services Act and Digital Markets Act, are all designed to establish a trusted environment where individuals and businesses can engage with digital platforms and online content without fear of manipulation, misinformation, or unfair practices. By prioritising these areas, the EU seeks to build an inclusive and secure digital economy that benefits both its citizens and the broader international community.

a) Media freedom policy

At the time of writing, the new media freedom policy was in the final stage of adoption. This has to be viewed against the backdrop of legislative initiatives since 2018 to address realising digital society and protecting EU democracy, the rule of law and EU values and rights.

For the past three years, the EU Commission has produced a Rule of Law Report examining the Member States' (i) justice systems (i.e., their independence, quality, and efficiency); (ii) their anti-corruption frameworks to prevent and fight corruption; (iii) media freedom and pluralism; and (iv) institutional issues related to checks and balances, focusing on key areas important to ensuring the rule of law. This Rule of Law Situation in the European Union Report is issued as a "communication" (i.e., devoid of legal obligations on the Member States). It stresses that:

"Safeguarding and upholding our democratic institutions and values is a shared responsibility of Member States and EU institutions. This is all the more important now when the EU and its Member States are threatened by hostile foreign actors using disinformation and cyber-attacks to try to undermine our democracies."[31]

Most recently, in August 2023, Poland and Malta were seen as hotbeds of legal harassment of journalists using intimidation and SLAPPs (i.e., strategic lawsuits against public participation), seen as a threat to democracy. A coalition of non-governmental organisation from across Europe, CASE, warned that SLAPPs *"attempt to intimidate and silence public watchdogs through lengthy and expensive litigation that drains a target's resources and chills critical voices [...]. We work to expose legal harassment and intimidation, protect the rights of those who speak out, and advocate for comprehensive protective measures and reform' using a three-pronged approach: exposure, building resilience, and advocating law reform and stronger safeguards in law"*. CASE is funded by

[31] European Commission (2022) Communication from the Commission to the European PArliament, the council, the European Economic and Social Committee and the Committee of the Regions, 2022 Rule of Law Report, Luxembourg 13.7.2022 COM(2022) 500 final https://eur-lex.europa.eu/illegal-content/EN/TXT/?uri-CELEX%3A52022DC0500

the Open Society Foundation (OSF), and makes technical assessments of the EU Commission's anti-SLAPP initiative, and gathers intel from EU institutions and networking[32].

b) Foreign information manipulation and interference

In September 2023, the EU urged Big Tech platforms to act against The Kremlin's "war of ideas" as it faced more disinformation on X (formerly Twitter) believed to originate or be inspired by Russia[33] for the purposes of undermining trust in EU authority and the rule of law. In this fast-moving field, the EU relies too much on rapid reaction by relevant nationally based authorities and agencies to block, take-down or fact check and correct disinformation and mal-information. Recent examples include measures to sanction and ban Russian RT (i.e., RT is a Russian state-controlled international news television network funded by the Russian government) [34]. On its borders, in the UK close to 8 percent of Google and Microsoft search results on specific topics in the United Kingdom were linked to foreign government actors, according to a report[35] for the country's telecommunications and media regulator.

Member States' governments are divided over banning, limiting access to, or censoring potentially harmful content[36] and this is a potent field for the dissemination of disinformation and misinformation.

c) Media literacy & responsible AI

The European Data Protection Supervisor (EDPS) locates the problems within the context of reflections over the kind of society being upheld or undermined. *"Even if fake news is spread heavily on social media, research has found that human behaviour ("word of mouth" marketing) contributes more to the spread of fake news than automated bots do. This shows that fighting the fake news sender is not the only approach. It also makes sense to increase the resilience to fake news on the side of the recipient and our society. Therefore, another important pillar of fake news detection is to increase citizens' awareness and media literacy"*.

Accordingly, the EDPS issued an opinion on 11 October 2023 (Opinion 42/2023) on the Commission's proposal of 28 September 2022 for two Directives on AI liability rules regarding the revision of the Product Liability Directive (PLD) and the Directive on adapting non contractual civil liability rules to artificial intelligence (AILD). The EDPS calls for equivalent (the same) protection for individuals who suffer damages caused by AI systems produced and/or used by EU institutions, bodies or agencies as used and/or produced by private actors or national authorities, irrespective of its classification as high-risk or non-high-risk.[37] It is against this context and the evolving landscape

[32] https://www.the-case.eu/about/
[33] https://euobserver.com/eu-political/157467
[34] https://www.euractiv.com/section/digital/news/eu-rolls-out-new-sanctions-banning-rt-and-sputnik/
[35] https://politico.us8.list-manage.com/track/click?u=e26c1a1c392386a968d02fdbc&id=604147d1b2&e=74e86d5ad0
[36] https://edps.europa.eu/press-publications/publications/techsonar/fake-news-detection_en
[37] EDPR (2023) Opinion 42/2023 on the Proposals for two Directives on AI Liability rules. https://edps.europa.eu/systemfiles/2023-10/2023-0622_d311_opinion_e.pdf

of AI legislation, declarations on AI Edinburgh Declaration on Responsibility for Responsible AI internet governance, standards, and protocols that the EU's Media Freedom decisions must be seen.

In January 2022, the EU Commission launched a public consultation on media freedom. In September, seeking to transform unbinding codes into a binding, directly applicable EU Regulation, the EU Commission adopted a European Media Freedom Act (EMFA) to protect media pluralism and independence, after the Commission included its initiative in its 2022 work programme, and concluded its communication on its annual Rule of Law report - the rule of law situation in the EU in July 2022[38]). On 15 July 2022, the Commission referred Hungary to the Court of Justice for breaching EU media freedom and telecoms rules. In June 2023, all this led to the establishment of a common framework for media services in the EU internal market. (https://www.consilium.europa.eu/en/press/press-releases/2023/06/21)

The situation remains somewhat fluid as the necessary legislative steps have yet to be completed. The aim is to prevent political interference in editorial decisions, protect journalists, notably against intrusive spyware (Pegasus), set requirements for audience measurement systems, and open allocation of state advertising, protect content against online content removal, and ensure transparency in media ownership, plus a European board for media services. The European Parliament has been active in this especially since 2020. In March 2022, the European Parliament set up a committee of inquiry (PEGA) to examine the use of spyware. In June 2023, it adopted a resolution, drafted by its special committee on foreign interference (ING2), stressing the need for an EU coordinated strategy against foreign interference and information manipulation, which it expected to increase during the elections in 2023 and the European Parliamentary elections in 2024. MEPs called for the establishment of a rapid alert system for MEPs and national MPs to counter online disinformation.

d) Digital Services Act

The legal basis for the proposal is Article 114 of the Treaty on the Functioning of the European Union (TFEU) and it also amends the rules of the Audiovisual Media Services (AVMSD) Directive and complements the Digital Services Act (DSA). Exceptions are to be allowed on the grounds of national security, and specific criminal offences under investigation (including terrorism, child abuse or murder, on a case-by-case basis. The DSA is a horizontal instrument aiming to create a safer and trusted online environment. Online platforms are required to be more open and accountable (on how content is recommended to users) and establish measures to ensure users' safety online, prevent interference and protect users from harmful and illegal content, goods, and services; or that designed to influence users' behaviour (dark patterns). Very large platforms (VLOPs) and search engines (VLOSE) must comply with stricter obligations under the DSA that also covers harmful content and disinformation.

[38] COM (2022) 500 final, 13.7.2022) (https://eur-lex.europa.eu/legal-content/EN/TXT?uri=CELE%3A52022DC0500

The DSA entered into force on 1 November 2022. Its implementation is to be helped by the European Centre for Algorithmic Transparency (ECAT) set up in April 2023. In May 2023, the Commission launched a consultation on draft rules on how independent audits should be done under the DSA for VLOPs and VLOSEs, followed by one in June 2023 on the required transparency database. In September 2023, the Commission launched the DSA Transparency Database. The Council adopted its negotiating position on 21 June 2023. The European Parliament's Committee on Culture and Education (CULT) adopted its own report on 7 September 2023. (COM(2022) 457 2022/0277(COD) The European Parliament is to adopt its negotiating mandate at the plenary in October 2023.

e) Digital Markets Act

These legal initiatives are located within the EU's commitment to making 'A Europe Fit for the Digital Age'. The Digital Markets Act (Regulation 2022/1925 OJ L 265 12.10.2022, p.0001; DMA) was adopted as a regulation on 13 October 2022 and in force as of 1 November 2022[39]. It applies to large companies - designated as gatekeepers - providing an array of services, including social networks, video sharing, virtual assistants, web browsers, search engines and online advertising and imposes new obligations on them, including an obligation to provide interoperable messaging services (up for review) and prohibiting various practices such as a self-preferencing, or reuse or private data collected during provision of one service for the purposes of another service. The Commission will be solely responsible for enforcing the DMA, assisted by a high-level group of digital regulators and in close cooperation and coordination with national authorities[40]. Gatekeepers can be fined up to 20% of worldwide turnover for failing to comply for repeat offences and in the event of systematic non-compliance face a fixed term ban of acquiring other companies. The Commission on 6 September 2023 formally designated 6 gatekeepers (Alphabet, Amazon, Apple, ByteDance, Meta and Microsoft for specific platform services. They have six months to comply with DMA rules. The DMA is one of the first regulatory tools to regulate the gatekeeper power of the largest digital companies, complementing competition rules.

1.3 Classification/detection of mis/disinformation

In AI4TRUST, we integrate an array of sophisticated methodologies, including text, audio, and visual classification techniques to discern between misinformation and disinformation. Given the exponential proliferation of data in the digital domain, the implementation of robust textual, audio,

[39] https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-digital-markets-act

[40] https://digital-markets-act.ec.europa.eu/dma-commission-launches-consultation-template-compliance-report-2023-06-06_en

and visual classification methodologies has become imperative in deciphering the veracity of information. In parallel, the crucial application of Social Network Analysis (SNA), further empowers us to untangle the intricate web of information dissemination, uncovering subtle patterns and interconnected nodes that often underlie the propagation of false narratives. This harmonious integration of diverse classification methodologies not only fosters a comprehensive understanding of the intricacies within the informational landscape but also equips us with the requisite analytical tools to bolster trust in our collective pursuit of truth.

1.3.1 Textual classification of mis/disinformation

In this section, we delve into the dynamic landscape of Natural Language Processing (NLP) and its role in combating misinformation and promoting trust in digital spaces. Initially, we outline the evolution of the NLP community's efforts in automating fact-checking processes, followed by an exploration of how the automatic detection of hate speech has enabled the identification of nuanced forms of disinformation. Additionally, we discuss how NLP continues to advance in its mission to detect logical fallacies, ensuring a more comprehensive understanding of information veracity. Finally, we offer insights into the firsthand experiences of the esteemed members of the AI4TRUST team, shedding light on their invaluable contributions to these pivotal areas.

Mis/disinformation is disseminated in the modern media ecosystem through various signals. Identifying these signals is of high importance to combat its spread. Within the NLP community, most of the efforts related to the detection of online mis/disinformation have focused on the task of assessing whether a specific claim is true or not, so-called *automated fact checking* (Guo et al., 2022), while the expression fake news has been generally used to encompass false information circulating online, without paying too much attention to sources (i.e., news headlines or posts on social media) and fine-grained distinctions among their intent, which may be harmful or not.

A broad range of tasks, datasets, and NLP approaches have been introduced in different parts of the fact-checking process to automatically determine the truthfulness of a claim in media. Common classification tasks are i) *claim detection*, to identify claims that require verification (Thorne et al., 2018, 2019; Aly et al., 2021), ii) *evidence retrieval*, to find sources supporting or refuting the claim (Wang, 2017), and iii) *veracity prediction*, to assess the veracity of the claim based on the retrieved evidence (Shu et al., 2020). For more information on NLP-fact checking, readers are referred to the recent review of Das et al. (2023). A long list of tools is available on the web to search for false/fake claims [41][42].

Claim detection is a crucial component of the pipeline. It aims to identify which text passages should be verified because the general public would be interested in knowing the truth (Alberto Barrón-

[41] https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html
[42] https://www.investintech.com/resources/blog/archives/9120-fact-check-tools-tips.html

Cedeño et al., 2020). Claims are based on appropriateness judgments, i.e. individual and collective mindsets that can be parametrized (Lazega, 1992). This task overlaps to some extent with *rumour detection* (Gorrell et al., 2019), whose goal is to identify an unverified story or statement typically circulating on social media. Indeed, rumours are usually check-worthy because, if not true, they may lead to mis/disinformation.

After detecting check-worthy claims, *evidence retrieval* is typically applied, i.e., to find existing information to confirm (or not) its veracity. Reliable information can be found in publicly released datasets, legal documents, trustworthy news sources, Wikipedia, etc. (Li et al., 2016). This step foresees the prior development of one or more knowledge sources, manually curated by experts, against which claims should be compared. This comparison leads to the third process step, which is *claim verification* and that, taking the claim and the evidence in input, outputs a label corresponding to the veracity of the former. Most NLP approaches to the task adopt a supervised framework, in which a classifier is trained starting from pairs of claims/evidence manually annotated as supported or refuted. In some cases, a third label may be introduced when, given the provided evidence, it is not possible to check whether the claim is true or not (Thorne et al., 2018).

Online hate speech can also be used as a major disinformation signal, based on offensive and hostile language to promote racism, sexism, violence, and gender discrimination (Jahan and Oussalah, 2023). In the case of disinformation, this applies in particular to subtle forms of online hate such as the use of stereotypes, sarcasm, or implicit hate.

Text classification of tweets and posts (e.g., YouTube, Facebook, Reddit, Gab) has been extensively conducted with machine learning algorithms and language models (LMs). Hate speech classification has been treated in the past as a binary classification task (e.g. Hate/Not Hate, Toxic/Non Toxic, Hateful/Non Hateful) (Basile et al., 2019; et al., 2021; Pavlopoulos et al., 2021), a multiclass problem (Hate speech/Offensive/Neither, Hate/Abusive/Spam/None) (Davidson et al., 2017; Founta et al., 2018; Grimminger et al., 2020) or a multilabel task covering different hate targets such as gender, race, origin, disability, religion, and sexual orientation (Mollas et al., 2022).

Identifying logical fallacies is another signal category that can determine whether online information is valid. A fallacy is generally difficult to be detected since the related statement or argument is often based on faulty reasoning. Thus, it may seem to be valid but is not so (Tindale, 2007). Furthermore, the existence of hundreds of fallacies[43] makes their detection a challenging task. LMs have been used in the past to perform multiclass classification on news to detect fallacies. For instance, Bonial et al. (2022) used 226 articles for Covid-19 news to detect 6 fallacies (ad hominem, appeal to emotion, red herring, hasty generalisation, irrelevant authority, non-fallacious). Jin et al. (2022) collected 2,449 samples of 13 logical fallacies from articles related to climate change and tested 12 existing LMs, attaining a variable performance (F1 score between 25 and 80), dependent on the fallacy type. Other related logical fallacy studies can be found in Musi et al. (2022), Sahai et al. (2021) and Sourati et al. (2022).

[43] https://iep.utm.edu/fallacy/#H2

In the past, members of the AI4TRUST consortium have extensively worked on the above-described tasks. Our partner NCSR-D (Greece) developed several tools related to text analysis, most of which have been part of the *Ellogon* language engineering platform (Petasis et al., 2002). Ellogon is a multi-lingual, cross-platform, general-purpose text engineering environment. Ellogon was designed for managing, storing, and exchanging textual data embedding and managing text processing components as well as visualising textual data and their associated linguistic information. The Ellogon platform contains preprocessing tools for the English and Greek language (HTML cleaning, language identification, tokenisation, sentence splitting, part-of-speech taggers, named-entity recognisers, sentiment analysis, etc.). Ellogon also provides support for annotating and curating corpora (Ntogramatzis et al., 2022). NCSR-D has significant experience in analysing low-resourced languages (such as Greek), and has performed research on transfer learning, and learning from small or imbalanced corpora (Papadopoulos et al., 2023). Finally, NCSR-D has significant expertise in identifying facts, claims and arguments. For example, Sardianos et al. (2015) researched argument extraction to identify segments that represent claims and premises on social Web texts (mainly news and blogs) in the Greek language, with a special focus on politics, economics, culture, various social issues, and sports. The argument mining tools, focusing on the Greek language, have been used to automatically extract argumentation, mainly from news in several European and national projects (Flouris et al., 2022; Ymeralli et al., 2022).

The project coordinator Fondazione Bruno Kessler (FBK, Italy) and its researchers have addressed online disinformation by focusing on two main challenges: the little availability of datasets in languages other than English and the limited robustness of claim verification approaches in terms of domain and genre. They have developed a dataset for the evaluation of claim verification that, starting from the Italian portion of the multilingual dataset X-FACT (Gupta and Srikumar, 2021), has been manually modified to include claims in news-like language, as well as the same statements rewritten as social media posts (Valer et al., 2023). Then, they have implemented a claim verification approach based on semantic similarity between a given input and evidence, which on the one hand reduces the need to collect a large training set for supervised claim verification, and on the other hand allows the classifier to abstain if not enough evidence is found. This preliminary study shows that automated claim verification is still far from being reliably applied in the real-world, and indicates the benefit of models' abstention in case of lacking evidence for verification.

As regards hate speech detection, FBK researchers have extensively worked at this task from different perspectives: for instance, the problem of online religious hate has been addressed creating datasets and a classifier for English and Italian tweets (Ramponi et al., 2022). Also, the problem of biases in hate speech classifiers has been explored (Ramponi and Tonelli, 2022), while a general evaluation of multilingual approaches has been presented (Corazza et al., 2020). More recent applications of generative LLMs to augment training datasets for hate speech detection have also been investigated (Casula and Tonelli, 2023).

GDI (Germany) has developed a new approach to disinformation detection using recent advances in Natural Language Processing (NLP) and Large Language Models (LLM). Embedding text using

LLMs allows for analysis of words and their relationship to each other in a sentence, as opposed to traditional counting-based text analysis techniques. LLM-based detection models are more accurate and more easily adapted to address the ever-evolving landscape of disinformation.

GDI's approach involves the tagging of sentences from previously assessed content which contains narratives that GDI tracks. The sentence tagging process is performed by third party researchers or GDI analysts trained to recognise disinformation. To ensure consistency, all analysts use a codebook that sets rules for the repeatable and measurable identification of potential disinformation. The data from which these sentences are selected is anonymised for domain, author and any other identifying attributes.

The tagged sentences are used to construct digital filters which are then encoded using an LLM. These encoded filters are then used to identify potential disinformation in newly seen content. When new content is analysed, each sentence contained in that content is also encoded using an LLM. This allows GDI to use our machine learning models to determine how "close" each sentence in the article is to the tagged sentence filters. A website is flagged for Manual Review when a significant number of articles on that website contain sentences that match tagged sentence filters.

1.3.2 Visual classification of mis/disinformation

In our examination of the current state of the art in visual classification of mis/disinformation, we first highlight the focus on tools that aid in identifying original videos used in misleading contexts, followed by an emphasis on the development of techniques to identify AI-generated or manipulated videos and images. One type of mis/disinformation, probably the easiest to do and thus one of the most found by journalists and fact-checkers, relies on the reuse of a video from an earlier event with the claim that it shows a recent or even ongoing event. The identification and debunking of such fakes, that are often further disseminated by an unaware public, require the detection of the original video by searching for prior occurrences of this video (or parts of it) on the Web.

Several technologies have been introduced over the last years, to facilitate this task. A couple of them (TinEye[44] and RevEye[45]) allow the user to perform reverse search on still images using the corresponding functionality of online search engines (such as Google), while other technologies (Berify[46] and Videntifier[47]) enable this reverse search only within closed collections of images and videos, thus significantly restricting the boundaries of investigation. The DataViewer of Amnesty

[44] https://tineye.com/
[45]https://chrome.google.com/webstore/detail/reveye-reverse-image-sear/keaaclcjhehbbapnphnmpiklalfhelgf?hl=en
[46] https://berify.com/
[47] http://www.videntifier.com

International[48] extends the online searching capability of the aforementioned solutions, by supporting the reverse search of YouTube videos using a (restricted) set of video thumbnails for reverse image search. Finally, tools such as the "Fake News Debunker by InVID & WeVerify"[49] represent the state of the art in detecting disinformation that relies on reusing old videos to deceive viewers about a recent/ongoing situation. Among other things, the “Keyframes” component of this tool allows users to process a video, extract a set of representative keyframes and use them for performing a reverse search on the Web using a variety of search engines, in order to find near duplicates of the video and spot cases of misrepresentation.

In terms of prior work, the AI4TRUST partner CERTH (Greece) has already developed the web-based technology behind the “Keyframes" component of the “Fake News Debunker by InVID & WeVerify”[50]. In AI4TRUST, CERTH will extend this technology by supporting interaction with multiple search engines. For this, it will automate: i) the submission of searching requests using a number of keyframes (or a small set of representative video thumbnails), ii) the processing of the search results for collecting video items and contextual information (e.g. publication date), iii) the analysis of the collected video items for finding near-duplicates of the query video, and iv) the exploitation of the contextual information of the spotted near-duplicates for classifying the query video as mis/disinformation or not.

Synthetic media is gradually recognised to be among the key risks for disinformation as a result of rapid advances in the field of generative AI. In terms of visual content, there have been many recent developments, often associated with the term “deep fakes”, which, even though starting from a very specific type of digital manipulation (face swapping in particular), it is now broadly (and in most cases inaccurately) used to refer to many types of fully synthetic or digitally manipulated images and videos (Tolosana et al., 2020). The most common types of visual synthetic media include the following:

- Fully synthetic images, most often of human faces: for several years, Generative Adversarial Networks (GANs) and more specifically the StyleGAN family of generators (Karras et al., 2019) has been the method of choice for generating synthetic images. However, as of 2022, Diffusion Models (Rombach et al., 2022) and most commonly publicly accessible models such as Stable Diffusion and Midjourney have become the most popular choice for generating a variety of realistic imagery using text prompting.
- Face attribute manipulation, e.g. modifying the age of a person or adding accessories (e.g. eye glasses, hats) on them, is another popular kind of AI manipulated media, where both GAN and Diffusion Model architectures have evolved a lot and offer numerous capabilities for editing an input image in terms of different attributes or for “interpolating” between two images (e.g.

[48] https://citizenevidence.amnestyusa.org/
[49]https://chrome.google.com/webstore/detail/fake-news-debunker-by-inv/mhccpoafgdgbhnjfhkcmgknndkeenfhe
[50] https://multimedia3.iti.gr/video_fragmentation/service/start.html

starting from an image of a young boy create intermediate versions of images that gradually end up depicting an elderly woman).

- Face swapping has been one of the first kinds of manipulations that popularised the term "deepfake" and aims to replace the face in an original image or video with a selected one. This kind of manipulation has been maliciously used, not only as a means to spread disinformation, but also in the context of image-based sexual abuse (IBSA).
- Face reenactment and lip synching are other very common types of video manipulation that aim at modifying the facial movements and expressions of a target person so that they present them in a specific way, e.g. present a politician as making a specific statement, which they did not.
- There are also other kinds of synthetic media models, for instance, fully synthetic video given a text description (the commercial service Runway ML provides such capabilities) and another emerging area of synthetic media is based on Neural Radiance Fields (NeRF) (Yu et al., 2021). However, neither of those have so far become popular and their use has not been recorded yet in relation to disinformation activities.

Given the variety of synthetic media and the rapid evolution of the field, it is natural that there is an increasing number of approaches for detecting whether an image or video is synthetic or synthetically generated. The survey by Tolosana et al. mentioned above offers a comprehensive overview of the field at the time it was published. In general, the following general trends can be noted with respect to synthetic media detection:

- Despite the large variety of methods in the literature, the most adopted approach is to train deep learning models using one or more of the popular public datasets in the literature, e.g. FaceForensics++ (Rossler et al., 2019), DFDC (Dolhansky et al., 2020), ForgeryNet (He et al., 2021).
- Most used detection architectures are typically based on convolutional networks such as ResNets, EfficientNets and XceptionNets, while recently Vision Transformers (ViT) have been extensively used.
- There is growing consensus that a key issue that detectors face is the generalisation to unseen generative architectures. While there have been some promising steps towards more general detectors (Chai et al., 2020) or by detecting synthetic videos as anomalies compared to the "real" ones (Haliassos et al., 2022), training specialised detectors that are focusing on specific kinds of manipulations seems to be the most practical and effective strategy to date.

In terms of prior work, CERTH has already developed a method for deepfake detection on videos (Baxevanakis et al., 2022), primarily focusing on face swap manipulations, and more recently, they have also developed detectors for fully synthetic images based on GANs and diffusion models (Dogoulis et al., 2023). There are numerous directions in need of further research in AI4TRUST, which according to the current plan include the following:

- Investigation of optimal architectures and setups, including the experimentation with distance metric learning approaches, for improving detection accuracy across benchmark datasets.
- Assessment of performance on cases of occluded faces, and investigation of new approaches that can improve the robustness of the detector.
- Investigation of audio-visual deep learning models that capture the dissonance across the visual and audio modalities for improving video deepfake detection performance.
- Assessment and enhancement of synthetic image detection models by leveraging large synthetic datasets using the methods by UNITN (described below).

As these technologies are designed to identify falsified content and distinguish between mis- and disinformation, it is essential to integrate their results with other tools capable of discerning the intent behind the dissemination of such falsified information.

Another important aspect to consider with respect to the challenges of visual classification of mis/disinformation is the fact that today's society is careful about ethical topics and with the raising of publicly available AI tools concerns about their fairness are also growing. Fairness in this context could be a way to ensure that the generative models are now prone to producing content that can be used for disinformation. In a supervised learning setting, the importance of the training data is well-known since the behaviour of the model at inference time is highly correlated to the seen data. Modern models can effectively learn and highly perform multiple downstream tasks generalising to unseen data. Besides the effectiveness of the pipeline, training data also brings unwanted side effects.

It has been proven that vision datasets contain biases (Torralba et al., 2011), thus the models learn the correlations present in the data which may be malignant (Bolukbasi et al., 2016, Boulamwini and Gebru, 2018, Hendricks et al., 2018, Zhao et al., 2017). In this context, studying the behaviour of deep learning models is crucial to avoid unwanted situations at inference time (Paleyes et al., 2022). Training fair discriminative models has become of paramount importance for the research community during the past years. Recent works have shown that not only do models learn the underlying bias present in the data (Jung et al., 2022, Stock et al., 2018), but they tend to often amplify it (Wang et al., 2020a). Multiple techniques have been proposed for mitigating the bias, from task-specific training, such as the introduction of regularisation terms or architectural approaches (Nam et al., 2020, Savani et al., 2020) to data augmentation strategies (Agarwal et al., 2022, Li et al., 2023). Recently, generative models, such as Generative Adversarial Networks (GANs), have been employed as a data augmentation technique to generate fairer data (Chaudhari et al., 2022, Xu et al., 2018, Xu et al., 2023), to generate counterfactuals (Abroshan et al., 2022, Dash et al., 2022) or to generate counterparts by editing sensitive attributes (Zhang et al., 2023). The above works train generative models from scratch which may be impractical, especially in low data regimes. Additionally, the pre-trained generative models are expected to reflect the bias that is inherent to the datasets where they have been trained on (Xu et al., 2018), challenging those methods that use them for bias mitigation.

Within the consortium of AI4TRUST, the University of Trento (UNITN, Italy) will address the above limitations by investigating an approach that leverages a pre-trained diffusion model (Preechakul et al., 2022, Song et al., 2020)· to edit sensitive attributes in images, e.g., facial images, in order to improve the fairness of existing (biased) datasets and, consequently, the fairness of a discriminative model trained on such datasets. By contrast to the previous works that train generative models (e.g., GANs) from scratch, we will incorporate the power of a fixed pretrained diffusion model to change sensitive attributes from a pool of generated images.

1.3.3 Audio classification of mis/disinformation

The creation of misinformation or disinformation through audio can involve various techniques, such as altering the speed or pitch of recordings, splicing audio clips, or utilising synthetic voice technologies. Our efforts within AI4TRUST are specifically focused on examining the domain of fully AI-generated audio. In this section, we elucidate how the ongoing work of the project can facilitate the identification and mitigation of such synthetic audio content, thereby enhancing our ability to effectively detect and counter audio-based misinformation or disinformation campaigns.

Disinformation (or misinformation) through the audio media is enabled by the ability to synthetically generate audio data. Audio generation techniques are constantly improving (Liu et al., 2023, Kim et al., 2023, Masood et al., 2023, inter alia), and while these advancements have many beneficial applications (such as allowing speech-impaired persons to recover their voices or creating digital art and entertainment content), they can also serve malicious purposes (e.g., cloning voices of celebrities to spread misinformation). A recent example is the deepfake video of Sam Bankman-Fried, in which the well-known entrepreneur and CEO appears to offer compensation to the users affected by the FTX collapse by pointing them to a scam website.

This section presents approaches for synthetic speech detection (also known as audio deep fake detection). This task attempts to prevent misuses of the technology by developing methods that can automatically estimate whether a given audio is real (bonafide) or fake (spoofed). There is a sustained ongoing effort on this task, for example, (Tak et al., 2022; Wang et al., 2022; Zhang et al., 2023). Here we focus on work that targets two properties, which are critical for applying these methods in practice, but many of the current models lack or neglect: generalisability and trustworthiness (i.e., good calibration).

Generalisable detection methods. Since synthesis methods are continuously evolving, it is unreasonable to expect that we will have access to training data similar to that encountered in practice. Generalisation is the capability of a model to perform well on data not seen during training. However, Müller et al. (2022) have recently shown that the generalisation abilities of popular fake audio detectors have been overestimated. They evaluate twelve top-performing detection models and show that none of them can generalise on an out-of-distribution dataset. A possible explanation for the poor generalisation performance are the preprocessing peculiarities exhibited

by the training dataset (ASVspoof, Wang et al., 2020b)—the silence duration (Müller et al., 2022) and the bitrate information (Borzi et al., 2022) correlate with the ground truth. Given that the best deepfake detection models are high-capacity, they can easily learn such low-level, but spurious features. In the context of image deepfake detection Ojha et al. (2022) have shown that strong generalisation results can be by leveraging strong pretrained representations (in their case CLIP (Radford et al., 2021) embeddings). Self-supervised representations have recently also been applied to the speech modality (Wang et al., 2022; Tak et al., 2022, Xie et al., 2023, Kawa et al., 2023, Rosello et al., 2023). Among those Xie et al. (2023) introduce a generalizable method for audio deepfake detection, based on large pre-trained representations and the hypothesis that fake audio samples have a wider distribution within the feature space. The natural samples, irrespective of the out-of-domain data presented to the system, should exhibit a more compact distribution. A separate contribution refers to using an additional loss measure based on triplet mining. The measure enforces the dispersion of learnt parameters for the positive vs the negative classes. Kawa et al. (2023) also address the use of pre-trained features, but base their experiments on the Whisper architecture. They evaluate three common deepfake detection algorithms when trained with these novel features. Their experiments also investigate the generalisation ability of their proposed setup, and show that the Whisper-derived features surpass the standard representations for both the in-domain and out-of-domain metrics. Finally, Rosello et al. (2023) fine-tune a conformer architecture using multi-lingual self-supervised model-derived features. The use of the conformer network enables the method to process variable-length input segments, yet still provide a decision for each four second chunk, like all previous studies.

Well-calibrated detection methods. Deep fake detectors will be used for taking critical decisions, so we want them to produce reliable, trustworthy scores. For example, if the detector outputs several fakeness scores of around 0.7, then we would expect that 70% of the inputs are indeed fake. A classifier that exhibits this property is known as well calibrated. Current research of general machine learning models addresses this aspect (Bhatt et al., 2021; Hüllermeier & Waegeman, 2021; Gawlikowski et al., 2023), but surprisingly little work addresses the topic of calibration of deep fake detectors, either audio or image-based. Recent work (Guillaro et al., 2023, Salvi et al., 2023a) has tackled a related problem of estimating the uncertainty (or conversely certainty) in a prediction. Both papers use a similar method: first train a deep fake detector, then train a second classifier (using a frozen representation extracted from the detector) to estimate whether the predictions of the first are correct or not.

Approach and results. Our methodology (submitted as a paper to ICASSP) addresses the two previously mentioned desired properties: generalisation and calibration. First, we propose to improve the generalisation capabilities by leveraging strong pretrained representations, namely self-supervised representations stemming from the wav2vec 2.0 method (Baevski et al., 2020). We keep these representations fixed and train only the final linear classification layer. Second, we investigate whether it is feasible to use the much more direct method of estimating the uncertainty from the output probabilities of the deep fake detector, for example, by computing the entropy over the outputs. We further evaluate whether the output probabilities can be used for assessing the

predictions' trustworthiness by attempting the task of uncertainty estimation (or reliability estimation).

We evaluate our approach on two publicly available datasets: ASVspoof'19 and In the Wild. ASVspoof'19 (ASV; Wang et al., 2020b) is a common dataset, which we use for both training and testing. This dataset consists of audio coming from 19 different speech synthesis systems (6 systems in the train and dev splits, and 13 in the evaluation split). In-the-Wild (ITW; Müller et al., 2022) is a recently introduced dataset, which we use only for testing purposes to benchmark the out-of-domain generalisation capabilities. ITW is collected from the internet, and is inherently noisy, containing speech artefacts, as well as other sounds. No information regarding the synthesis methods is provided. Moreover, we evaluate the methods from two perspectives: (i) their discriminative power over fake and real samples; and (ii) their ability to produce calibrated predictions. These perspectives are measured by the equal error rate (EER) and expected calibration error (ECE).

Our results show that we can attain state-of-the-art generalisation and calibration performance (0.6% EER/1.8% ECE on the in-domain ASVspoof'19 dataset and 7.9% EER/16.1% ECE for the out-of-domain ITW dataset) by leveraging the 2B-parameter multilingual variant of the wav2vec 2.0 feature extractor, which is the largest, and was pretrained on the most diverse data. Furthermore, we obtain much better uncertainty estimation than the alternative method of Salvi et al. (2023a) of in terms of both the fraction of data selected as reliable and the accuracy on this data. Next, we plan to confirm the results obtained by our method on more out-of-domain datasets (such as, TIMIT-TTS, Salvi et al., 2023b, or FoR, Reimao & Tzerpos, 2019), and investigate whether implicit localisation of partially tampered signals is possible.

1.3.4 Social network analysis

In AI4TRUST, SNA is not included among the AI tools developed in WP3 but rather constitutes an integral part of the activities conducted in WP2, which will be integrated into the preprocessing analysis run automatically on the social media data collected. The primary objective of the social network analysis is to identify indicators of coordinated malicious behaviour, which can be both structural (e.g., identifying the sources disseminating the content) and dynamic (e.g., understanding how the content is being diffused across the social network). The identification of these markers will significantly strengthen the AI4TRUST platform, particularly the Disinformation Warning System, by highlighting potentially unreliable news pieces associated with disinformation campaigns. Further insights into the current landscape of this topic will be provided in Deliverable 4.1, while the methods developed will be described in D2.3, as the endeavour to identify the social dynamics of disinformation is shared between WP2 and WP4.

1.4 Automatic countering of mis/disinformation

Among the objectives of AI4TRUST we also have the automated generation of verdicts, and in particular their adaptation to different social contexts. Based on the idea of "Social Correction" (Bode & Vraga, 2018), we want to build an AI tool able to write potential responses as those that can be found in online social media platforms that are the social-media counterpart of the “journalistic verdicts”, produced by fact-checkers, obtained by using ‘claim+article’ content. These “social correction verdicts” are primarily used to address misinformation, but in principle they can be also used to address concealed disinformation and malinformation. This is particularly relevant when we want to address not only the user who posted the content but also the bystanders that can be misled by deceptive posts.

The idea of providing AI-based suggestion tools is rooted in the assumption that mis/disinformation cannot be fought only by debunking it, but such debunking should be actively disseminated using verdicts on social media platforms to prevent misinformation spreading. Additionally, the sheer amount of deceptive content produced daily is simply too much to be dealt with manually. Thus, it is essential for stakeholders fighting mis/disinformation to be helped with appropriate tools that can make their activity much more effective and efficient (Chung *et al.*, 2021).

Methodologies. For the task of verdict generation, several methodologies have been explored, ranging from logic-based approaches (Gad-Elrabvet al., 2019; Ahmadi et al., 2019) to deep learning techniques (Popat et al., 2018; Yang et al., 2019; Shu et al., 2019; Lu and Li, 2020). More recently, He et al. (2023) introduced a reinforcement learning-based framework which generates counter-misinformation responses, rewarding the generator to enhance its politeness, credibility, and refutation attitude while maintaining text fluency and relevancy. Previous works have shown how casting this problem as a summarization task – starting from a claim and a corresponding fact-checking article – appears to be the most promising approach (Kotonya & Toni, 2020a).

Under such framing, the explanations are either extracted from the relevant portions of manually written fact-checking articles or generated ex-novo; these two approaches correspond, respectively, to extractive and abstractive summarization. Extractive and abstractive approaches suffer from known limitations: on the one hand, extractive summarization cannot provide sufficiently contextualised explanations; on the other, the abstractive alternatives can be prone to hallucinations undermining the justification's faithfulness. Following the idea described in (Kotonya & Toni, 2020a) we extensively experimented with both extractive and abstractive summarization and proposed a pipeline that obtained SOTA results by mixing the two approaches and that is driven by claim content in both steps (Russo et al., 2023).

Data. While the abstractive approach remains the most promising -- also in light of the current advances in LLMs development -- the problem of collecting an adequate amount of training examples persists: the few datasets available for explanation production are limited in size, domain coverage or quality.

The most used datasets are either machine-generated, e.g., e-FEVER by Stammbach and Ash (2020), or silver data as for LIAR-PLUS by Alhindi et al. (2018). To the best of our knowledge, only two datasets include gold explanations, i.e. PubHealth by Kotonya and Toni (2020b) and the MisinfoCorrect's crowdsourced dataset by He et al. (2023). However, both datasets are limited to a specific domain (respectively, health and COVID-19), and only the latter comprises textual data written in an SMPs style (informal, personal, and empathetic if required). This style is very different from a journalistic style, more direct and concise, meant for the general public, but the dataset misses an accompanying news article for each entry that would allow for grounded verdict generation. For this reason we are currently working on a larger dataset that not only accounts for emotions, SMP style and empathy but also provides a background fact-checking article for each entry.

2. AI tools and preprocessing requirements

In AI4TRUST, we will advance and extend a set of AI tools for analysing textual, audio, and visual data to assist the detection of disinformation according to various content use or manipulation scenarios that are typically found in disinformation campaigns (such as the use of hate/emotional/provocative speech, the generation of deep fake image/video/audio content, and the re-contextualization of image/videos through their association with misleading textual descriptions). Given that the focus of the tools we are developing primarily lies in assessing the veracity of content rather than the intention behind its dissemination, these tools effectively identify misinformative content.

In this section, we provide an outline of the AI tools to be integrated into the AI4TRUST platform, along with the generative tools for their development and bias testing. These tools will be further detailed in D3.1, D3.2, and D3.3 of WP3. We also identify the pre-processing requirements within the analysis pipeline of the platform. As discussed in Deliverable 5.4, most AI tools will operate independently on the respective partners' machines, with the platform accessing them through APIs. Here, we delineate the characteristics of data pre-processing necessary on the platform to prepare social and news media data that feed the AI platform for the AI methodologies aiming at automatically identifying disinformation content. Additionally, we offer an overview of the pre-processing carried out on the partners' servers.

In the following Table 2, we provide a list of the AI4TRUST AI tools that will be used for assessing the trustworthiness of the collected content and detecting disinformation.

Table 2. List of AI tools for data analysis and disinformation detection

Technology	Description
Speech to text	This technology transcribes an audio file (or the audio track of a video file) including speech, into text. Its output will be processed by the text analysis methods of AI4TRUST.
Hate speech detection	Given a post on social media, this technology outputs a label indicating if the post contains hate speech or not; hate speech spreads, incites, promotes, or justifies hate against a person or group of people due to characteristics that they share (e.g., race, religion, gender) and is frequently found in disinformation campaigns.
Disinformation detection	Given a post on social media, this technology outputs a label indicating if the post potentially contains mis/disinformation or not; it can be useful for fact-checkers allowing to spot possible cases of fake news online but it needs human supervision to confirm the presence (or absence) of mis/disinformation.
Disinformation	Given a post on social media containing mis/disinformation and an article that

countering	debunks this post, this technology outputs a text that shortly explains why the post was classified as mis/disinformation; it can make fact-checkers more efficient in countering the spread of fake news with proper explanations rather than simply claim that "this is not true".
Document social intelligence and contextualization	This is a suite of tools that is able to analyse textual input (news, posts, tweets, comments, etc.) and identify clues such as semantic indicators of identifications to reference groups, of authorities, of normative choices or beliefs.
Document Intelligence Level 1	This is a suite of tools that is able to analyse textual input (news, posts, tweets, comments, etc.) and identify clues such as, emotional text, provocative information, hate speech; given some textual input, this technology labels specific segments of the text according to the existence of each of the aforementioned clues.
Document Intelligence Level 2	This tool analyses textual input and identifies information that seems unsupported, such as arguments that are not supported by premises; it can be used to spot information that is presented as factual but without supporting evidence, thus indicating the need for fact-checking or verification.
Document Intelligence Level 3	This tool analyses textual input and tries to identify some common logical fallacies in it, such as slippery slope, red herring, etc.
Reverse video search on the Web	This tool segments an input video into fragments, extracts a set of representative keyframes and uses these keyframes for reverse image search on the Web; it can assists the detection of duplicates of a given video on the Web and the identification of cases where an old video has been re-used to mislead viewers about a contemporary event.
Sensational content detection	Given a video file, this technology outputs a score representing the probability that the video contains sensational (shocking, scary, exciting) content, which is usually part of disinformation campaigns
Deepfake audio detection	Given an audio file, this technology outputs a score indicating the probability that the audio file was artificially generated, either entirely or partially.
Deepfake image/video detection	Given an image or video file, this technology outputs a score representing the probability that this file was generated using a given set of known deepfake generation models; in the case of videos, such scores are produced at the shot-level helping fact-checkers to localise the deep lake in the video.
Visual-text misalignment detection	This technology evaluates pairs of images/videos and their textual descriptions, to assist the detection of cases where an image/video has been re-framed or re-contextualized to mislead the viewers about an event
Textual sentence level analysis	This technology processes an article at the sentence level and classifies it as disinformation, based on the existence of adversarial narratives within the text.

During the initial months of the project, we conducted an internal and exploratory survey across the consortium to assess the perceived necessity for these tools among the various user categories represented within. This survey was completed by 16 media experts, 11 fact-checkers, and 19 researchers. In Table 3, we present the average survey results, based on a rating scale ranging from 1 to 5.

Table 3. Result of the internal exploratory survey on the usefulness of AI tools

	Media experts	Fact-checkers	Researchers
Speech to text	4	4.6	3.2
Hate speech detection	3.9	3.4	3.4
Textual Disinformation detection	4.4	4.7	4.5
Highlight text segments as hateful or provocative	3.9	3.6	3.6
Highlight text segments as unsupported information	4.1	4.1	3.9
Highlights text segments as logical fallacies	4.3	3.7	3.5
Disinformation countering	4.2	4.5	4.2
Reverse video search on the web	4.2	4.9	3.7
Sensational content detection	3.6	3.5	3.1
Deepfake audio detection	4.5	4.5	4.3
Deepfake video detection	4.5	4.8	4.3
Visual-Text misalignment detection	3.7	4.1	3.7
Textual sentence level analysis	3.7	3.9	3.8

Apart from the AI tools listed in Table 2, in AI4TRUST we will build a set of generative AI technologies that are shown in Table 4. These technologies will not be exposed to the users of the AI4TRUST platform. They will be utilised solely for generating or manipulating data of different modalities (i.e., images, videos, speech/audio, as well as pairs of image/video and textual information), that will be used for training and evaluating our AI tools for deep fake/manipulated image/video/audio content detection, and for video-text misalignment detection.

Table 4. List of Generative AI technologies that will be used for assisting the training and evaluation of various AI tools for data analysis and disinformation detection

Deepfake audio generation	This technology will be used to generate a natural and expressive audio recording of a speech, given an input text and a speaker identity and while using only a small amount of data about the target speaker.
Person Image Generation	This technology will be used to generate realistic images showing persons, by editing various facial attributes, such as the skin colour, the hair style and colour, the age, and the beard or moustache.
Semantic-Guided Scene Generation	This technology will be used to generate realistic images showing a scene, by applying object-level editing of the visual content according to a textual prompt that indicates the semantics of the targeted object.
Playable Video Generation	This technology will be used to generate realistic videos by selecting a series of discrete actions that should be shown at every time step of the video.

2.1 Preprocessing of textual content

Since different models often require different preprocessing, we first aim to keep a version of the text that is as close as possible to its raw version in case new data needs arise. The only requirement for this version is that individuals may not be directly identifiable from reading it. That is why the unique identifier of the users defined by the platforms will be hashed, using, for instance, the HMAC algorithm with a random generated secret[51]. This way, all mentions of other users, email addresses or other identifiers in the text can be removed and replaced by these new unique identifiers, which do not allow direct identification of the actual individuals.

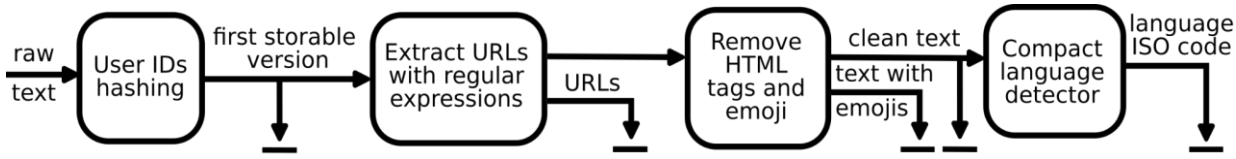
Once this first version has been generated and stored, we may proceed with further preprocessing. A common requirement that the models have is to remove any content which has not been generated by the users themselves. URLs will be extracted to a separate data field using regular expressions, since they do not contain linguistic content in themselves. We will also remove HTML

[51] https://docs.python.org/3/library/hmac.html

tags using regular expressions, and in particular leveraging the implementation of the Trafilatura[52] Python library. Emoji and other special characters will be removed for the input of some models, but will be kept for others as they may hold useful semantic content.

Finally, we need to detect the language the text was written in, if not given by the platform. To that end we can use the version of the Compact Language Detector made freely available by Google[53]. After this step, a new field should be present in the data in the form of a language ISO code, or, if the language detector does not output a language with high-enough confidence, a null value.

Fig. 2. Schematic summary of our textual preprocessing pipeline. Stored data are visually represented by an arrow pointing down to a horizontal line



2.2 Visual classification of mis/disinformation

The AI tools for image and video analysis will be compatible with most of the known formats of image (i.e., jpg, png, tiff, bmp) and video (i.e., mp4, webm, avi, mov, wmv, mpg, flv, mkv) files. Images and videos of high resolution and low compression rate should be preferred, in order to facilitate the analysis and enable the production of more reliable results; especially by the developed AI tools for deepfake image/video detection. It should be noted that below a certain resolution and quality level, deepfake detection is highly unreliable so it is suggested that a check is performed prior to submitting an image/video for analysis and in case it does not meet certain criteria, the end user should be made aware that this is not an acceptable image/video for further analysis.

These tools will be hosted in the processing servers of the associated technology-providing partners, and exposed to the AI4TRUST platform through APIs. The analysis of a given image/video will only require the provision of the URL that points to the image/video, and any required data-preprocessing steps (such as video-to-frame decomposition, frame sampling, frame resizing, etc.) will be performed at the server's side. The output of these tools will be provided to the envisioned Disinformation Warning System – a core system of our platform aiming at identifying what are the

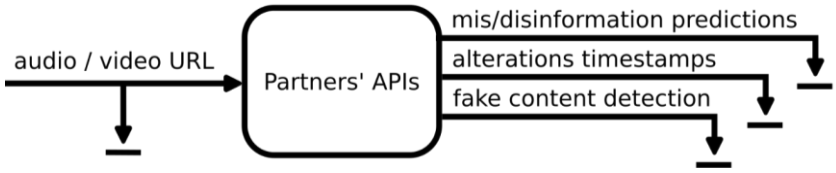
[52] https://trafilatura.readthedocs.io
[53] https://github.com/google/cld3

possibly disinformation narratives that our fact-checkers will have to verify and eventually debunk – and stored in the AI4TRUST platform, to allow further use by and integration with other components of the platform.

2.3 Audio classification of mis/disinformation

The developed tools will focus on using low-quality speech data (i.e. 16 kHz sampling rate, 8 bits/sample) such that the audio-based classification of mis/disinformation content can be performed on any audio-visual content provided by the stakeholders. The classification method will be exposed as an API platform in which the end-users will be able to upload audio samples, as well as to provide link for audio or audio-visual data from the most common online platforms (e.g. YouTube, Facebook, TikTok, etc.). The links will be processed by the API, the audio content retrieved, resampled, and then passed through the audio mis/disinformation system. Its response will entail a sample-level prediction of fake/non-fake content, as well as, in line with the project's future developments, the localisation of modified/altered/manipulated subsegments of the audio data. If needed, the processed audio content, original link, and predictions will be stored in the central platform for additional analysis and processing.

Fig. 3. Schematic summary of our audio and visual preprocessing pipelines. Stored data are visually represented by an arrow pointing down to a horizontal line.



2.4 Social network analysis

In the AI4TRUST framework, the incorporation of Social Network Analysis assumes a crucial role within the operations of WP2. Its primary objective is to automatically examine the dissemination patterns of unreliable content extracted from social and news media data. This analysis is focused on identifying orchestrated malicious activities, tracing the origins of content dissemination, and comprehending how information propagates across the social network. As previously highlighted in section 1.3.4, the recognition of these indicators substantially strengthens the AI4TRUST platform, particularly reinforcing the Disinformation Warning System as through these grounded

methodologies it is possible to shed light on the diffusion mechanism behind the spread of misinformation and disinformation.

2.4.1 Social network analysis as a “preprocessing step” in the platform

Social network analysis is not categorised as an "AI tool" and consequently does not form part of the activities conducted in WP3. This distinction arises from the fact that the associated methodologies primarily rely on the development of quantitative indicators rooted in both mathematical and social science theories, rather than the training of machine learning algorithms. Consequently, the endeavour to develop innovative social network indicators and insights into the current landscape of this domain will be a collaborative effort between WP2 (T2.5) and WP4, specifically within the parallel Task 4.2, "Mapping of social production of misinformation." Notwithstanding, such indicators may improve and feed SNA features to the AI-based dis/misinformation detection techniques developed within the project.

The discussion on the socio-contextual basis fostering the spread of mis/disinformation will be further elaborated in D4.1 of WP4, while an extensive exploration of the technical aspects of the Social Network Analysis will be conducted in D2.3 of WP2. Here, we delineate the key steps for the intended network analysis of a generic social media source, recognizing that this framework may need adjustments to accommodate platforms other than Twitter/X, which has historically served as the benchmark data source for such studies. The recent policy changes implemented by Twitter/X have necessitated a reassessment of the data collection tasks. If, for instance, we were to transition to channel-based platforms like Telegram, the nature of social network (or hypernetwork/social group) data would manifest differently. It is imperative to conduct further investigation into this shift before formulating more detailed data processing recommendations and requirements. D4.1 will explore the inherent disparities associated with diverse social network sources, elucidating how the subsequent social steps will be tailored to the respective ontologies.

Regardless of the specificities of networks that can be traced with respect to different platforms, networked patterns of circulation of unreliable contents can be examined within a common analytic framework composed of three main blocks:

a) Initial Steps in mis/disinformation dissemination analysis:

- Identify the social groups and subnetworks at the root of mis/disinformation dissemination.
- Gather comprehensive social and content data in the surrounding groups and networks, in line with the requirements for text processing set out in section 2.1.
- Analyse the corresponding social network structure and socio-semantic network structure i.e., in terms of sets of staging of reference groups, authorities and beliefs, or norms, or claims associated with surrounding alters or groups.
- Evaluate the diversity of information sources and elements these groups engage with, including negative links.

b) Behavioural pattern identification for misinformation spread:

- Identify behavioural patterns of misinformation spreaders, including super-spreaders, social bots, but also casual spreaders.
- Take into account successful as well as unsuccessful dissemination events.
- Explore cohesive subsets of users and their role in disseminating misinformation within echo chambers and polarised communities.
- Analyse local and broader positioning of groups within the wider system of misinformation propagation e.g., within the constellation of cohesive groups or groups of groups.
- Identify groups sharing similar social intelligence appropriateness judgments facilitating mis/disinformation spread by sharing implicit contextualization.

c) Longitudinal and multilevel behaviour observation and network analysis:

- Observe how actors modify, update, and rewire their affiliations over time in response to multiple mis/disinformation dissemination events.
- Focus on how actors may aim at reducing uncertainty and adjust their understanding over time, for example by changing their beliefs or normative choices through longitudinal data analysis.
- Gather data over a specific time period to map local communities within their broader context.
- Observe how a multilevel, stratigraphic organisation of controversies emerges in the aggregation of individual mindsets into collective mindsets of groups talking at each other. Dis/misinformation dissemination could be facilitated by synchronisation and progressive within-group alignments between levels.

The first two items primarily focus on analysing individual events related to dis/misinformation dissemination. In contrast, the last pillar takes a longitudinal approach, examining behaviour across multiple events, making it more actor-centric at several levels. To support these efforts, comprehensive data regarding who is disseminating what, including social links, content, and longitudinal information over a specified timeframe, is necessary to effectively map local communities within their broader context. To this end, we need to be able to unambiguously assign content to nodes over time: we expect this task to require additional ethical scrutiny in terms of making sure that no personal data is being used in contravention of the applicable law and the good practices implemented across the project

2.4.2 Preprocessing necessary for social network analysis

Within the AI4TRUST platform, the analysis of social media and social networks represents a critical "preprocessing" phase. Simultaneously, the data sourced from social media necessitates processing before any analytical endeavours can take place. Typically, Social Network Analysis indicators rely on the unbiased examination of well-defined mathematical networks, comprising

nodes (representing actors), and edges (reflecting the relationships between these actors) or hyperedges (reflecting higher-order relationships). Consequently, the foremost preprocessing task involves reconstructing this network from raw data from which edges (or hyperedges) may be extracted.

Adapting the data preprocessing techniques to the distinct characteristics of various social media platforms such as Facebook, TikTok, and YouTube is imperative for the accurate cleansing, organisation, and arrangement of information prior to analysis. Considering that each platform may possess its own data structure and attributes, the preprocessing procedures may vary. Nonetheless, the following represent some fundamental procedures to be considered:

1. Data collection and storage:

- Define user and group perimeters enabling a systematic collection of node and content data related to mis/disinformation target keywords
- Gather data using APIs or web scraping tools specific to each platform.
- Store the data in a structured format, such as csv, json, parquet, for easier processing.

2. Data cleaning:

- Remove duplicates: Social media data often contains duplicate records. Identify and remove them to maintain data integrity.
- Handle missing data: Address missing values by either imputing them or removing incomplete records.
- Correct errors: Correct any obvious data errors or inconsistencies.

3. Removing irrelevant content:

- Filter out irrelevant content that does not contribute to our analysis, either for ontological reasons (e.g., irrelevant meta-data) or for semantic reasons (typically, outside of topical scope, user/group perimeter or temporal boundaries).

4. Privacy and ethical considerations:

- Ensure that we adhere to privacy regulations and ethical guidelines when working with social media data.
- This topic will be further addressed in Section 4.

The specific preprocessing steps may vary depending on our research or analysis goals. It is important to tailor our preprocessing to our specific needs and be aware of the limitations and potential biases in the data collected from social media platforms.

3. Topics selected and keywords

In this section, we introduce the three topics selected for analysis in the AI4TRUST platform and provide an overview of the adopted approach to single out instances of information disorders. Such approach passes primarily through the employment of a "living list" of keywords (i.e., a list in continuous evolution) that contains terms and expressions that have been found to be frequently employed within mis/dis/malinformation dynamics in these three thematic areas. For illustrative purposes, the current version of the lists displays words in English, while terms and expressions in the other languages included in our project (French, German, Greek, Italian, Polish, Romanian, Spanish) can be found in the Appendix.

We first discuss the selection criteria that lead us to the topics selected - Climate Change, Public Health, and Migrants - to develop and test the AI4TRUST platform, with an aim to outline a selection framework that will be replicable in the future with respect to other topics. We also outline the procedures necessary to actively mine these topics from social and news media, which will be further elaborated in D2.2 0f WP2. Subsequently, we present the current version of the "living list" of keywords in English commenting on the logic through which it has been drafted.

3.1 Topic selection criteria

We devise a set of selection criteria that intend to provide a topic-neutral replicable schema. The chosen topics appear as illustrative samples of this schema, which can be adapted and repurposed either for derived sub-topics or for entirely different themes. Although users should be able to inspect any topic, the initial selection should heed not only the needs of the users but also the requirements of the developers and researchers. In this case, it is fundamental to train the models on socially relevant and broad topics that cut across different social and cultural contexts but also exhibit sufficient stability across a theoretical timeframe that coincides with the development of the analysed spaces, namely online social media platforms and information outlets.

Criteria for topic-selection rationale:

- **Scope:** The topic should be as generic as possible but lend itself to segmentation into specific derived sub-topics. Marginal narratives (that could become mainstream) should be identified within the context of the broader topics. Not all instances of misinformation are part of a coordinated disinformation campaign. Likewise, a disinformation campaign on certain topics does not resort to all pieces of specific misinformation circulating about the general topic. For example, false information circulating about vaccines should appear as an artefact of the more general discussion on vaccines.
- **Relevance:** The topic should be socially relevant across the languages, geographic and online spaces covered by the project. Many issues, from climate change to vaccines,

women's rights, migration, LGBTQ+ rights etc, have become recurrent topics in the transnational public sphere.

- **Historicity:** Selected topics should not correspond to discrete stories (or events) but to continuous discussions traceable in time and space. Providing researchers have appropriate access to longitudinal data, this historicity will allow us to study not only the present dynamics of mis/disinformation but also the evolution of the phenomenon. This is a vital criterion that determines the selection of certain topics (e.g. climate change, migration) over others.
- **Falsifiability:** The stories and narratives contained within the topic should be reasonably falsifiable. By falsifiability we mean establishing the validity of a truth claim being disseminated online. This is not akin to Popperian scientific validity but to the factual accuracy of the claim. For example, even though we cannot ascertain the exact effects of climate change, there is a scientific and social consensus regarding the anthropomorphic causes of climate change that does not require our own scientific or forensic analysis.
- **Feasibility**: The feasibility of all stages of the project, from data collection to processing and analysis, depends on the existing technical capacity, linguistic knowledge and academic expertise to address the selected topic. For example, while the consortium can build on a strong track record on health-related disinformation by some of the partners, it has neither the expertise nor the linguistic knowledge to address a whole range of highly prominent topics (e.g., War in Ukraine; War in Yemen, Palestine-Israel conflict)

3.1.2 From text to context

To understand the dynamics of mis/disinformation diffusion, we must look at the messages, the actors that (re)produce them, and the spaces of interaction. Technically speaking, this poses a challenge and an opportunity to automated processes. A toolkit focused on discrete units of content (e.g. texts, images) can successfully detect instances of mis/disinformation but will not be able to fully address the problem of intentional and/or coordinated dissemination of incorrect information (disinformation) at scale. For that to happen we should focus not only on discrete units of produced content but also on contexts of reproduction. The units of analysis to identify disinformation are not texts but contexts, not isolated post but semantic networks of analogous false narratives, not individual actors but interconnected actors with different positions and roles that form dissemination clusters, not specific social media platforms but the links across the information ecosystem begging for multimodal analysis. A tool that intends to address both misinformation and disinformation must thus look at social dynamics at scale, resorting to network visualisation and analysis tools that combine machine-driven and human-led processes. Providing that there are humans in the loop with basic network literacy at all stages of the process, from preprocessing to analysis and use, the developed tool can aspire to identify and counter not only misinformation instances but also disinformation operations. This feature is also determinant from the end-user perspective. Prospective users will want not only to detect misinformation but, fundamentally, to map how disinformation is spreading and who are the main actors and clusters of actors driving the circulation of the false narratives.

3.2 Keywords seeds

In practical terms, the effort of tracking and mining instances of information disorders can be done in three ways: 1) tracking a set of keywords; 2) tracking a set of users or creators; 3) tracking a set of groups. In this report, we focus on the first way, as it represents the most direct way for filtering content on online social media's API. Starting from a set of filter keywords (1), it is then possible to identify the actors (2) that use these keywords or engage with content where these keywords are used, and further identify the groups (3) to which these actors belong.

In order to start mining out the aforementioned contexts of a generic social media such as YouTube, Facebook, Twitter/X or TikTok without any prior knowledge, it is advisable to adopt a content-based approach that starts from a wide array of keywords adapted to the specific topic and language we are working on. The working premise here is that a broader list of keywords (i.e., terms and expressions often employed in mis/dis/malinformative content) ranging from broader to more specific - even slang - terms, has more value than a specific one as it allows to scout more extensively public and group conversations unfolding on digital platforms in search for dysfunctional items that exhibit both typical and non-traditional (yet significant) markers of information disorders.

As the practices sustaining mis/dis/malinformation dynamics are ever-evolving, these extensive lists cannot be thought of as fixed but, rather, need to be conceived of as "living lists" of keywords that are constantly refined in two, complementary ways:

a) by removal: evaluating the relevance of results extracted from public API based on the keywords in the list and filtering out a posteriori those that are irrelevant together with the keywords strictly leading to them;
b) by addition: leveraging on patterns of co-occurrences between keywords and other terms/expressions in the same piece of content to identify neologisms used in disinformative context to hide malicious intentions.

The identification of optimal update procedures necessary to identify key actors or groups, and track the appearance of new search keywords to be included in the API filtering task, is part of the research tasks of our project. Here below, we list a set of proposed keywords in English which can be helpful to identify instances of mis/dis/malinformation with respect to the three topics of interest we selected for the AI4TRUST platform. This list has been generated by integrating contributions from participant teams based on extant research in the different domains under investigations, our media experts, fact-checkers, and social media researchers with lists used by GDI for their internal activities (and a few further keywords suggested by ChatGPT). These lists must necessarily be translated in the other 7 languages of our project, and adapted to the media, social and political environments where these languages are spoken. For this purpose, in the annex we offer a first attempt of translating and integrating these keywords performed by our Media and Fact checking partners. Also, as explained above, the lists will be refined during the project activities.

3.3 Topic: Climate change

Climate change is being selected as a topic as frequently targeted by disinformation campaigns due to its polarising nature and significant implications for various industries and policies. False narratives often seek to downplay the severity of the issue, create doubt about scientific consensus, or promote alternative agendas, making it challenging to implement effective environmental policies and initiatives.

3.3.1 Proposed Keywords (English)

Adaptation strategies, Air pollution, Biodiversity loss, Carless cities, Carbon dioxide emissions, Carbon emissions, Carbon footprint, Carbon pricing, Carbon sequestration, Chemical spill, Clean energy, Clean energy technologies, Climate activism, Climate action, Climate adaptation, Climate change, Climate change adaptation, Climate change hoax, Climate crisis, Climate education, Climate finance, Climate impact on ecosystems, Climate justice, Climate mitigation, Climate modelling, Climate policy, Climate refugees, Climate resilience, Climate science, Climate variability, Copernicus, Deforestation, Droughts, Eco-friendly, Emissions, Environment, Environmental activism, Environmental disaster, Environmental impact, Environmental sustainability, Erosion control, Extreme weather, Extreme weather events, Forest fires, Fossil fuels, Friday for Future, Geoengineering, Glaciers, Global temperature, Global warming, Greenhouse effect, Greenhouse gases, Greta Thunberg, Habitat conservation, Habitat destruction, Heatwaves, Ice melting, Marine conservation, Microplastics, Mitigation measures, Natural disasters, Net-zero emissions, Ocean acidification, Overfishing, Ozone depletion, Paris Agreement, Photovoltaics, Rainforest, Recycling initiatives, Renewable energy, Sea level rise, Sustainability efforts, Sustainable development, Sustainable energy, United Nations Climate Change Conference, Waste management, Water pollution, Water scarcity, Weather, Weather modification, Wind turbines, Wildfires.

3.4 Topic: Public health

Public health is frequently targeted by disinformation campaigns, especially concerning topics like vaccines, treatments, and disease origins. Misinformation can erode trust in healthcare systems, lead to harmful behaviours, and hinder effective disease management, emphasising the critical need to combat falsehoods for the well-being of communities and individuals.

3.4.1 Proposed Keywords (English)

5G, 5th gen wireless, 5th-generation wireless, Alien, Alien Agenda, AlienLeaks, Anthony Fauci, Big Pharma, Bill Gates, CDC, CHD, COVID, COVID-19, COVID vax, COVID19, Dr Bryan Ardis, DNA, Emergency, FDA, George Soros, Johnson & Johnson, Marburg virus, Moderna, Morgellons, Pfizer, RNA, Ron Watkins, SARS-CoV-2, Transhuman, Transhumanism, UAP, UFO, UFO report, Unidentified Aerial Phenomena, adrenochrome, adverse events, adverse reaction, antibodies, anti-vax, anti-vaxx, anti-vaxxer, antivaxx, antivaxxer, antibodies, antibody, antivaxx, antivaxxer, AstraZeneca, autism, bad medicine, bad science, bells palsy, big pharma, big tech, big tech censorship, BioNTech, blood, blood clots, blood sample, cancer, cancer cells, cardiac arrest, cardiac problem, ccp bioweapon, ccp virus, chemtrails, chemotherapy, chemical abortion, chemical-free, childhood vaccine schedule, chlorine dioxide, chondroitin sulphate, chronic fatigue syndrome, cleveland clinic, clinical trials, clotshot, codemonkeyz, contact tracing, corona, corona virus, coronavirus, covid, covid passport, covid19, cures, cure, dangerous medicine, dangerous product liability, death, death jab, death rate, death shot, depopulation, detox, dna, drugs, electromagnetism, epidemic, epidemic curve, face mask, foetal tissue, foetus, foetuses, fifth generation wireless, fifth-gen wireless, fifth-generation wireless, flu, forced vaccines, gene therapy, genes, graphene, green arrivals, green list, guinea pigs, health education, health risk, heart attack, herd immunity, homeopathic, horse paste, incubation period, immune system, immunization, infection, injections, inoculation, intensive care unit, jab, jabgate, janssen, kill shot, lateral flow test, lethal injection, let the bodies hit the floor, long-Covid, malaria, mask mandates, masks, masks are for slaves, measles, medical abortion, medical tyranny, medicines, meghathakur, millimeter wave, mitochondria, monkeypox, monkey pox, mortality, mRNA, natural medicine, nuremberg 2, nuremberg 2.0, nuremberg code, nuremberg trials for covid, no jab, no kids vaccinated, pandemic, patient zero, pentagon ufo report, pharmaceutical companies, pills, plandemic, political agenda, prevention, protein spike, pro-SAFE vaccine, prochoice, pro-life, prochoice, prochoice, prolife, prolife, radiation, reaction, red arrivals, red list, reversal pill, robert malone, rochelle walensky, roe v wade, sanitation, self isolation, side effects, snake blood, snake venom, south african variant, spike, spike protein, stop state genocide, study, sudden death, surgical face mask, the great culling, toxin removal, transhuman, unvax, unvaxxed lives matter, unvaccinated, vaids, vaccine, vaccine adverse reporting system, vaccine choice, vaccine deaths, vaccine injuries, vaccine mandates, vaccine side effects, vaccine victims, vaccination, vaccinations, vaccines, vax, vax death, vaxx, watch the water, wuhan, yellow fever.

3.5 Topic: Migrants

Migrant-related issues often fall prey to disinformation, leading to the perpetuation of stereotypes and biases that can fuel social tensions and hinder inclusive policymaking. Misinformation about

migrants can exacerbate divisions, hinder integration efforts, and undermine the promotion of diverse and inclusive communities.

3.5.1 Proposed Keywords (English)

Asylum seekers, attack Europe, banderisation, borders, child migrants, civilization threat, danger civilisation, displacement, economic immigrants, economic migrants, Economic impact of migration, europhobie, greece_defends_europe, greece_under_attack, great replacement, great remplacement, illegal, illegal alien, illegal aliens, immigrant, immigrants, immigration, Immigration law, Immigration policies, Integration programs, invasion, invasion Europe, invaders, invade, migrant danger women, migrant killer, migrants, Migration and development, Migration management, Migration patterns, migratory invasion, Migrant children, Migrant detention, Migrant health, Migrant rights, Muslims, rape migrant, rape refugee, rapefugees, refugee camp, Refugee crisis, Refugee resettlement, refugee danger women, relocate, remigration, resettlement action, social care, Social inclusion, terrorists, threat Europe, threat west, threat western values, white genocide, Xenophobia.

(To this list can be added specific nationalities, such as "Ukrainians" or "Syrians", or indication of origins, such as "Africans" or "Arabs" depending on the period and country of analysis)

3.6 Intersectional perspective

As argued above, attention has gone towards the element of intentionality when it comes to distinguishing mis/dis/mal-information, albeit the debate on how to map and trace it remains wide open. Admittedly, less attention has been paid to the element of harm which inevitably flows from informational disorders of all natures and configurations and irrespectively of intentionality. In a deeply mediated environment like the one we live in today (Hepp 2020), public digital narratives that enclose biased, stereotyped or even tactically framed characterizations of individuals, subjectivities, groups, social and political dynamics, and processes are harmful to the extent to which they contribute to reinforce and, possibly, augment existing patterns of inequality, discrimination, invisibilisation, and exclusion (e.g., Felmlee et al.2020, Sobieraj 2020).
As it aims to address seriously the harmful implications of mis/dis/mal-information, AI4TRUST adopts an intersectional perspective, starting from the assumption that informational lacks that characterise and, at the same time, are fed by information disorders build on systems of intersecting axes of discrimination (Crenshaw 1989, Hill Collins 2019) based on race, gender, sexual orientation, class, religious affiliation, ethnicity and geographical provenance, physical and mental ability, bodily configuration, etc. . Intersectionality is a conceptual framework, theorised by Professor Kimberle Crenshaw, which outlines individuals can be targeted on the basis of more than one social identity (race, religion, disability, gender identity, sexual orientation, class…) playing simultaneously into dynamics of multiple systems of oppression.

In this sense, AI4TRUST flanks the identification of dysfunctional informational dynamics with a thorough attention for the contents that spread through them paying particular attention to how multiple elements (for example, attention to both gender and racist connotations) concur to creating harm and for whom. In doing so, the project strives to uncover how informational disorders contribute to de-democratization trends by fostering not only political polarisation (Tucker et al. 2018) but, more radically, the polarisation of social identities along normative cleavages that feed deep antagonism and adversarial dynamics. Finally, AI4TRUST consortium partners recognise the importance of adopting an intersectional framework throughout the overall project.

4. Legal, ethical and security compliance

AI4TRUST is a multi-disciplinary project that addresses the challenges of supporting media professionals (i.e., fact-checkers and journalists) and policymakers in tackling disinformation, misinformation and malinformation. Respect for legal, ethical and security requirements is embedded in the project. The project approach to this cross-project theme has already been outlined in the first instance in deliverable D1.2 “Data Management Plan”. This section presents the real-world evidence of the AI4TRUST appreciation of challenges and opportunities, and approach to realising legal, ethical and security practice for addressing information veracity for social media and media.

One of the central concerns of AI4TRUST is how new technologies and AI can be used in practical ways compatible with the law, privacy, ethical and security requirements. AI4TRUST underscores the necessity of considering the inherent risks of automated decision-making when creating and advancing algorithms and systems. The project stresses the importance of integrating privacy by design (PbD), ethics by design (EdB), and security by design (SdB) principles from the outset, and suggests conducting an open review at the beginning (more specifically, this will take place in WP4 and WP5) to proactively recognize and address potential risks and precarious situations. This proactive approach aims to prevent and mitigate unintended harm and errors that could impact decisions made by machines or humans at a later stage.

The AI4TRUST project places a strong emphasis on incorporating privacy, ethics, and security into the design of its tools, particularly when assessing how they might affect shared values and objectives. In this context, the consortium will consider the following key factors:

1. **Understanding the impact:** Before developing AI4TRUST tools, it's crucial to comprehensively understand how they may affect shared values and goals. This involves a thorough analysis of the potential consequences of the tool's application, both positive and negative, with a focus on protecting EU values, rights, democracy, and human society. This understanding should guide the tool's development and deployment.
2. **Data selection and use:** The selection of data for training and scaling AI4TRUST tools is a critical factor. Ensuring that the data is representative and unbiased is essential to avoid perpetuating biases and ensuring that the tool respects the dignity and autonomy of individuals. It's important to carefully curate and review the data used, taking into account ethical considerations.
3. **Ongoing data review:** Data collected from the Web should be continuously reviewed to ensure it aligns with the intended purposes of the tool. This review process helps in maintaining data quality and relevance, minimizing potential biases that can emerge over time, and upholding ethical standards.
4. **Alignment with international, EU, and national shared values:** Throughout the development and deployment of AI4TRUST tools, it is imperative to continuously assess and align the project

with shared values. This ensures that the tool remains consistent with the principles and goals set out to protect EU values, rights, democracy, and human society.

5. **Avoiding bias:** Guarding against bias in AI systems is critical. It requires regular monitoring, testing, and adjustments to the tool's algorithms to minimize any discriminatory or unfair outcomes. Bias can emerge in various stages, including data collection, model training, and real-world application.
6. **Respect for human dignity and autonomy:** AI4TRUST tools should be designed and deployed in a way that respects the dignity and autonomy of individuals. This involves ensuring that the tool does not infringe on privacy, dehumanise individuals, or undermine their decision-making capabilities.
7. **Piloting and scientific research:** Testing these tools in a controlled scientific research environment is a good practice. It allows for the refinement and validation of the technology without immediate real-world consequences. This stage is essential for improving the tool's performance, reducing biases, and enhancing its alignment with the project's goals and shared values.

In summary, the development of AI4TRUST tools should be a meticulously planned and executed process, considering not only their technical aspects but also their societal, ethical, security and legal implications. Regular assessments, reviews, and adherence to shared values are vital in ensuring that these tools effectively protect EU values, rights, democracy, and human society.

Hence, the project's objective is to align with the EU AI Strategy, which seeks to position the EU as a global leader in AI, with a strong emphasis on ensuring that AI remains human-centered and reliable, serving the interests of individuals and contributing positively to society. In this regard, the Vice-President of the EU Commission Věra Jourová at the "Fighting Misinformation Online Event" on 26 October 2023 said: "*If designed and used in accordance with democratic systems and fundamental rights, AI systems can become a central technology to support the work of professionals [...]. Our priority is to enforce the Digital Services Act and work with the signatories of the Anti-Disinformation Code to make the online space safer, more transparent and more accountable*".

Recognising the societal impact of using new technologies and AI for the fight against so-called fake news and the wider challenges to fundamental rights, AI4TRUST actions in this regard, which will be contained in deliverables D4.2 'Explainability and Transparency report and AI tools' and D4.3 'Final Explainability and Transparency report and AI tools' of WP4 as well as in the deliverables of WP5 concerning the release of the different versions of the AI4TRUST platform, will enable the recommendations informed by the project evidence to be transformed into feasible, workable, relevant operational guidance that will be fit-for-purpose, necessary and proportionate.

4.1. Responsibility for privacy, ethics, and security in practice

The AI4TRUST partners are committed to considering and incorporating the principles of privacy by design, ethics by design, and security by design throughout the project. Firstly, privacy by design entails integrating privacy considerations into the development of various systems, products, services, and organisations right from the outset. It emphasises the proactive safeguarding of privacy interests, ensuring they are a fundamental part of the system development process, rather than being added as an afterthought. The EU has acknowledged the significance of privacy by design and underscored its importance through various legal frameworks. One pivotal legislation that champions this concept is the General Data Protection Regulation (GDPR), which became enforceable in May 2018. Article 25 of the GDPR establishes privacy by design as a fundamental principle, mandating organizations to implement suitable technical and organizational measures to protect individuals' personal data. By adopting privacy-friendly practices, organisations can mitigate the risks of privacy breaches and uphold their legal obligations.

Secondly, the principles of security by design extend the established concept of privacy by design, where technology and technical measures are employed to optimise privacy in accordance with legal frameworks. Security by design encompasses this notion and underscores the necessity of fortifying AI-assisted decision-making systems against unauthorised intrusion and manipulation, thereby emphasising resilience. Thirdly, ethics by design entails purposefully integrating ethical and humane use principles into the entire lifecycle of designing, developing, and delivering software and services. In practical terms, this approach empowers researchers, developers, and practitioners to incorporate ethical use principles, including human rights, privacy, safety, honesty, and inclusion, into their activities throughout the project.

Specifically, there are six overarching principles that any AI system must uphold, grounded in fundamental rights established in the Charter of Fundamental Rights of the European Union (EU Charter) and relevant international human rights law, according to the guidelines of the document "Ethics By Design and Ethics of Use Approaches for Artificial Intelligence" of the European Commission (2023):

1. **Respect for human agency:** Human beings must have their autonomy, dignity, and freedom respected, allowing them to make their own decisions and carry out their actions.
2. **Privacy and data governance:** People have a fundamental right to privacy and data protection, which must be upheld and always respected.
3. **Fairness:** AI systems should ensure that individuals are provided with equal rights and opportunities, without undeserved advantages or disadvantages.
4. **Individual, social, and environmental well-being:** AI systems should contribute to, rather than harm, the well-being of individuals, society, and the environment.
5. **Transparency:** The purpose, inputs, and operations of AI programs should be transparent and understandable to all relevant stakeholders.

6. **Accountability and oversight:** Humans should be able to comprehend, supervise, and control the design and operation of AI-based systems. The actors involved in the development and operation of these applications should take responsibility for their functioning and the resulting consequences.

These principles serve as a foundational framework to ensure that AI technologies such as those to be developed by AI4TRUST align with human rights, ethical standards, and societal values.

The implication here is that, before creating an AI tool, it is essential to proactively assess and address the potential for biases and intrusion. This proactive approach involves considering privacy, ethics, and security by design principles, with the understanding that there is a risk in assuming that responsibility for the legal, ethical, and secure use of these tools can simply be shifted from those who administer or use them to those who design them. In practice, while privacy, ethics, and security by design are crucial components of building trustworthy AI systems, they are not standalone solutions. They are part of a broader toolbox aimed at ensuring that AI usage is legal, ethical, and secure, ultimately serving the greater good of society. This approach recognises that responsibility for legal, ethical, and secure AI use should be a shared effort involving multiple stakeholders, including designers, administrators, and users.

4.1.1. The importance of privacy, ethics, and security by design for AI4TRUST

Discussions are ongoing throughout the project to deliberate on how to incorporate legal, ethical, and security requirements into the selection of technical options within Work Packages 3 (WP3) and 5 (WP5), as well as into the choice of words and hashtags used in the development and training of the AI4TRUST platform and its components, as outlined in Deliverable D2.1 and spanning across Work Packages 2 (WP2) and 4 (WP4). The technical methods for securing data for training these tools will align with legal mandates concerning privacy, uphold ethical values, and ensure that data is handled securely from the project's inception. Furthermore, activities related to building the socio-technical foundation for training AI tools will adhere to the aforementioned legal, ethical, and security requirements. These endeavours will also be characterised by a co-creative approach involving all project stakeholders, encompassing ICT and SSH researchers, fact-checkers, and media professionals. These discussions have been and will be documented in Deliverables D1.2 of WP1, D4.3 and D4.4 of WP4, and D5.4 of WP5, underscoring the project's commitment to addressing and integrating these crucial considerations.

Partners are clear that the tools that will be developed within the project inform and are informed by the emerging legislative framework that policymakers would hope would be a good and robust model that follows international, European, and national dictates on privacy, ethics, and security respecting secure practice for worldwide adoption. Partners are aware that these may not be fully reconcilable with the exponential increase in technological capabilities to impersonate humans

with, or without, the intention to deceive. Privacy, ethics, and security by design therefore becomes an aspiration that is challenged constantly by fast-adapting technology whose scope quickly may outstrip the purpose for which it was originally created.

The AI4TRUST consortium is actively engaged in discussing and will continue to deliberate on the values that are inherently embedded in the model and design of the AI platform and its components. To address this, several questions are being considered, such as:

1. **Unconscious bias**: Is there any unconscious bias within the training data, and can measures be taken to mitigate it?
2. **Values reflection**: What values should the AI platform reflect and uphold, and can they be prioritised without introducing bias?
3. **Criteria for selection:** What criteria underlie the selection of these values or their hierarchical ranking?

Clarity regarding these values is crucial as it helps AI4TRUST partners identify any potential deviations from them. The project's efforts to combat disinformation, misinformation, and malinformation rely on an implicit set of values, which are continuously being made explicit throughout the project's lifecycle, anticipated in this deliverable D2.1 and in D1.2 'Data Management Plan'. This transparency aims to support reflective practice by all stakeholders. Furthermore, the project's emphasis on privacy, ethics, and security by design makes clear the partners' approach to managing risks. Implicit in this approach is the anticipation and visualisation of potential harms, rather than relying solely on litigation or mitigation strategies after the deployment of AI tools if harm is alleged. This proactive stance underscores the project's commitment to addressing issues before they materialise and ensuring the responsible and ethical use of AI technologies.

4.2. Platform practical implications

The legal, ethical, and security guidelines presented define the boundaries within which the AI4TRUST project and its technical solutions are expected to operate. These guidelines, as also detailed in D1.2 'Data Management Plan,' emphasize that all components of the AI4TRUST platform must take these considerations into account throughout the entire processing pipeline. Adequate strategies should be established to ensure that solutions are designed to be ethical, privacy-compliant, and secure by design. When it comes to the actual implementation of these solutions, there are two levels at which their impact can be assessed:

1. **Single component/Algorithm level:** At this level, each individual component or algorithm within the platform is examined to ensure that it aligns with the specified legal, ethical, and security standards. Assessing the impact of each component provides valuable insights into

their compliance and allows for the identification of specific areas for improvement or adjustment.

2. **Platform level:** At this broader platform level, the overall impact of the combined components and algorithms is considered. This assessment provides a more comprehensive view of how the platform operates in terms of privacy, ethics, and security. It allows for the identification of any systemic issues and the development of overarching functional requirements to enhance the platform's privacy-related, ethical, and secure performance.

Both levels of assessment offer valuable insights that can be translated into concrete technical requirements, which are then implemented in the AI4TRUST platform to ensure that it operates in a manner that aligns with the legal, ethical, and security principles outlined in the EU guidelines. For what concerns the first level of analysis, each component, algorithm, tool, or methodology needs a careful analysis related to privacy, ethics and security-related aspects, which will be conducted in WP6 for ensuring that their design is in fact compliant to the described guidelines. Aspects such as bias in the training data sets, or use of synthetic data, for instance, need a specific analysis of the workflow related to the specific component under development and need to be addressed within the scope of that component, rather than at a higher level. On the other side, at a platform level, certain approaches have already been identified in D5.4 “Platform Specification”, to account for the more general aspects that can be collectively addressed for several components and tools at a higher level. In practice, this will imply that:

1. **Data ingestion phase:** The data ingestion phase, as detailed in Section 3.2 of D5.4 “Platform Specification”, will be conducted in a manner that adheres to privacy and security requirements. This approach will align with the legal and ethical considerations outlined in the AI4TRUST D1.2 “Data Management Plan”. Additionally, the data collection process will be purpose-limited, focusing on specific topics and tools training. It will only draw data from public sources and ensure data minimisation.
2. **Data elaboration phase:** In the data elaboration phase, as described in Section 3.3 of D5.4 “Platform Specification”, roles such as "controllers" and "processors" will be clearly defined in compliance with GDPR regulations. "Controllers" will be responsible partners ensuring data processing compliance, while "processors" will be authorised partners for personal data processing. Furthermore, this phase will handle data de-identification through techniques like anonymisation, pseudonymisation, aggregation, and detail reduction. It will exclusively support platform functionalities that align with EU ethical principles, such as preserving human autonomy, dignity, equality, and non-discrimination.
3. **Data storage and access:** The management of data storage and access, outlined in Section 3.4 of D5.4 “Platform Specification”, will incorporate appropriate authentication and authorisation techniques. All consortium stakeholders will receive access to one or more interfaces based on data sharing agreements. The management of this access will ensure the secure handling of credentials, authentication, authorisation, and will prioritise adherence to data licenses, usage rights, and privacy considerations.

4. **Data exposed to end-users:** The data exposed to end-users, as discussed in Section 3.6 of D5.4 “Platform Specification”, will be presented without personal identifiers. De-identification techniques adopted during the data elaboration phase will facilitate this. User identity information will be determined by the data imported during the data ingestion process. If the data is already anonymised or pseudonymised at the source, additional techniques for generating output will not be necessary.

These measures collectively underline the AI4TRUST project's commitment to privacy, security, and ethical considerations throughout its data processing pipeline and platform functionalities. They aim to ensure that the project operates in a manner that respects ethical principles, data privacy and security, and compliance with relevant regulations.

4.3. Scientific implications and partners’ best practices

The development of the AI4TRUST platform is a collaborative, interdisciplinary endeavour that places a strong emphasis on adhering to legal, ethical, and security standards established by both the European Union and academic/practitioners best practices. This commitment is evident in various aspects of our work. As we navigate the challenges associated with the development of a Human-Centered AI platform that utilises diverse data sources, including social media data, we remain acutely aware of the multifaceted risks inherent in AI development. These risks encompass several key areas:

1. **Bias mitigation:** We are actively addressing the potential biases that can be present in AI training data, working to ensure that our AI models produce fair and unbiased results.
2. **Responsible use:** We are vigilant about the risk of our tools being misused in ways that run counter to our platform's mission of combating disinformation, misinformation, and malinformation. We aim to develop safeguards and promote responsible use.
3. **Privacy protection:** We recognise the importance of safeguarding individual privacy, particularly in the context of social media research. We are committed to respecting privacy boundaries and adhering to ethical and security practices when working with such data.

By proactively addressing these challenges and remaining committed to our mission, the AI4TRUST project strives to develop a platform that not only meets high legal, ethical, and security standards but also contributes to the responsible and ethical use of AI in combatting disinformation, misinformation and malinformation and upholding individual privacy. AI4TRUST will naturally comply with the latest EU regulations in terms of data collection, dissemination, and protection, as well as with respect to the use of AI and deep learning methods.

If there is an intention to share non-public data within the consortium, such as for the purpose of further training AI models or processing data during the inference stage, the establishment of Data Processing Agreements will be a prerequisite before any data sharing occurs. These agreements

will outline the terms and conditions governing data sharing, ensuring that data is handled in a manner that complies with legal and ethical standards. Moreover, if there is a need to qualitatively evaluate the performance of our technologies through user cases (e.g., in WP6), involving individuals from our partners, all necessary technical and organisational measures will be implemented to obtain explicit consent from data subjects. These measures will be put in place to protect the rights and privacy of data subjects, emphasising the importance of ethical and responsible data handling.

In line with the project's research outputs, it is advisable to implement the following measures:

1. **Secure storage:** Store research outputs on file servers that are protected by standard security measures to prevent unauthorized access. Keep security software up-to-date with the latest security patches and establish backup provisions to ensure data integrity and availability.
2. **Public availability of AI Models:** Selected trained AI models and the software for building and retraining these models can be made publicly available on platforms like GitHub, promoting transparency and collaboration.
3. **Limited accessibility for sensitive AI Models:** Some AI models, particularly deep face detection models, should not be made publicly available due to their susceptibility to adversarial attacks. These models can be securely stored and accessed only by authorised personnel to mitigate security risks.
4. **User notifications:** Ensure that AI4TRUST users are informed about the limitations of the technologies provided. This can be done through notifications or disclaimers in the user interface, emphasising that the output of the AI tools should be used as an aid in their own assessments and final judgments, rather than as the sole basis for trustworthiness evaluations.

These measures collectively contribute to responsible data management, cybersecurity, transparency, and user awareness, aligning with the project's goals of building trustworthy AI technologies while safeguarding security and ethical considerations.

In the subsequent subsections, we provide a brief overview of good practices drawn from various fields. These practices will be further elaborated upon and tailored to the specific requirements of the AI4TRUST platform development during the activities of WP3 and WP5.

4.3.1 Interdisciplinary framework

The mitigation of disinformation demands rigorous validation and verification processes to maintain the integrity and credibility of social media data, upholding empirical research standards in the face of constantly evolving misleading content. Similarly, counteracting informational disorders requires a deep understanding of the cultural underpinnings that influence the spread and reception of false narratives. Our goals cannot be achieved outside a systematic framework of interdisciplinary collaboration between social scientists, data scientists, and information technologists, which not only leverages and connects qualitative and quantitative analytical methods but, even more

relevantly, cross-fertilizes domain knowledge and competencies. The aim is to produce a comprehensive analysis of mis/dis/mal-information and a set of interdisciplinary practices that augment the quality of the notions and solutions proposed. Equally relevant, the cross-fertilization of perspectives allows the combination of a clear focus on cutting-edge technical solutions with ethical and data protection concerns, which are pivotal when addressing a pervasive phenomenon like disinformation, its manifold cultural nuances and the variety of sensibilities and experiences involved. Upholding transparency in research methodologies and findings remains critical in fostering trust and accountability, facilitating an inclusive discourse within the scientific community, while simultaneously respecting the cultural intricacies and uniqueness of the communities affected by disinformation.

4.3.2 Computer vision

As a best practice in terms of data protection, it is suggested to base AI4TRUST research on existing publicly available datasets for training AI models for: a) deep fake image and video detection, b) sensational visual content detection, c) condensed video representation, d) visual-text misalignment detection, and e) explainable image and video classification. To ensure data protection best practices in AI4TRUST research, the University of Trento (UNITN, Italy) will adopt publicly available datasets for training AI models for image and video generation tasks.

4.3.3 Audio AI

In terms of audio deepfake generation, all training data and models will use publicly available repositories. All generated samples will be labelled as such, and any potential impersonating model will not be shared with any external partners. If need be, the synthesised samples may also contain digital watermarks, making them easily identifiable as generated data, rather than real data. Any tests conducted with human subjects (e.g., listening tests, recording sessions) will ensure the fair use and protection of any personal details of the participants. Data encryption and protection will also be enforced for any speech data collected by volunteering or remunerated means.

4.3.4 Natural language processing

NCSR-D and FBK will comply with legal regulations and adhere to best practices to ensure that NLP applications are developed and deployed in an ethical and secure manner, safeguarding user privacy and system integrity. Regarding ethics, NCSR-D will additionally adhere to the Association for Computational Linguistics (ACL) responsible NLP guidelines, as represented by the ACL 2023

Responsible NLP Checklist[54] and the ACM Code of Ethics and Professional Conduct[55]. According to these guidelines, the limitations and potential risks of each work should be considered (and discussed in scientific publications). Regarding the creation of scientific artefacts (e.g., datasets, models, etc.), proper citation practices of re-used artefacts should be followed, and usage consistent with the intended use of re-used artefacts should be ensured, while clear licensing must accompany artefact distribution. Additionally, precautions should be taken to protect anonymized, confidential, or unauthorised data from accidental disclosure. Particular attention is paid to data from social media, for which we avoid user profiling and we present, whenever possible, data analysis in aggregated format removing user information. During the development of the models, the details of model set up should be listed (model version, details of train / test / dev splits, number of total experiments, hyperparameter search and best-found values, computational budget (e.g., GPU hours), computing infrastructure used, packages used, descriptive statistics of results). With regard to annotation, the full text of instructions given to human annotators should be given, as well the demographic and geographic characteristics of the annotators. Data collection protocol should be approved -if applicable - by an ethics review board, while people, whose data are used/curated, should give a consent. As for the last partner that will take over the NLP work,, GDI has various policies in place to address security and ethical concerns in its work. Firstly, GDI has a data protection policy[56] available publicly which outlines Data protection and security principles implemented at GDI. Additionally GDI has internal policies which shape its work including a private and cybersecurity policy, a data retention policy, as well as a guideline to train analysts and avoid bias being integrated in trained data.

4.3.5 Social network analysis

Relational data needed for social network analysis typically relies on connections between user/nodes. By definition of the project, which revolves around digital public spaces, the targeted platforms are such that the data is entirely publicly-available at the time of collection. Users are aware by design that this information will be made public and potentially accessible by any other Internet user, such as academics. This potentially includes any form of relational data as well i.e., links denoting explicit affiliation (followers, members) and implicit interaction (replies, likes). In this context, users are publicly identified by labels of their choice (diversely denoted as “user names”, “account names”, “screen names”, “login names”): this information will be collected and may contain personal names (for instance, if a user chooses "Jane doe" as a username, they implicitly make it possible to trace their personal name). They may also refer to other user names and cite URLs which point to other users of the target platform, or another identifiable platform.

[54] https://aclrollingreview.org/responsibleNLPresearch/
[55] https://www.acm.org/code-of-ethics
[56] https://www.disinformationindex.org/privacy/

As a result, AI4TRUST endeavours to ensure that the data describing any type of user names will be kept to a bare minimum and that in all instances where it is not needed, it will be de-identified resorting to the appropriate technique (e.g., anonymisation, pseudonymisation, detail reduction, noise addition, aggregation). For example, posts could be numerically identified according to a numbering proper to AI4TRUST (i.e., distinct from the unique identifiers potentially attributed by the respective platforms and which would make it possible to easily trace back the origin of the post and its author). In practice and in general, pseudonymised IDs are typically generated by hashing the non-pseudonymised data, requesting an encryption key from a separate database that associates small ranges of hash values with specific keys, then encrypts the non-pseudonymised data with the key corresponding to a given hash value range. The separate key database thus connects, in essence, a large number of small sets of possible user ID values with distinct keys. To decode a pseudonymised ID, one thus needs to know (1) the encrypted pseudonymised ID together with its hash range and (2) the encryption key associated with the hash range in the separate key database. Without the key database, it is technically reasonably intractable to retrieve the non-pseudonymised ID from an pseudonymised ID. As such, this process ensures that the two types of information are kept in separate databases, whereby none alone is sufficient to access non-pseudonymised data. This yields a database equivalent to a so-called “user database”, separate from pseudonymised data sets usually denoted as “content databases”.

References

Abroshan, M., Khalili, M. M., & Elliott, A. (2022, October). Counterfactual Fairness in Synthetic Data Generation. In NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research.

Agarwal, S., Muku, S., Anand, S., & Arora, C. (2022). Does data repair lead to fair models? curating contextually fair data to reduce model bias. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3298-3307).

Ahmadi, N., Lee, J., Papotti, P., & Saeed, M. (2019). Explainable fact checking with probabilistic answer set programming. arXiv preprint arXiv:1906.09198.

Alhindi, T., Petridis, S., and Muresan, S. (2018). Where is your evidence: Improving fact-checking by justification modeling. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-236.

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

Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S. & Ali, Z.S. (2020). Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Proceedings 11 (pp. 215-236). Springer International Publishing.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R. & Sanguinetti, M. (2019, June). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 54-63).

Baxevanakis, S., Kordopatis-Zilos, G., Galopoulos, P., Apostolidis, L., Levacher, K., Baris Schlicht, I. & Papadopoulos, S. (2022, June). The mever deepfake detection service: Lessons learnt from developing and deploying in the wild. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation (pp. 59-68).

Betzel, M., Nyakas, L., Papp, T., Kelemen, L., Monori, Z., Varga, Á., Marrazzo, F., Matějka, S., Ó Fathaigh, R., & Helberger, N. (2020). Notions of Disinformation and Related Concepts (ERGA Report) [Report]. European Regulators Group for Audiovisual Media Services. https://erga-online.eu/wp-content/uploads/2021/03/ERGA-SG2-Report-2020-Notions-of-disinformation-and-related-concepts-final.pdf.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R. & Xiang, A. (2021, July). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 401-413).

Bode, L. & Vraga E. K. (2018). See something, say something: Correction of global health misinformation

on social media. Health communication 33, 9 (2018), 1131–1140.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V.,& Kalai, A.T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

Bonial, C., Blodgett, A., Hudson, T., Lukin, S., Micher, J., Summers-Stay, D. & Voss, C. (2022, June). The Search for Agreement on Logical Fallacy Annotation of an Infodemic. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 4430-4438). Available: https://aclanthology.org/2022.lrec-1.471.

Borzì, S., Giudice, O., Stanco, F., & Allegra, D. (2022). Is synthetic voice detection research going into the right direction?. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 71-80).

Buning, M.D.C. (2018). A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation. Publications Office of the European Union.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Casula, C. & Tonelli, S. (2023). Generation-Based Data Augmentation for Offensive Language Detection: Is It Worth It?. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3359–3377, Dubrovnik, Croatia. Association for Computational Linguistics.

Chai, L., Bau, D., Lim, S. N., & Isola, P. (2020). What makes fake images detectable? understanding properties that generalize. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16 (pp. 103-120). Springer International Publishing.

Chaudhari, B., Choudhary, H., Agarwal, A., Meena, K., & Bhowmik, T. (2022). FairGen: Fair Synthetic Data Generation. arXiv preprint arXiv:2210.13023.

Chung, Y. L., Tekiroğlu, S. S., Tonelli, S., & Guerini, M. (2021). Empowering NGOs in countering online hate messages. Online Social Networks and Media, 24, 100150.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S. & Villata, S. (2020) A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology (TOIT) 20 (2), 1-22.

Das, A., Liu, H., Kovatchev, V., & Lease, M. (2023). The state of human-centered NLP technology for fact-checking. *Information processing & management*, *60*(2), 103219.

Dash, S., Balasubramanian, V. N., & Sharma, A. (2022). Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 915-924).

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955.

de Oliveira, N.R., Pisa, P.S., Lopez, M.A., de Medeiros, D.S.V., & Mattos, D.M.F. (2021). Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. Information, 12, 38. https://doi.org/10.3390/info12010038.

Dogoulis, P., Kordopatis-Zilos, G., Kompatsiaris, I., & Papadopoulos, S. (2023, June). Improving Synthetically Generated Image Detection in Cross-Concept Settings. In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation (pp. 28-35).

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C.C. (2020). The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397.

European Commission (2023). Code of Practice on Disinformation:new reports available in Transparency Centre. Retrieved from: https://digital-strategy.ec.europa.eu/en/news/code-practice-disinformation-new-reports-available-transparency-centre.

Fathaigh, R.Ó., Helberger, N., & Appelman, N. (2021). The perils of legally defining disinformation. Internet Policy Review, 10(4). https://doi.org/10.14763/2021.4.1584.

Flouris, G., Efthymiou, V., Papantoniou, K., Patkos, T., Petasis, G., Pittaras, N., Plexousakis, D., Roussakis, G., Tzortzakakis, E., & Ymeralli, E. (2022). DebateLab Tools for E-Journalism and Informed Citizenship. In GEC-22. Retrieved from http://users.ics.forth.gr/~fgeo/files/GEC22.pdf.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior, in: Proceedings of the International AAAI Conference on Web and Social Media.

Gad-Elrab, M.H., Stepanova, D., Urbani, J., & Weikum, G. (2019). Exfakt: A framework for explaining facts over knowledge graphs and text. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM'19, page 87–95, New York, NY, USA. Association for Computing Machinery.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J. & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 1-77.

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., & Derczynski, L. (2019, June). SemEval-2019 Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours. In Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019 (pp. 845-854). Association for Computational Linguistics.

Grimminger, L., & Klinger, R. (2021). Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection. https://www.ims.uni-stuttgart.de/.

Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., & Verdoliva, L. (2023). TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20606-20615).

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. Transactions of the Association for Computational Linguistics, 10, 178-206. https://doi.org/10.1162/tacl_a_00454.

Gupta, A., and Srikumar, V. (2021, August). X-Fact: A New Benchmark Dataset for Multilingual Fact Checking. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 675-682). http://dx.doi.org/10.18653/v1/2021.acl-short.86.

Haliassos, A., Mira, R., Petridis, S., & Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14950-14962).

He, B., Ahamad, M., and Kumar, S. (2023). Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In Proceedings of the ACM Web Conference 2023, pages 2698–2709.

He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., ... & Liu, Z. (2021). Forgerynet: A versatile benchmark for comprehensive forgery analysis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4360-4369).

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. In Proceedings of the European conference on computer vision (ECCV) (pp. 771-787).

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110, 457-506.

Ireton, C., & Posetti, J. (2018), A short guide to the history of 'fake news': A learning module for journalists and journalism educators. UNESCO.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing*, 126232. https://doi.org/10.1016/j.neucom.2023.126232.

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R. & Schölkopf, B. (2022). Logical fallacy detection. Findings of the Association for Computational Linguistics: EMNLP 2022, 7209–27. https://doi.org/10.18653/v1/2022.findings-emnlp.532.

Jung, S., Chun, S., & Moon, T. (2022). Learning fair classifiers with partially annotated group labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10348-10357).

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

Kawa, P., Plata, M., Czuba, M., Szymański, P., Syga, P. (2023) Improved DeepFake Detection Using Whisper Features. Proc. INTERSPEECH 2023, 4009-4013, doi: 10.21437/Interspeech.2023-1537.

Kim, H., Kim, S., Yeom, J., & Yoon, S. (2023). UnitSpeech: Speaker-adaptive Speech Synthesis with Untranscribed Data. arXiv preprint arXiv:2306.16083.

Kotonya, N. and Toni, F. (2020a). Explainable automated fact-checking: A survey. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kotonya, N. and Toni, F. (2020b). Explainable automated fact-checking for public health claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7740–7754, Online. Association for Computational Linguistics.

Lazega, E. (1990), "Internal Politics and the Interactive Elaboration of Information in Workgroups: An Exploratory Study", *Human Relations*, 43:87-101.

Lazega, E. (1992), Micropolitics of Knowledge, New York: Aldine de Gruyter.

Li, R., Fontanini, T., Prati, A., & Bhanu, B. (2022). Face Synthesis With a Focus on Facial Attributes Translation Using Attention Mechanisms. IEEE Transactions on Biometrics, Behavior, and Identity Science, 5(1), 76-90.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2016). A survey on truth discovery. ACM SIGKDD Explorations Newsletter, 17(2), 1-16. https://doi.org/10.1145/2897350.2897352.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., & Plumbley, M. D. (2023). AudioLDM: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503.

Lodge, J (2010) Quantum Surveillance and 'Shared Secrets: A biometric Step too far? CEPS.ISBN 978-94-6138-009-8.

Lu, Y.-J. and Li, C.-T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 505–514, Online. Association for Computational Linguistics.

McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *The Journal of Abnormal and Social Psychology*, *62*(2), 327.

Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2020). Ethos: an online hate speech detection dataset. https://doi.org/10.1007/s40747-021-00608-2.

Momina, M., et al. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence 53.4, 3974-4026.

Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). Does audio deepfake detection generalize?. Interspeech.

Musi, E., & Reed, C. (2022). From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*, *33*(3), 349-370.

Nam, J., Cha, H., Ahn, S., Lee, J., & Shin, J. (2020). Learning from failure: De-biasing classifier from biased classifier. Advances in Neural Information Processing Systems, 33, 20673-20684.

Ntogramatzis, A. F., Gradou, A., Petasis, G., and Kokol, M. (2022, June). The Ellogon Web Annotation Tool: Annotating Moral Values and Arguments. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 3442-3450).

Ojha, Utkarsh, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Oneață, D., Stan, A., Pascu, O., Oneață, E. & Cucu, H. (2023). Towards generalisable and calibrated synthetic speech detection with self-supervised representations. Submitted to ICASSP.

Oneată, E., Oneață, D. & Tînțaru, D. (2023). Weakly-supervised deepfake localization in diffusion-generated images. Submitted to WACV.

Oshikawa, R., Jing Qian, and William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6086–6093. https://aclanthology.org/2020.lrec-1.747/.

Paleyes, A., Urma, R. G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: a survey of case studies. ACM Computing Surveys, 55(6), 1-29.

Papadopoulos, G., Kokol, M., Dagioglou, M., and Petasis, G. (2023, July). Andronicus of Rhodes at SemEval-2023 Task 4: Transformer-Based Human Value Detection Using Four Different Neural Network Architectures. In Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 542-548).

Pavlopoulos, J., Sorensen, J., Laugier, L., & Androutsopoulos, I. (2021, August). SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)* (pp. 59-69). https://aclanthology.org/2021.semeval-1.6/.

Petasis, G., Karkaletsis V., Paliouras G., Androutsopoulos, I., & Spyropoulos C. (2002). Ellogon: A New Text Engineering Platform. https://arxiv.org/abs/cs/0205017.

Pîrlogeanu, G., Oneaţă, D., Georgescu, A., & Cucu, H. (2023). The SpeeD–ZevoTech submission at DISPLACE 2023. Interspeech.

Popat, K., Mukherjee, S., Yates, A., and Weikum, G. (2018). DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Preechakul, K., Chatthee, N., Wizadwongsa, S., & Suwajanakorn, S. (2022). Diffusion autoencoders: Toward a meaningful and decodable representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10619-10629).

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

Ramponi, A. & Tonelli, S. (2022). Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3027–3040, Seattle, United States. Association for Computational Linguistics.

Ramponi, A., Testa, B., Tonelli, S. & Jezek, E. (2022) Addressing religious hate online: from taxonomy creation to automated detection. PeerJ Computer Science 8, e1128.

Reimao, R., & Tzerpos, V. (2019, October). For: A dataset for synthetic speech detection. In 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) (pp. 1-10).

Rodríguez Perez, C (2019). Una reflexión sobre la epistemología del fact-checking journalism: retos y dilemas (pp. 245).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

Rosello, E., Gomez-Alanis, A., Gomez, A.M., Peinado, A. (2023) A conformer-based classifier for variable-length utterance processing in anti-spoofing. Proc. INTERSPEECH 2023, 5281-5285, doi: 10.21437/Interspeech.2023-1820.

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

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H. & Pierrehumbert, J.B. (2020). HateCheck: Functional Tests for Hate Speech Detection Models. 10.18653/v1/2021.acl-long.4.

Russo, D., Tekiroglu, S.S. and Guerini, M. (2023) Benchmarking the Generation of Fact Checking Explanations. Transactions of the Association for Computational Linguistics Journal.

Sahai, S., Balalau, O., & Horincar, R. (2021), Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions, in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Apr. 2021, pp. 644–657. doi: 10.18653/v1/2021.acl-long.53.

Salvi, D., Bestagini, P., & Tubaro, S. (2023a). Reliability Estimation for Synthetic Speech Detection. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5).

Salvi, D., Hosler, B., Bestagini, P., Stamm, M.C., & Tubaro, S. (2023b). TIMIT-TTS: a Text-to-Speech Dataset for Multimodal Synthetic Media Detection. IEEE Access.

Sardianos, C., Katakis, I.M., Petasis, G., & Karkaletsis, V. (2015). Argument extraction from news. In Proceedings of the 2nd Workshop on Argumentation Mining, Denver, CO, USA, 56–66.

Savani, Y., White, C., & Govindarajulu, N.S. (2020). Intra-processing methods for debiasing neural networks. Advances in neural information processing systems, 33, 2798-2810.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). Defend: Explainable fake news detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data 8 3, 171–188.

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

Sourati, Z., Venkatesh, V. P. P., Deshpande, D., Rawlani, H., Ilievski, F., Sandlin, H. Â., & Mermoud, A. (2023).

Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, *266*, 110418. doi: https://doi.org/10.1016/j.knosys.2023.110418.

Stammbach, D., & Ash, E. (2020). e-fever: Explanations and summaries for automated fact checking. Proceedings of the 2020 Truth and Trust Online (TTO 2020), pages 32–43.

Stock, P., & Cisse, M. (2018). Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 498-512).

Stowarzyszenie Demagog. (2019). Krytyczny umysł. Problem fake news w Polsce. https://krytycznyumysl.pl/

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., & Larcher, A. (2021). End-to-end anti-spoofing with RawNet2. In ICASSP (pp. 6369-6373).

The Newsreel Project Consortium (2021). Newsreel2. New Teaching Fields for the Next Generation of Journalists. Research report. Dortmund: Erich Brost Institute for International Journalism, https://eldorado.tu-dortmund.de/handle/2003/40586.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018, June). FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 809-819). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074

Tindale, C.W. (2007). Fallacies and argument appraisal. Cambridge University Press. https://doi.org/10.1017/CBO9780511806544.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

Torralba, A., & Efros, A. A. (2011, June). Unbiased look at dataset bias. In CVPR 2011 (pp. 1521-1528). IEEE.

Valer, G., Ramponi, A., & Tonelli, S. (2023). When you doubt, abstain: A Study of automated fact-checking in Italian under domain shift. In Proceedings of the Ninth Italian Conference on Computational Linguistics. [In press].

Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global challenges*, *1*(2), 1600008.

Wang, W.Y. (2017). liar, liar pants on fire: A new benchmark dataset for fake news detection, in: ACL.

Wang, X., & Yamagishi, J. (2022). Investigating self-supervised front ends for speech spoofing countermeasures. Odyssey 2022: The Speaker and Language Recognition Workshop.

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020a). Towards fairness in visual recognition: Effective strategies for bias mitigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8919-8928).

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N. & Ling, Z. H. (2020b). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech &

Language, 64, 101114.

Wardle, C. (2020, September 22). Understanding Information disorder. First Draft. https://firstdraftnews.org/long-form-article/understanding-information-disorder/.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

Xie, Y., Cheng, H., Wang, Y., Ye, L. (2023) Learning A Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection. Proc. INTERSPEECH 2023, 2808-2812, doi: 10.21437/Interspeech.2023-1383.

Xu, D., Yuan, S., Zhang, L., & Wu, X. (2018, December). Fairgan: Fairness-aware generative adversarial networks. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 570-575). IEEE.

Xu, W., Zhao, J., Iannacci, F., & Wang, B. (2023). FFPDG: Fast, fair and private data generation. arXiv preprint arXiv:2307.00161.

Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., and Hu, X. B. (2019). Xfake: Explainable fake news detector with visualizations. In The World Wide Web Conference, WWW '19, page 3600–3604, New York, NY, USA. Association for Computing Machinery.

Ymeralli, E., Flouris, G., Efthymiou, V., Papantoniou, K., & Patkos, T. (2022). Representing Online Debates in the Context of E-Journalism." http://thinkmind.org/index.php?view=article&articleid=semapro_2022_1_10_30005.

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4578-4587).

Zhang, F., Kuang, K., Chen, L., Liu, Y., Wu, C., & Xiao, J. (2022, September). Fairness-aware contrastive learning with partially annotated sensitive attributes. In The Eleventh International Conference on Learning Representations.

Zhang, X., Yi, J., Tao, J., Wang, C., & Zhang, C.Y. (2023). Do You Remember? Overcoming Catastrophic Forgetting for Fake Audio Detection. ICML.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.

Annex I

In this Annex, we present the initial translation and adaptation efforts undertaken by our media and fact-checker partners to facilitate the initial data collection process. As described in Section 3 of this deliverable, it is important to view these lists as dynamic and continuously evolving resources. They will be regularly updated and enhanced with insights from the active collection of social media and news media data, as well as feedback from our expert collaborators.

Proposed keywords in French

Climate change

Changement climatique, co2, catastrophe environnementale, villes sans voiture, déversement de produits chimiques, canular sur le changement climatique, Greta Thunberg, le réchauffement climatique.

Public health

Adrénochrome, anti-avortement, anti-vaccin, anticorps, antivaxer, arme biologique du ccp, arrêt cardiaque, asymptomatique, aucun enfant vacciné, auto-isolement, avortement, biotech, blessures vaccinales, caillot sanguin, calendrier des vaccinations infantiles, choix du vaccin, Code de Nuremberg, congé menstruel, convention istanbul, courbe épidémique, COVID-19 crimes contre l'humanité, décès vaccin, effets indésirables, effets secondaires du vaccin, essais cliniques, extraterrestre, fièvre jaune, fœtus, fuites extraterrestres, industrie pharmaceutique, grippe Wuhan, homéopathique, immunité collective, influenceur anti-vaccin, johnson et johnson,vaccin variole, vaccin variole, masques esclaves, mauvais médicament, médecine naturelle, médecine naturelle, médicament dangereux, mort vaccin, mutation virale, non vacciné, ordre du jour extraterrestre, OVNI, pas de piqûre, passeport covid, patient zéro, période d'incubation, passeport covid, patient zéro, période d'incubation, pro vie, procès de nuremberg covid, prochoix, Rapport d'OVNI, Roe Wade, rougeole, sang de serpent, souche delta, SRAS-CoV-2, test de flux latéral, tyrannie médicale, vaccin, vacciné, variant mongole, variant sud-africaine, variole du singe, virus ccp, virus jaune, virus de wuhan, pillule abortive, pillule avortement, avortement genocide, avortement meutre, avortement assasinat, avortement cancer, avortement sterile, avortement danger, ivg genocide, ivg meurtre, ivg cancer, ivg sterile, ivg danger, ivg assassinat, sidaique, avortement meurtre, protege vie avortement, protege vie ivg, #BigPharma, #CeVaccinNeSertaRien, #JeNeMeVaccineraiPas, #NonAuVaccin, #StopVaccin, BigPharma vaccin, bill gates vaccin, complot vaccin, complot vacciniation, crise cardiaque vaccin, diktat vaccin, elite mondiale vaccin, elite mondiale vaccin, dépeuplement vaccin, faux vaccin, mort vaccin, franc-macon vaccin, holdup, juif vaccin, laboratoires pharmaceutiques complot vaccin,mensonge vaccin, mort AstraZeneca, mort Moderna, mort Pfzier, mort Spoutnik, mort vaccin, non Passeport Vaccinal, OMS complice vaccin, Pharma vaccin complot, Raoult vaccin, rat laboratoire vaccin, Soros vaccin, Soros covid, Soros pandemie, vaccin 5G, vaccin assassin, vaccin cobaye francais, vaccin criminel, vaccin danger, vaccin desobeissance civile, dictature

sanitaire, vaccin reduire population, vaccin mort, vaccin poison, vaccin puce, vaccin Qanon, vaccin toxique, vaccin tue, vaccin tuer population, vaccination danger, vaccin danger, vaccin cobaye, faux vaccin, complot vaccin, faux vaccin, complot vaccin, OMS complice vaccin, complot vaccin, OMS complice vaccin, vaccination danger, complot vaccination, #JeNeMeVaccineraiPas, non vaccin, bill gates vaccin, cobaye covid, cobaye coronavirus, non vaccination, #StopVaccin, #NonAuVaccin, mort vaccin, vaccin tue, #CeVaccinNeSertaRien, #BigPharma, vaccin Qanon, covid Qanon, big pharma, vaccin puce, vaccin 5G, Rothschild vaccin, vaccin 666, vaccin satan, vaccin golem, OMS meurtre, OMS danger, controle population OMS, Non Au TestPCR, génocide sanitaire, fausse pandemie, Desobeissance Civile, non au port du masque obligatoire, couvre-feu dictature, cobaye covid, masques révoltez-vous, NonAuMasque, chloroquine, coronagate, vaccins obligatoires, veran demission, chinois responsable, covid collabos, fausse pandemie, covid mensonge, remède éthanol, remède javel, révolte corona.

Migrants

Migrant, immigrant, enfants migrants, tommy robinson, immigrant, invasion migratoire, migration, immigration, étranger en situation irrégulière, Étrangers illégaux, grand replacement, regroupement familiale crime originel, réfugié dehors, envahisseur, reimigration, invasion migratoire, menace migratoire, france aux francais, migrants dehors, sale migrant, sale réfugié, dégage migrant, dégage réfugié, pseudo migrant, faux migrant, faux réfugié, pseudo migrant, pseudo réfugié, insécurité migrant, invasion africaine, Invasion Migratoire, reconquista, déferlement migratoire, immigrationdemasse, immigration terrorisme, menace migratoire, france aux francais, migrants dehors, immigration anarchique, oqtf, urgent rétablir frontières, invasion africaine, #francaisreveillezvous, Déferlement migratoire, métèque.

Proposed keywords in German

Climate change

Greta Thunberg, Energiewende, CO2, Umweltkatastrophe, autolose Städte, Klimawandel, die globale Erwärmung, Chemieunfall, Klimawandel-Schwindel, Klima Hysterie, Klima Elite, Klima-Irre, Fridays for Future, Klimareligion, Luisa Neubauer, Klima-Ideologen, Klima-Terrroristen, Klima-Chaoten, Klimakontrolle.

Public health

Arzneimittel, Impfopfer, Impftote, Schlangenblut, Adrenochrom, Impfung, Lambda-Stamm, Haftung für gefährliche Produkte, Nürnberger Prozesse wegen Covid, Virusmutation, Spike-Protein, Vax-Tod, mRNA-Brühe, Epidemiekurve, Gen-Impfung, Herzmuskelentzündung , Genspritze, zusammengebrochen, Todesschuss, Nürnberg 2.0, Nebenwirkung, Stoppen Sie den Völkermord, kein Stich, Alien-Agenda, Masern, Masken sind für Sklaven, PSA, UFO-Bericht des Pentagons, keine Kinder geimpft, Covid-Pass, Maskenpflicht, Wahl des Impfstoffs, Robert Koch-Institut, RKI, WHO-Diktatur, Nürnberg 2, geimpft, Impfverletzungen, Spritze, Nürnberger Code, die große Keulung, System zur Meldung unerwünschter Impfstoffe, KHK, Karl Lauterbach, EMA, Pharma-Lobby, Impfverbrecher, medizinische Tyrannei, Gerinnsel, Unbekanntes

Flugobjekt, gefährliche Medizin, Wuhan-Virus, südafrikanische Variante, krebserregend, Pharmaindustrie, Big Pharma, Pfizer, Biontech, Geldpocken, Bernsteinliste, Johnson und Johnson, Koalition zur Seuchenvorsorge, Ungeimpfte Leben sind wichtig, Impfplan für Kinder, Pferdepaste, Sicherheitsprobleme, Todesstoß, Impfvorschriften, Krebs, Zwangsimpfungen, WLAN der 5. Generation, Gelbfieber, mongolische Variante, mongolische Variante, Impfungen, Wuhan-Grippe, China-Virus, UFO-Bericht, ungeimpft, Unbekannte Luftphänomene, Salbei, Heilung, Medizin, Bell-Lähmung, Millimeterwelle, WLAN der fünften Generation, Impftote, Tötungsschuss, Selbstisolation, klinische Versuche, Impftote, Tötungsschuss, Selbstisolation, klinische Versuche, klinische Versuche, Schlangengift, Nebenwirkungen des Impfstoffs, Sucharit Bhakdi, Verbrechen gegen die Menschheit, Abflachung der Kurve, Anti-Impfstoff-Influencer, Außerirdischer, absichtliches Gift, absichtliches Gift, CCP-Biowaffe, CCP-Biowaffe, Affenpocken, Inkubationszeitraum, schlechte Medizin, Patient null, natürliche Medizin, Antikörper, Versuchskaninchen, Herzstillstand, Nebenwirkungen, CCP-Virus, schlechte Wissenschaft, Delta-Stamm, Big-Tech-Zensur, Lass die Körper auf den Boden fallen, rote Ankünfte, Pro-SAFE-Impfstoff, Herdenimmunität, Vaers-Datenbank, Kontaktverfolgung, homöopathisch, asymptomatisch, Corona-Diktatur, Coronavirus, Mandat, Gift Cocktail, Schnelltests, Pro-Wahl, Abtreibung, Gynäkologie, Pro-Choice, prochoice, Fötus, Roe gegen Wade, Istanbuler Konvention, Menstruationsurlaub, Menstruation.

Migrants

Anti-Schwuchtel-Gesetze, Immigrant, Einwanderung, Migranten, Einwanderungswelle, Invasion, Einwanderer, Kindermigranten werden vermisst, Kriminalität, Undankbarkeit, undankbar, Vergewaltigung, vergewaltigen, Sozialtourismus, Sozial-Tourismus, Sozialtouristen, Sozialleistung ausnutzen, auf Staatskosten, Sozialstaat ausnutzen, Wohnungsnot, Luxusflüchtling, Luxus Flüchtling, Luxusunterkünfte, Messer Migranten, Messermigranten, Ausländer, Horde, Ausländerkriminalität.

Proposed keywords in Greek

Climate change

Στρατηγικές προσαρμογής, μόλυνση του αέρα, απώλεια βιοποικολότητας, πόλεις χωρίς αυτοκίνητα, εκπομπές διοξειδίου του άνθρακα, εκπομπές άνθρακα, αποτύπωμα άνθρακα, τιμολόγηση άνθρακα, δέσμευση άνθρακα, διαρροή χημικών, καθαρή ενέργεια, τεχνολογία καθαρής ενέργειας, ακτιβισμός για το κλίμα, δράση για το κλίμα, προσαρμογή στην κλιματική αλλαγή, κλιματική αλλαγή, απάτη της κλιματικής αλλαγής, κλιματική κρίση, περιβαλλοντική εκπαίδευση, χρηματοδότηση για το κλίμα, επίδραση του κλίματος στο οικοσύστημα, δικαιοσύνη για το κλίμα, μετριασμός των επιπτώσεων της κλιματικής αλλαγής, μοντελοποίηση του κλίματος, κλιματική πολιτική, κλιματικοί πρόσφυγες, κλιματική ανθεκτικότητα, επιστήμη του κλίματος, κλιματική μεταβλητότητα, Copernicus, αποδάσωση, ξηρασίες, φιλικό προς το περιβάλλον, εκπομπές, περιβάλλον, έλεγχος της διάβρωσης, ακραίος καιρός, ακραία καιρικά φαινόμενα, δασοπυρκαγίες, ορυκτά καύσιμα, Παρασκευή για το κλίμα, μηχανική κλίματος, παγετώνες, παγκόσμια θερμοκρασία, παγκόμσια υπερθέρμανση, φαινόμενο του θερμοκηπίου, αέρια του θερμοκηπίου, Γκρέτα Τούνμπεργκ, διατήρηση ειδών και οικοτόπων, καταστροφή ειδών και οικοτόπων, καύσωνες, λιώσιμο των πάγων, διατήρηση της θαλάσσιας φύσης, μικροπλαστικά, μέτρα μετριασμού, φυσικές

καταστροφές, μηδενικές εκπομπές αερίων, οξίνιση των ωκεανών, υπεραλίευση, Τρύπα του όζοντος, συμφωνία του Παρισιού, φωτοβολταικά, τροπικό δάσος, πρωτοβουλίες ανακύκλωσεις, ανανεώσιμη ενέργεια, αύξηση της στάθμης της θάλασσας, βιώσιμες προσπάθειες, βιώσιμη ανάπτυξη, βιώσιμη ενέργεια, Διάσκεψη των Ηνωμένων Εθνών για την Κλιματική Αλλαγή, διαχείριση αποβλήτων, μόλυνση του νερού, λειψυδρία, καιρός, τροποποίηση του καιρού, ανεμογεννήτριες, πυρκαγιές.

Public health

5g, ασύρματα δίκτυα πέμπτης γενιάς, εξωγήινοι, ατζέντα εξωγήινων, AlienLeaks, Άντονι Φάουτσι, Big Pharma, μεγάλες φαρμακευτικές, Μπιλ Γκέιτς, CDC, CHD, COVID, COVID-19, εμβόλιο COVID, COVID19, Δρ. Μπράιν Άντρις, DNA, έκτακτη ανάγκη, FDA, Τζορτζ Σόρος, Johnson & Johnson, ιός Marburg, Moderna, Morgellons, Pfizer, RNA, Ρον Γουάτκινς, SARS-CoV-2, μετάνθρωπος, διανθρωπισμός, ATIA, UFO, UFO Report, μη αναγνωρισμένα εναέρια φαινόμενα, αδρενοχρώμιο, ανεπιθύμητα γεγονότα, παρενέργειες, αντιβιωτικά, κατά των εμβολίων, anti-vax, εντιεμβολιαστής, AstraZeneca, αυτισμός, κακή θεραπεία, κακή επιστήμη, παράλυση Bell, μεγάλες εταιρείες τεχνολογίας, λογοκρισία από μεγάλες εταιρείες τεχνολογίας, biontech, αίμα, θρόμβοι αίματος, δείγμα αίματος, καρκίνος, καρκινικά κύτταρα, καρδιακή προσβολή, καρδιακό πρόβλημα, βιολογικό όπλο του Κομμουνιστικού Κόμματος της Κίνας, ιός του Κομμουνιστικού Κόμματος της Κίνας, αεροψεκασμοί, χημειοθεραπεία, χημική άμβλωση, χωρίς χημικά, πρόγραμμα εμβολιασμών παιδικής ηλικίας, διοξείδιο του χλωρίου, θειική χονδροϊτίνη, σύνδρομο χρόνιας κόπωσης, κλινική Cleveland, κλινικές δοκιμές, εμβόλιο που προκαλεί θρόμβους, codemonkeyz, ανίχνευση επαφών, κορωνοϊός, κορονοϊός, διαβατήριο covid, θεραπείες, θεραπεία, επικίνδυνα φάρμακα, ευθύνη για επικίνδυνα προϊόντα, θάνατος, ποσοστό θανάτου, ένεση θανάτου, μείωση του πληθυσμού, αποτοξίνωση, dna, ναρκωτικά, ηλεκτρομαγνητισμός, επιδημία, επιδημιολογική καμπύλη, μάσκα προσώπου, εμβρυϊκός ιστός, έμβρυο, έμβρυα, γρίπη, υποχρεωτικά εμβόλια, θεραπεία γονιδίων, γονίδια, γραφένιο, πράσινες αφίξεις, πράσινη λίστα, εκπαίδευση σχετικά με την υγεία, κίνδυνος για την υγεία, καρδιακή προσβολή, ανοσία της αγέλης, ομοιοπαθητικός, αλογόπαστα, περίοδος επώασης, ανοσοποιητικό σύστημα, ανοσοποίηση, μόλυνση, ενέσεις, εμβολιασμός, μονάδα εντατικής θεραπείας, ένεση, σκάνδαλο εμβολίου, Janssen, θανατηφόρα ένεση, τεστ πλευρικής ροής, θανατηφόρα ένεση, μακρά covid, ελονοσία, υποχρεωτική χρήση μάσκας, μάσκες, οι μάσκες είναι για σκλάβους, ιλαρά, ιατρική άμβλωση, ιατρική τυραννία, φάρμακα, Megha Thakur, χιλιοστομετρικό κύμα, μιτοχόνδρια, Ευλογιά των Πιθήκων, θνησιμότητα, mRNA, φυσική ιατρική, Νυρεμβέργη, Νυρεμβέργη 2.0, κώδικας της Νυρεμβέργης, δίκες της Νυρεμβέργης για τον κορωνοϊό, χωρίς εμβόλιο, κανένα παιδί δεν έχει εμβολιαστεί, πανδημία, ασθενής μηδέν, αναφορά του Πενταγώνου για UFO, φαρμακευτικές εταιρείες, χάπια, πλανδημία, πολιτική ατζέντα, πρόληψη, πρωτεΐνη ακίδα, υπέρ του ασφαλούς εμβολίου, υπέρ της επιλογής, υπέρ της ζωής, ακτινοβολία, αντίδραση, κόκκινες αφίξεις, κόκκινη λίστα, χάπι αντιστροφής φαρμακευτικής άμβλωσης, Ρόμπερτ Μαλόουν, Ροσέλ Ουαλένσκι, Ρόου εναντίον Ουέιντ, απολύμανση, απομόνωση, αίμα φιδιού, δηλητήριο φιδιού, μετάλλαξη Νότιας Αφρικής, ακίδα, σταματήστε τη γενοκτονία των αγαλμάτων, μελέτη, ξαφνικός θάνατος, χειρουργική μάσκα, αφαίρεση τοξινών, οι ζωές των ανεμβολίαστων μετρούν, ανεμβολίαστοι, vaids, σύνδρομο επίκτητης ανοσοανεπάρκειας από το εμβόλιο, εμβόλιο, VAERS, Σύστημα Αναφοράς Ανεπιθύμητων Παρενεργειών των Εμβολίων, επιλογή για το εμβόλιο, θάνατοι από το εμβόλιο, τραυματισμοί από το εμβόλιο, παρενέργειες του εμβολίου, θύματα του εμβολίου, εμβολιασμός, εμβολιασμοί, Ουχάν, Γιουχάν, Κίτρινος Πυρετός.

Specific Neologisms: Ξαφνικίτιδα, μπόλι, μπολιασμένος, πρωτόκολλο θανάτου, κοροϊδοιός, σφράγισμα, τσίμπιμα, τσιπάκι.

Migrants

Αφρικανοί, Άραβες, αιτούντες άσυλο, επίθεση στην Ευρώπη, Banderisation, βάρβαροι, σύνορα, παιδιά μετανάστες, απειλή για τον πολιτισμό, επικίνδυνος πολιτισμός, εκτοπισμός, οικονομικοί μετανάστες, οικονομικές επιπτώσεις της μετανάστευσης, φοβία για την Ευρώπη, Η_Ελλάδα_προστατεύει_την_Ευρώπη, Η_Ελλάδα_δέχεται_επίθεση, μεγάλη αντικατάσταση, παράνομος, παράνομος μετανάστης, μετανάστες, μετανάστευση, μεταναστευτικός νόμος, μεταναστευτικές πολιτικές, μεταναστευτικά προγράμματα, εισβολή, προγράμματα ενσωμάτωσης, εισβολή στην Ευρώπη, εισβολείς, μετανάστες κίνδυνος για τις γυναίκες, μετανάστης δολοφόνος, μετανάστευση και ανάπτυξη, διαχείριση της μετανάστευσης, μεταναστευτικά μοτίβα, μεταναστευτική εισβολή, κράτηση μεταναστών, υγεία των μεταναστών, δικαιώματα των μεταναστών, Μουσουλμάνοι, μετανάστης βιαστής, πρόσφυγας βιαστής, κέντρο προσφύγων, προσφυγική κρίση, μετεγκατάσταση προσφύγων, πρόσφυγες κίνδυνος για τις γυναίκες, μετεγκατεστημένος, επαναπατρισμός, δράση μετεγκατάστασης, κοινωνική μέριμνα, κοινωνική ενσωμάτωση, τρομοκράτες, απειλή για την Ευρώπη, απειλή για Δύση, απειλή για τις Δυτικές αξίες, Τόμι Ρόμπινσον, Ουκρανοί, Ουκρανοποίηση, γενοκτονία των λευκών, ξενοφοβία.

Specific Neologisms: Λαθροεισβολέας, λαθρομετανάστης, λαθροεπενδυτές, επενδυτές.

Proposed keywords in Italian

Climate change

Città senza auto, fuoriuscita di sostanze chimiche, bufala del cambiamento climatico, Greta Thunberg, cambiamento climatico, riscaldamento globale, co2, disastro ambientale Strategie di adattamento, inquinamento atmosferico, perdita biodiversità, città senza auto, emissioni Co2, emissioni carbonio, impronta carbonio, prezzo carbonio, sequestro carbonio, fuoriuscita sostanza chimiche, energia pulita, tecnologie energia pulita, attivismo climatico, inganno cambiamento climatico, crisi climatiche, educazione climatica, finanza climatica, impatto climatico ecosistemi, giustizia climatica, migrazioni climatiche, moderazione climatica, politiche climatiche, rifugiati climatici, resilienza climatica, scienza, clima, Copernicus, deforestazione, siccità, eco-friendly, emissioni, ambiente, attivismo ambientale, disastro ambientale, impatto ambientale, sostenibilità ambientale, controllo erosione, tempo estremo, eventi meteorologici estremi, incendi boschivi, carburanti fossili, Friday for Future, geoingegneria, ghiacciai, temperature globali, riscaldamento globale, effetto serra, gas serra, Greta Thunberg, conversazione habitat, distruzione habitat, ondate caldo, scioglimento ghiacciai, conservazione marina, microplastiche, misure di contenimento, disastri naturali, emissioni nette zero, acidificazione oceano, sovrapesca, riduzione ozono, accordi di Parigi, fotovoltaico, foreste pluviali, iniziative riciclo, energia rinnovabile, aumento livello mare, sforzi sostenibilità, sviluppo sostenibile, energia sostenibile, Conferenza cambiamento climatico Onu, gestione rifiuti, inquinamento acqua, mancanza acqua, tempo atmosferico, modificazione tempo atmosferico, turbine eoliche, incendi boschivi.

Public health

5g, wireless quinta generazione, wireless 5 gen, alieno, agenda aliena, AlienLeaks, Anthony Fauci, Big Pharma, Bill Gates, CDC, CHD, Covid, Covid-19, vaccino Covid, Covid19, Dr Bryan Ardis, DNA, emergenza, FDA, George Soros, Johnson & Johnson, virus di Marburg, Moderna, Morgellons, Pfizer, RNA, Ron Watkins, SARS-CoV-2, transumano, transumanesimo, UAP, UFO, report UFO, fenomeno aereo non identificato, adrenocromo, eventi avversi, reazioni avverse, anticorpi, anticorpo, Astrazeneca, autismo, cattiva medicina, cattiva scienza, paralisi di Bell, big Pharma, big tech, censura big tech, biontech, sangue, coaguli di sangue, campione sangue, cancro, cellule cancerogene, arresto cardiaco, problema cardiaco, arma biologica ccp, virus ccp, scie chimiche, chemioterapia, aborto chimico, privo di sostanze chimiche, programma vaccinale infantile, biossido di cloro, condroitin solfato, sindrome affaticamento cronico, clinica Cleveland, test clinici, coagulato, codemonkeyz, tracciamento contatti, corona, corona virus, coronavirus, covid, passaporto covid, covid19, cure, cura, medicina pericolosa, responsabilità prodotti pericolosi, morte, iniezione mortale, tasso mortalità, colpo mortale, spopolamento, detox, farmaci, elettromagnetismo, epidemia, curva epidemica, mascherina, tessuto fetale, feto, feti, influenza, vaccini forzati, terapia genetica, geni, grafene, green list, porcellini d'india, educazione alla salute, rischio salute, infarto, immunità di gregge, omeopatia, pasta di cavallo, periodo di incubazione, sistema immunitario, immunizzazione, infezione, iniezioni, inoculazione, terapia intensiva, jabgate, janssen, prova di flusso laterale, iniezione letale, long covid, malaria, obblighi mascherina, mascherine, mascherine sono per gli schiavi, morbillo, aborto medico, tirannia medica, medicine, megha thakur, onda millimetrica, mitocondri, vaiolo delle scimmie, mortalità mRNA, medicina naturale, Norimberga 2, Norimberga 2.0, codice di Norimberga, processo di Norimberga per covid, no iniezione, no bambini vaccinati, pandemia, paziente zero, report pentagono ufo, compagnie farmaceutiche, pillole, plandemic, agenda politica, prevenzione, proteina spike, pro vaccino sicuro, prochoice, pro-life, prolife, radiazione, reazione, pillola di inversione, robert malone, rochelle walensky, roe contro wade, servizi igienico-sanitari, auto isolamento, effetti collaterali, sangue di serpente, veleno di serpente, variante sudafricana, spike, proteina spike, studio, morte improvvisa, mascherina chirurgica, grande abbattimento, rimozione tossine, transumano, non vaccinato, unvaxxed lives matter, unvaccinated, immunodeficienza acquisita da vaccino, vaccino, sistema di segnalazione eventi avversi da vaccino, scelta vaccinale, morti da vaccino, infortuni da vaccino, obblighi vaccinali, effetti collaterali vaccino, vittime vaccino, vaccinazione, vaccinazioni, vaccini, vax, wuhan, febbre gialla

Migrants

I bambini migranti scompaiono, extracomunitari, migrazione, immigrazione, invasione, immigrato, clandestini, migranti, straniero clandestino, Tommy Robinson, immigrati, africani, arabi, richiedenti asilo, attacchi Europa, banderizzazione, confini, bambini migranti, minaccia alla civiltà, pericolo civiltà, dislocamento, immigrati economici, migranti economici, impatto economico della migrazione, eurofobia, Grecia difende europa, Grecia sotto attacco, grande sostituzione, illegale, clandestino, clandestini, immigrato, immigrati, immigrazione, leggi immigrazione, politiche immigrazione, programmi integrazione, invasione, invasione Europa, invasore, invasione, donne a rischio migranti, migranti assassini, migranti, migrazione e sviluppo, gestione migrazione, modelli di migrazione, invasione migratoria, bambini migranti, detenzione migranti, salute migranti, diritti migranti, musulmani, migrante stupratore, rifugiato stupratore, campo profughi, crisi rifugiati, reinsediamento rifugiati, rifugiati pericolo donne, trasferiti, emigrazione, attività di reinsediamento, assistenza sociale,

inclusione sociale, terroristi, pericolo Europa, pericolo Occidente, pericolo valori occidentali, Tommy Robinson, ucraini, ucrainizzazione, genocidio bianco, xenofobia, Lampedusa, sbarchi, barche, barchini, naufragio, centro di accoglienza, ong, Centro di Primo Soccorso e Accoglienza, accoglienza, Libia, Tunisia, Mediterraneo, Mar Mediterraneo.

Proposed keywords in Polish

Climate change

Strategie adaptacyjne/środki przystosowawcze, zanieczyszczenie powietrza, utrata różnorodności biologicznej, miasta bez samochodów, emisja dwutlenku węgla, ślad węglowy, ceny emisji dwutlenku węgla, sekwestracja dwutlenku węgla, wyciek substancji chemicznych, czysta energia, technologia czystej energii, Clean energy technologies, aktywizm klimatyczny, działanie/działania na rzecz klimatu, adaptacja klimatyczna, zmiany klimatu/zmiana klimatu, adaptacja do zmian klimatycznych, oszustwo dotyczące zmian klimatu/spisek klimatyczny, kryzys klimatyczny, edukacja klimatyczna, zrównoważone finanse, wpływ klimatu na ekosystemy, sprawiedliwość klimatyczna, łagodzenie zmian klimatu/mitygacyja zmian klimatu, modelowanie klimatu, polityka klimatyczna, uchodźcy klimatyczni, odporność klimatyczna, nauka o klimacie, zmienność klimatu, Copernicus, wylesianie/deforestacja, susze, ekologiczny, emisje, środowisko, aktywizm ekologiczny, katastrofa ekologiczna, wpływ/oddziaływanie na środowisko, zrównoważony rozwój środowiska, kontrola erozji/zapobieganie erozji, ekstremalna pogoda, ekstremalne zjawiska pogodowe, pożary lasw, paliwa kopalne, Friday for Future/Młodzieżowy Strajk Klimatyczny, geoinżynieria, lodowce, globalna temperatura, globalne ocieplenie, efekt cieplarniany, gazy cieplarniane, Greta Thunberg, ochrona siedlisk, niszczenie siedlisk, fala/fale upałów, topnienie lodu/topnienie lodowców, ochrona mórz, mikroplastik, środki łagodzące, klęska żywiołowa, zerowe emisje netto, zakwaszanie oceanu, zbyt intensywne połowy/przełowienie, zubożenie warstwy ozonowej, Porozumienie paryskie, fotowoltaika, las deszczowy, inicjatywy recyklingowe/inicjatywy proekologiczne, energia odnawialna, wzrost poziomu morza, wysiłki na rzecz zrównoważonego rozwoju, zrównoważony rozwój, energia odnawialna, Konferencja Narodów Zjednoczonych w sprawie zmian klimatycznych, gospodarowanie odpadami, zanieczyszczenie wody, ziedobór wody, pogoda, modyfikacja pogody/wpływanie na pogodę, turbiny wiatrowe, niekontrolowany ogień.

Public health

5G, technologia bezprzewodowa piątej generacji, technologia bezprzewodowa 5-tej generacji, obcy/kosmita, agenda obcych, Anthony Fauci, Big Pharma, Bill Gates, CDC, CHD, COVID, COVID-19, szczepionka na COVID/szczepionka przeciw COVID, COVID19, Dr Bryan Ardis, DNA, nagły wypadek, FDA, George Soros, Johnson & Johnson, wirus Marburg, Moderna, Morgellony/Morgellonowie, Pfizer, RNA, Ron Watkins, SARS-CoV-2, transczłowiek, transhumanizm, UAP, UFO, raport o UFO, Niezidentyfikowane zjawiska latający, adrenochrom, zdarzenia niepożądane, działanie niepożądane, przeciwciała, antyszczepionkowy, antyszczepionkowiec, przeciwciało, astrazeneca, autyzm, paramedycyna/medycyna niekonwencjonalna, para

nauka, Porażenie Bella, big pharma, big tech, cenzura big techów, biontech, krew, skrzepy/zakrzepy, próbka krwi, rak/nowotwór, komórki nowotworowe, zatrzymanie akcji serca, problem kardiologiczny/problem sercowy, broń biologiczna ccp, wirus ccp, chemtrails/smugi chemiczne, chemoterapia, aborcja chemiczna, bez chemii, harmonogram szczepień dla dzieci, dwutlenek chloru, siarczan chondroityny, zespół chronicznego zmęczenia, cleveland clinic, badania kliniczne, zakrzep, codemonkeyz, śledzenie kontaktów, korona, koronawirus, covid, paszport covidowy/paszport cowidowy, covid19, leki/legarstwa, lek/lekarstwo, niebezpieczny lek, odpowiedzialność za produkt niebezpieczny, śmierć, zatrzyk śmierci, śmiertleność, śmiertleny zastrzyk, depopulacja, detoks, dna, narkotyki, elektromagnetyzm, epidemia, krzywa epidemiologiczna, maska ochronna/maseczka ochronna, tkanka płodowa/tkanka płodu, płód, płody, grypa, przymusowe szczepionki, terapia genowa, geny, garfen, świnki morskie, edukacja zdrowotna, ryzyko dla zdrowia, zawał serca, odporność stadna/odporność zbiorowiskowa, homeopatyczny, maść końska, okres wylęgania/okres inkubacji, układ odpornościowy, immunizacja/uodpornienie, infekcja, zastrzyki, szczepienie ochronne, oddział intensywnej terapii, ukłucie/szczepionka, janssen, zabójczy strzał, badanie metodą przepływu bocznego, przewlekły Covid, malaria, nakaz noszenia maski, maski, maski są dla niewolników, odra, aborcja medyczna, tyrania medyczna, leki, meghathakur, fala milimetrowa, mitochondria, małpia ospa, śmiertelność, mRNA, medycyna naturalna, norymberga 2, norymberga 2.0, kodeks norymberski, procesy norymberskie za covid, pandemia, pacjent zero, raport Pentagonu o UFO, firmy farmaceutyczne, pigułki, plandemia, program polityczny, zapobieganie/profilktyka, białko kolca, prochoice, pro-life, prochoice, prochoice, prolife, prolife, promieniowanie, reakcja, robert malone, rochelle walensky, roe przeciwko wade, warunki sanitarne, samoizolacja, skutki uboczne, krew węża, jad węża, wariant południowoafrykański, kolec, białko kolczaste, zatrzymac ludobójstwo, badanie, nagła śmierć, chirurgiczna maska, wielki ubój, usuwanie toksyn, transludzki, nieszczepiony, ~~vaids,~~ nieszczepione życie ma znaczenie, nieszczepiony, szczepionka, system zgłaszania działań niepożądanych szczepionek, wybór czy zostać zaszczepionym, zgonów poszczepienne, urazy poszczepienne, upoważnienie do zaszczepienia, skutki uboczne szczepionki/niepożądany odczyn poszczepienny/nop, ofiar szczepionek, szczepienie, szczepienia, szczepionki, śmierć szczepionkowa, wuhan, żółta febra.

Migrants

Afrykanie, Arabowie, osoby ubiegające się o azyl, atakować Europę, banderyzacja, granice, dzieci-migranci, zagrożenie cywilizacyjne, niebezpieczna cywilizacja, przemieszczenie, imigranci ekonomiczni, emigranci ekonomiczni, ekonomiczne skutki migracji, eurofob, grecja_chroni_europę, grecja_atakowana, wielkie zastąpienie, nielegalny, nielegalny obcy, nielegalni obcy, imigrant, imigranci, imigracja, prawo imigracyjne, polityka imigracyjna, programy integracyjne, inwazja, inwazja na Europę, najeźdźcy, najechać, migranci zagrażają kobietom, migrant zabójca, migranci, Migracja i rozwój, zarządzanie migracją, wzorce migracji, inwazja migracyjna, dzieci migrantów, zatrzymanie migracyjne, zdrowie migrantów, prawa migrantów, muzułmanie, migrant-gwałciciel, uchodźca-gwałciciel, ~~rapefugees,~~ obóz dla uchodźców, kryzys uchodźczy, przesiedlenie uchodźców, uchodźcy zagrażają kobietom, relokacja, ponowna migracja, akcja przesiedleńcza, opieka społeczna, integracja społeczna, terroryści, zagrażać Europie, zagrażać Zachodowi, zagrażać zachodnim wartościom, tommy robinson, Ukraińcy, ukrainizacja, ludobójstwo białych, ksenofobia.

Proposed keywords in Romanian

Climate change

Strategii de adaptare, Poluarea aerului, Poluarea atmosferică, Pierderea biodiversității, Distrugerea biodiversității, Orașe fără mașini, Emisii de dioxid de carbon, Emisiile de dioxid de carbon, Amprenta de carbon, Prețul carbonului, Stocarea carbonului, Deversare de substanțe chimice, Energie curată, Tehnologii energetice curate, Activism climatic, Acțiune climatică, Adaptare la schimbările climatice, Schimbări climatice, Adaptare la schimbările climatice, Caniculă privind schimbările climatice, Criza climatică, Educație climatică, Finanțarea climei, Impactul climei asupra ecosistemelor, Justiție climatică, Atenuarea schimbărilor climatice, Modelarea climatică, Politica climatică, Politica de mediu, Refugiați de mediu, Reziliența climatică, Știința mediului, Variabilitatea climei, Copernicus, Defrișări, Secete, Ecologic, Emisii, Mediu, Activism de mediu, Dezastru ecologic, Impact asupra mediului, Impact climatic, Impact asupra climei, Sustenabilitatea mediului, Mediu durabil, Combaterea eroziunii, Vreme extremă, Fenomene meteorologice extreme, Incendii de pădure, Combustibili fosili, Vineri pentru viitor, Geoinginerie, Ghețari, Temperatura globală, Încălzirea globală, Efectul de seră, Gazele cu efect de seră, Greta Thunberg, Conservarea habitatelor, Distrugerea habitatelor, Valuri de căldură, Topirea gheții, Conservarea mediului marin, Microplastice, Măsuri de atenuare, Dezastre naturale, Emisii nete zero, Acidificarea oceanelor, Pescuitul excesiv, Epuizarea stratului de ozon, Acordul de la Paris, Energie fotovoltaică, Fotovoltaice, Pădurea tropicală, Inițiative de reciclare, Energie regenerabilă, Creșterea nivelului mării, Eforturi de durabilitate, Dezvoltare durabilă, Energie durabilă, Conferința Națiunilor Unite privind schimbările climatice, Gestionarea deșeurilor, Poluarea apei, Scăderea apei, Vreme, Schimbarea vremii, Turbine eolienne.

Public health

5G, a 5a generație wireless, wireless 5 gen, Alien, Alien Agenda, AlienLeaks, Anthony Fauci, Big Pharma, Bill Gates, CDC, CHD, COVID, COVID-19, COVID vax, COVID19, Dr. Bryan Ardis, DNA, Urgență, FDA, George Soros, Johnson & Johnson, virusul Marburg, Moderna, Morgellons, Pfizer, ARN, Ron Watkins, SARS-CoV-2, Transuman, Transumanism, UAP, OZN, raport OZN, Fenomene aeriene neidentificate, adrenocrom, evenimente adverse, reacții adverse, anticorpi, anti-vax, anti-vaxx, anti-vaxxer, antivaxx, antivaxxer, anticorpi, anticorpi, anticorpi, antivaxx, antivaxxer, astrazeneca, autism, medicină proastă, știință proastă, paralizia clopotelor, big pharma, big tech, cenzura big tech, biontech, sânge, cheaguri de sânge, probă de sânge, cancer, celule canceroase, stop cardiac, problemă cardiacă, ccp bioweapon, ccp virus, chemtrails, chimioterapie, avort chimic, fără chimicale, calendarul vaccinurilor pentru copii, dioxid de clor, sulfat de condroitină, sindromul oboselii cronice, clinica Cleveland, studii clinice, clotshot, codemonkeyz, contact tracing, corona, corona virus, coronavirus, covid, covid passport, covid19, cure, leac, leac, medicament periculos, răspundere pentru produse periculoase, deces, death jab, mortalitate, mortalitate, death shot, depopulare, detoxifiere, ADN, medicamente, electromagnetism, epidemie, curbă epidemică, mască de față, țesut fetal, fetus, fetuși, fetuși, a cincea generație de wireless, gripă, vaccinuri forțate, terapie genică, gene, gene, grafenă, sosiri verzi, listă verde, cobai, educație pentru sănătate, risc pentru sănătate, atac de cord, imunitate de turmă, homeopatic, cremă de cal, perioadă de incubație, sistem imunitar, imunizare, infecție, injecții, inoculare, unitate de terapie intensivă, jab, jabgate, janssen, kill shot, test de flux lateral, injecție letală, să lăsăm cadavrele să zacă pe podea, long-Covid, malarie, mandate de măști, măști, măștile sunt pentru sclavi, rujeolă, avort medical, tiranie medicală, medicamente, meghathakur, unde milimetrice, mitocondrie, pojarul maimuței, mortalitate, ARNm,

medicină naturală, nuremberg 2, nuremberg 2. 0, codul de la nürnberg, procesele de la nürnberg pentru covid, no jab, nu copii vaccinați, pandemie, pacient zero, pentagon ufo report, companii farmaceutice, pastile, plandemic, agenda politică, prevenție, vârf de proteină, vaccin pro-SAFE, prochoice, pro-viață, pro-life, prochoice, pro-life, prolife, prolife, radiații, reacție, sosiri roșii, lista roșie, pilula de inversare, robert malone, rochelle walensky, roe v wade, igienă, autoizolare, efecte secundare, sânge de șarpe, venin de șarpe, varianta sud-africană, spike, spike protein, opriți genocidul statuii, studiu, moarte subită, mască de față chirurgicală, marea sacrificare, eliminarea toxinelor, transuman, unvax, unvaxxed lives matter, nevaccinat, *vaids*, vaccin, sistemul de raportare a efectelor adverse ale vaccinurilor, alegerea vaccinului, decese prin vaccinare, leziuni prin vaccinare, mandate de vaccinare, efecte secundare ale vaccinurilor, victime ale vaccinurilor, vaccinare, vaccinuri, vax, vax moarte, vaxx, *watch the water*, wuhan, febra galbenă

Migrants

Africani, arabi, solicitanți de azil, atacă Europa, banderizare, granițe, copii migranți, amenințarea civilizației, pericolul civilizației, strămutare, imigranți economici, migranți economici, impactul economic al migrației, eurofobie, grecia_apară_europa, grecia_sub_atac, marea înlocuire, marea relocare, ilegal, străin ilegal, străini illegali, imigrant, imigranți, imigrare, imigrație, Legea imigrației, Politici de migrație, Programe de integrare, invazie, invazie Europa, invadatori, invadare, invadează, pericol migranți femei, ucigaș de migranți, migranți, Migrație și dezvoltare, Managementul migrației, Modele de migrație, invazie migratorie, Copii migranți, Detenția migranților, Sănătatea migranților, Drepturile migranților, Musulmani, migrant violat, refugiat violat, refugiați violați, tabără de refugiați, Criza refugiaților, relocare refugiați, pericolul refugiaților, femei refugiate, relocare, remigrație, acțiune de relocare, asistență socială, incluziune socială, teroriști, amenințare Europa, amenințare vest, amenințare valori occidentale, tommy robinson, ucraineni, ucrainizare, genocid alb, xenofobie.

Proposed keywords in Spanish

Climate change

calentamiento global, engaño del cambio climático, cambio climático, desastre ambiental, ciudades sin coches, , co2, greta thunberg, derrame de sustancias químicas, cambio cromático, timo climático, presas, sequía, temperaturas, inundación, agenda 2030, nivel del mar, chemtrails, geoingeniería, planeta B, ciclón, tormenta, huracán, HAARP, DANA, insectos, emergencia climática, pantano, incendio, reciclaje, glaciar, temperaturas extremas, embalses, paneles solares, hidroeléctrica, molinos de viento, electricidad, baterías, ola de calor, coches eléctricos, cobalto, contaminación, temperatura, explosiones de baterías, sueño verde, litio, hidrógeno, estela de condensación, combustión, domo atmosférico, atmósfera, CO2, farsa climática, hielo, granjas, ganadería, cargarse el campo, carne, yoduro de plata, cambio climatico antropogénico, agua, racionamiento, fumigación, energía verde, destrucción de lluvias, afrenta 2030, climodemia, climademia, ecodictadura, ecofascismo, ecorrégimen dictatorial, farsemia climática, Paranoia climática, Timo climático, 2030 agenda

Public health

5g, aborto, adrenocromo, Agenda alienígena, anticuerpos, antivacunas, antivax, antivaxx, aplanando la curva, arma biológica del ccp, asintomático, astrazeneca, autoaislamiento, base de datos vaers, Bill Gates, biotecnología, calendario vacunal infantil, cáncer, cardiopatía coronaria, cepa lambda, clínica de cleveland, coágulo, coalición para la preparación ante epidemias, cobayas, codigo de nuremberg, códigomonkeyz, colapsado, conejillos de indias, convención de estambul, corona, coronavirus, COVID-19, crímenes contra la humanidad, curar, curva epidémica,, disparo mortal, efectos secundarios,, elección de vacuna, códigomonkeyz, colapsado, conejillos de indias, EPP, eventos adversos, feto, fiebre amarilla, golpe de muerte, gran censura tecnológica, gran farmacéutica, gran tecnología, gripe de wuhan, homeopático, inalámbrica de quinta generación, influencer antivacunas, inmunizaciones, Jabgate, jansen, johnson y johnson, juicios de nuremberg por covid, la gran matanza, la inmunidad de grupo, las mascaras son para los esclavos, vacunas, vacunas forzadas, VAERS, variante mongola, variante sudafricana, vax, veneno de serpiente, veneno intencional, víctimas de la vacuna, viruela de dinero, viruela del simio, virus chino, virus de wuhan, virus del pcc, wuhan, las vidas no vacunadas importan, lesiones por vacunas, mala ciencia, mandato, mandatos de máscara, mandatos de vacunación, medicamento, medicina mala, medicina natural, medicina peligrosa, mimo, mira el agua, Roe contra Wade, sabio, sangre de serpiente, sarampión, SARS-CoV-2, SB-277, sin jab, sistema de notificación de efectos adversos, tiranía médica, Tiro de muerte,, vacuna, moderno, muerte vax, muertes por vacunas, mutación viral, ningún niño vacunado, no vacunado, Núremberg 2, Núremberg 2.0, paciente cero, parálisis de campanas, paro cardiaco, pasaporte covid, PAU, período de incubación, período de incubación, pfizer, pinchazo, Planificación familiar, Pro vida,, productos farmacéuticos, preselección, prueba de flujo lateral, rastreo de contactos, reacción adversa, Reacciones adversas, responsabilidad por productos peligrosos,

Specific neologisms: aRn, bozal, covidiano, despiertos, disidentes , enflautamientos, inoculados, kakunas, masónico, mundialista, neonega , NOM, Nuevo Orden Mundia), Nueva Normalidad, Opoficción, plandemia, Repentinitis, Tapabocas, Úrsula Von der Pfizer, Vacuñao, Zombi covidiano, Zoociedad

Migrants

Migración, migrantes, niños migrantes, inmigración, extranjeros ilegales, niños migrantes desaparecen, inmigrantes, invasión, marroquíes, moros, musulmanes, árabe, Alá es grande, ilegales, magrebí, islamismo, pateras, barco nodriza, mahoma, ramadan, meca, inmigrante okupa, africanos, invasión islámica, mezquita, negro, persona negra, burka, ayudas, paguitas, puertas ilegales, reemplazo social, terrorismo.