AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "AI FOR COUNTERACTING DISINFORMATION"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Deutsche Welle** (DW)

This report is part of the deliverable D2.3 - "*AI technologies and applications in media: State of Play, Foresight, and Research Directions*" of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Copyright

© Copyright 2022 AI4Media Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the AI4Media Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.
All rights reserved.

AI for counteracting disinformation

Current status

The phenomenon of online ***disinformation*** has evolved since around 2010 and is defined as "*false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit*"[1]. While the spreading of false or manipulative information has occurred for centuries, the significance and negative impact of this activity/phenomenon has increased with the emergence of social media, digital information production and consumption as well as advances in technology, including Artificial Intelligence. Although the effects of online *disinformation* have been addressed by fact checking and verification specialists for almost one decade, events such as the US presidential election in 2016 and the Covid-19 pandemic have brought the significant risks for society, democracy, and individuals to mainstream, academic and political attention.

Many different stakeholders are engaged in counteracting *disinformation:* not only social media platforms, fact-checking initiatives, open-source intelligence specialists and news media organisations, but also academia, governments, educational institutions and civil society initiatives. One or more of the following interrelated approaches come generally into use for the purpose of counteracting *disinformation*:

- Verifying content (e.g., videos, photos, or posts) and social media accounts;
- Checking statements (claims) made by public figures against facts;
- Identifying disinformation narratives/stories in social media;
- Conducting media literacy and education/training programmes;
- Establishing self-regulation schemes and regulatory frameworks;
- Developing counteractive methods, technologies, and support tools.

For many years, AI technologies have played an important role in counteracting *disinformation*, especially in tools and systems used for content verification, fact-checking and social media disinformation analysis (Figure 1). The need for AI support has recently increased: On one hand the frequency and scope of disinformation has grown to a level that manual approaches cannot handle. On the other hand, adversaries use advanced AI technologies and automation for targeted campaigns, content manipulation or synthetic media production, which in many cases are only detectable with AI-powered systems.

[1] There are many definitions for *disinformation.* We have chosen the one first defined by Wardle, Claire, and Hossein Derakhshan. "INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making." (2017). This definition is also used in the report issued by the European Commission's High Level Expert Group on Fake News and Online Disinformation (Source: European Commission. "A Multi-Dimensional Approach to Disinformation: Report of the Independent High-Level Group on Fake News and Online Disinformation." (2018).

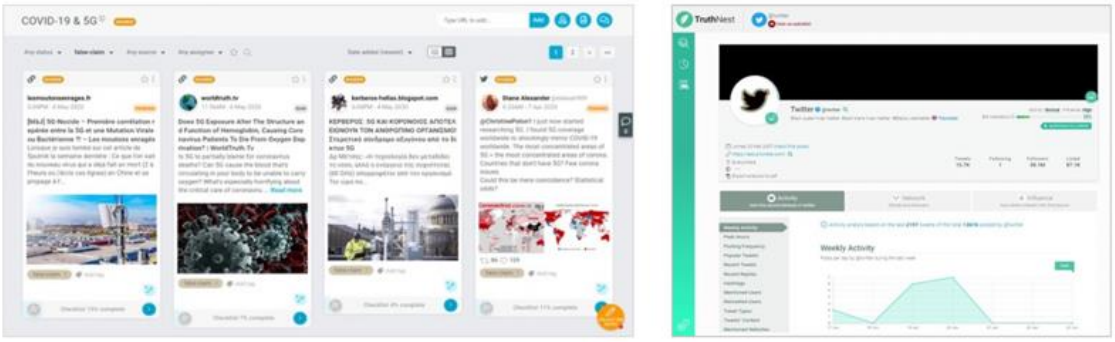

Figure 1: Examples for current support tools to counteract disinformation: Truly Media[2] (left) and TruthNest[3] (right).

Despite multiple AI based support functions being available, there are several shortcomings, limitations, and missing elements to ensure long-term success in counteracting online disinformation. The ***current areas of limitation*** are presented in Figure 2 and discussed below.

AI functions and solutions. Most AI solutions today are good at specific, narrowly defined tasks that can help to identify disinformation elements and claims in social media. Examples are reverse image and geo search, detection of bot accounts, comparing digital content for detecting changes/manipulation (e.g., text, video, and photos), detecting deepfake face-swaps in videos or photos, analysing audio to detect manipulation, scanning large data repositories for specific keywords, and analysing certain aspects of content in social networks and relationships between accounts. **What AI cannot yet deliver for practitioners in fact checking and verification is the detection and analysis of entire, complex disinformation narratives, handling more complex tasks across social/digital platforms and involving multimodal data types, and covering all aspects and types of synthetic media detection/manipulation.**

Underlying datasets. Although there is research into multimodal approaches[4], at present, the underlying datasets for AI solutions that are *practically* used to counteract *disinformation* often relate to one data type (e.g., either text, video, images, or audio) or only one content source (e.g., Twitter). Further, it can be generally challenging to collect quality datasets[5] and in many cases, they are related to only one domain (e.g., politics). In addition, the usage of certain datasets is subject to ethical concerns and many datasets are difficult to obtain (and to maintain for longer periods of analysis) in the light of regulation, Intellectual Property Rights (IPR), ethical requirements and the terms and conditions from media platforms. **Despite regulatory initiatives, there are currently limitations regarding datasets that are easily and openly available to those researchers/developers that produce AI solutions against *disinformation*,**

[2] Truly Media: https://www.truly.media/
[3] TruthNest: https://www.truthnest.com/
[4] An example for such research is: A. Giachanou, G. Zhang, and P. Rosso. "Multimodal multi-image fake news detection." 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2020.
[5] F. Torabi Asr, and M. Taboada."Big data and quality data for fake news and misinformation detection." *Big Data & Society* 6.1 (2019): 2053951719843310.

but which are yet ethically and legally compliant. Further, there are requirements for multimodal, cross-platform and multi-domain datasets.

Human-AI Collaboration. Current AI-powered tools largely provide machine support for humans who need to conduct complex fact-checking and verification workflows. They enable otherwise (humanly) impossible analysis (detection) and reduce time, stress, or cost, but are largely based on manual human oversight and manual pre-detection (e.g., presenting the AI-function with a video in which a deepfake face-swap is suspected). Where semi or full automation could be technically achieved, there is usually associated human distrust in the capability of the AI-powered solutions to make an accurate and/or contextually acceptable decision. **So far, there are limited approaches for true human-machine collaboration or (socially acceptable) forms of automation.**

Responsible and Trustworthy AI. Most current AI functions and services used in the context of counteracting *disinformation* are accuracy and performance oriented, with little or no information given (by third-party providers) about the AI function/model itself, its legal compliance or measures taken to provide explainability, to mitigate bias or to increase reliability/robustness. Such limitations related to *Responsible/Trustworthy AI* can impact on successfully counteracting *disinformation* for various reasons: 1) In this domain, many decisions are related to the comparatively vague and complex concept of "truth". 2) Unlike some commercial AI application domains, the work of fact checkers, verification specialists or journalists is also influenced by immaterial aspects (e.g., societal and public value systems). 3) It is typical for fact checkers, verification specialists, journalists, or open-source intelligence analysts to be curious, show attention to detail and question any presented information prior to further using it. 4) All stakeholders (specialist staff, editorial managers or the board of management of the organisation) are bound by editorial control rules (e.g., dual control principles), journalistic codes and specific organisational values as well as legal frameworks related to publishing/journalism. **Current shortcomings related to the integration of Trustworthy AI principles into AI functions that help to counteract *disinformation* can raise (justified) questions among affected stakeholders and therefore reduce their acceptance and use by related specialists and/or their organisations.**

Usability and User Experience. While there are (and will be) many useful stand-alone AI functions or solutions available, it remains difficult to transform their technical output/predictions into suitable user interfaces within the tools used by practitioners and to create satisfactory user experiences that are adequate for non-technical users in counteracting *disinformation*. **This difficulty is due to a gap of knowledge and funding resources that occurs between any existing (or future) AI function presented in the format of code, a dockerised container or Application Programming Interface (API) and the user interface of an end-user tool that makes the result of this function usable.**

End user tools and systems. Various specialist initiatives and projects conducted research and delivered AI-functions for counteracting *disinformation* at the level of research outcomes, piloted prototypes, or open-source solutions. The (global) supply market is highly fragmented and consists of many small (or even ultra-small) players. There are few European commercial solutions in the market specifically targeted at the fact checking and verification workflow (e.g.,

such as Truly Media[6]). Other off-the-shelf products in that direction are headquartered outside Europe, are not yet commercially mature enough or target a different customer base (e.g., corporate reputational analysis and PR). **There remains a lack of tailored, end-to-end, and mature products for all relevant stakeholders that are suitable (in business terms), accessible (in financial and integration terms) and sustainable (in both maintenance and energy terms).**

Figure 2: Areas of current limitation for counteracting disinformation with AI.

Research challenges

While existing AI approaches and tools are already invaluable to counter *disinformation*, there is a need for improvement in areas such as AI technology advancement, datasets, human-AI collaboration, trustworthiness, user interface transfer and industry products. These research areas are presented in Figure 3 and discussed below. To develop AI systems for countering disinformation that have this wider capacity, further research is needed over the next decade to overcome limitations. This research is required from diverse academic fields, related to technology, business, and society. Due to the fact that this chapter focuses on AI technologies, other important research challenges are not covered, e.g., cultural, psychological or cognitive aspects. The following research areas and challenges should be addressed:

AI Technology Advancement: This research area relates to next generation AI approaches and functions that fill current technology gaps in counteracting *disinformation*. Examples for research subjects are:

- Multimodal content analysis (e.g., image with integrated text)
- Cross-platform content and network analysis
- Linguistic and country-specific environment analysis
- Detection of content manipulation by means of synthetic media
- Automatic synthetic content detection/flagging
- Dynamic AI-updates in counteraction tools (to match disinformation actors)
- Early detection of arising disinformation narratives/elements
- Causal, contextual, and cultural analysis of complex statements

[6] Truly Media: https://www.truly.media/

- Analysis of complex narratives / disinformation stories over time
- Automatic identification of check-worthy, potentially harmful elements
- Integrated analysis with Blockchain based authentication approaches.

Next Generation Datasets: This research area relates to next generation datasets used for the training and evaluation of AI functions that help to counteract *disinformation*. It involves technology research (e.g., synthetic data) and societal research (e.g., policies enabling long-term access to social media platform data). Examples for research subjects include:

- Multimodal and multilingual datasets
- Cross platform datasets
- Datasets that enable early or even real-time detection
- Synthetic datasets (overcoming issues of real datasets)
- Legal, ethical and IPR compliance certification for datasets
- Regulated datasets for specific uses/users (public value)
- Specialised datasets for *disinformation* detection purposes.

Human-AI Collaboration and Automation: This research area relates to enabling true human-AI collaboration and acceptable automation of fact checking and verification workflows in terms of journalistic/content environments, legal, ethical, and business issues. The research involves the fields of AI technology, human-computer interaction, interface design, AI-based product design and trustworthy AI. Examples for research subjects include:

- Automatic filters to select suspicious content
- Automation & collaboration approaches to fact checking/verification
    - Role of humans / human-in-the-loop / oversight
    - Workflows with no, minimal, semi or full automation
    - Seamless conceptual integration of the above
- Resolution of editorial/legal responsibility conflicts (human vs machine)
- Issues of censorship and freedom of speech related to the use of AI
- Issues of editorial control, journalistic values, and legal frameworks
- Role of Trustworthy AI in Human-AI collaboration in *disinformation* domain
- Characteristics of “acceptable” automation in the *disinformation* domain.

Trustworthy AI Capability: This research area relates to increasing the overall transparency of AI functions used in the context of counteracting *disinformation* and integrating specific trustworthy AI approaches/tools to enable a responsible and accepted use of AI in this field (and better human-AI collaboration, see point 3 above). Examples for research subjects include:

- Role of transparent/trustworthy AI in the acceptance of AI tool support
- Tailored AI transparency certifications (provider, model, data, legal)
- Tailored Trustworthy AI certifications (explainability, fairness, robustness)
- Translation of trustworthy AI output for interfaces / non-technical users
- Balancing decisions between effectiveness and trustworthiness
- Can transparency/trustworthy elements enable full automation?

- How to avoid misuse of AI technologies employed against *disinformation.*

Function-to-Interface Transfer: This research area relates to bridging the knowledge gap between a delivered AI function and the user interface of an end-user tool that makes the result of this function well usable for non-technical users. Examples for research subjects include:

- Expertise and staff roles to overcome this challenge (user-side)
- Alternative ways of presenting output of AI functions (provider-side)
- Translation of AI output for interfaces and into user language
- Translation of trustworthy AI output for interfaces / non-technical users
- Personalised approaches: matching AI affinity/expertise of end users
- Dashboard approaches for AI analysis outcomes
- Reduction of complex AI analysis outcomes.

Tailored European Products: This research area relates to the market for AI-powered tools used in fact checking and verification workflows and ways of enabling dedicated, tailored European, end-to-end products for this purpose that are suitable and accessible for large and small European professional stakeholders (e.g., fact checking organisations, media companies, self-employed journalists, or civic initiatives). Examples for research subjects include:

- Existing AI-powered tools/functions and their providers
- Multilingual products
- Product characteristics for realistic adoption by users/organisations
- Opportunities/barriers: European public sector / public service products
- Opportunities/barriers: European commercial products
- Multi-faceted products: one AI-powered back end with multiple front ends.

Figure 3: Suggested research areas to support the counteracting of disinformation with AI.

The scenario below illustrates a vision for counteracting *disinformation* in 15-20 years from now, from the perspective of a media industry stakeholder and related to the six research areas suggested above. To realise this vision, it is important to achieve improvements in the aforementioned areas. The scenario also indicates the overall societal and media impact that the successful realisation of such research activities could have.

Societal and media industry drivers

Vignette: Counteracting disinformation in the 2030s with the AI-powered CADI-Tool

Carmen is a freelance medical journalist, working in the media environment of the late 2030s. During her 20-year career in this sector, she has seen continued growth of online *disinformation*, affecting all genres of published content on any platform. Not surprisingly, this led to major public upskilling programmes for information workers, journalists, pupils, and the public as well as regulatory measures and global agreements between media platforms and governments. At the same time, Carmen saw the development and uptake of complex AI-powered support systems for dealing with and counteracting *disinformation* daily.

For over 10 years now, Carmen has had a single-user subscription to a web-based product called *CADI*, which provides similar features as other commercial systems available that also integrate with corporate content systems. All these AI-powered systems for counteracting *disinformation* have in common that they are widely adopted and socially accepted, make use of synthetic or trust-certified datasets, automatically update to state-of-the-art functions (also to keep up with *disinformation* adversaries), come with transparency and trustworthy AI certifications tailored to this domain, provide easy-to-grasp assistance via personalised, visually dynamic and flexible end-user interfaces, based on a user experience that is driven by seamless human-AI collaboration, involving a high level of workflow automation and flexible levels of human oversight.

All types of information workers, including Carmen as a freelancer, can easily deal with all types of disinformation related workflows and tasks: from verifying content items, to checking claims against facts and analysing complex social media narratives, including those that are rapidly emerging.

It is Carmen's first task of the day to check the information agenda. She will then research, produce and submit by the end of the day a video story on emerging reports about a virus outbreak in a neighbouring country. Checking the news agenda is quickly done with her personalised *CADI* dashboard, already set to her preferences (mid-level information detail and low-level technology affinity). Carmen added two required languages and geographic regions, to achieve cultural and linguistic analysis matches, as well as the required content keywords related to her medical topic. Carmen quickly glances over the resulting data visualisations, showing in an integrated way the breaking news coverage, trending social stories around it, suspected disinformation narratives, already debunked claims and a list of key media items that are either shown as suspicious or already verified by other information workers. The latter list is divided into fully synthetic, synthetically manipulated, and non-synthetic media items. Based on the news (and disinformation) overview that she had obtained earlier via the dashboard, Carmen uses the *CADI* system to conduct a further universal search job across multiple media platforms and media types such as text, video, images, or audio. Apart from reporting the news of the virus outbreak, she will also contrast official statements with circulating theories and highlight selected *disinformation* elements as it is common practice. The *CADI* system acts as an early-warning system in this breaking news situation and automatically suggests "suspicious" statements, narratives, and media items for her (human) review and for use in fact-checking

reports. Carmen is particularly pleased to have this function as it took over a decade for AI systems to identify what humans might regard as "suspicious". At the same time, to avoid overload, the system has automatically deleted several *disinformation* elements in her results feed, based on transparent, certified approaches she is fully aware of.

While Carmen is accepting some of *CADI*'s decisions related to *disinformation* elements (as she knows the AI-technology is trustworthy and has been certified), she decides to follow up the transparency and trustworthy AI information provided for others. One reason for checking some specific aspects manually is that the media editor to whom she submits her video report will do the same for the purpose of editorial control. In particular, she double checks the AI system's decision that a popular video featuring the health minister of the neighbouring country is a deepfake, because getting this wrong may have legal implications for the media company publishing her video.

After having spent a little too long on research, Carmen now quickly produces the video. Prior to finalising the video, she asks *CADI* for an update of both news and *disinformation* developments. Luckily, there weren't any major developments. She presses the submit button, leaves her desk and while she walks home, Carmen's mind wanders back to the early days of her career in the early 2020s. She can hardly believe that counteracting *disinformation* was a difficult, time-consuming and complex workflow, sometimes with limited success, not possible in live or breaking news situations, hindered by language barriers and conducted by a few specialists in the media sector, who played a game of catch-up with the ever-advancing disinformation actors.

Future trends for the media sector

Further technical advances will also help to drive the production and distribution of false or misleading information. However, conducting extensive multidisciplinary research into the next generation of AI-powered solutions as described above can lead to a turning point, giving way to a range of opportunities and benefits for the media industry – as well as other related domains that can benefit from similar functions/tools. The following future trends can be anticipated:

- Increasing acceptance of and trust in AI-powered tools for counteracting *disinformation* in media and society through ***transparency and trustworthy AI certifications***, that can be easily used by end-users, stakeholders, journalistic codes of conducts and regulatory frameworks.
- Removing barriers to workflow automation in an area that involves the complex concept of "truth" by establishing successful ***Human-AI collaboration models***.
- Providing significantly more information workers in media and society with access to powerful and suitable ***support tools*** to deal with and counteract *disinformation*.
- Enabling ***earlier or even real-time detection***, which solves the societal problem of reactive fact-checking and verification after *disinformation* has already spread.
- Allowing creative media and information workers to focus on their core tasks, while the complex, time-consuming ***workflows*** related to analysing *disinformation* are largely ***automated*** – in a responsible, trusted way.

Goals for next 10 or 20 years

By 2040, all relevant stakeholders will be involved in counteracting *disinformation*, ranging from information workers in all domains (media, government, business, and society), to members of the public and (global) media platforms – benefiting from widely automated, trusted and seamlessly integrated counteraction workflows, early identification of *disinformation*, and significantly more impactful, accessible, and user-friendly support tools.

By then, AI-powered support systems for counteracting *disinformation* will have capabilities such as

- Multimodal and cross-platform analysis;
- Linguistic, country, culture, and context analysis;
- Full synthetic content and synthetic manipulation analysis;
- Automatic and early (real-time) detection of *disinformation;*
- Automatic detection of check-worthy items, claims or narratives
- Seamless and flexible human-AI collaboration workflows;
- Certified information related to Transparency, Trustworthy AI, Datasets;
- Automatic technology upgrades to match tools of disinformation actors;
- Interoperability with content authentication systems (e.g., Blockchain-based).

The above capabilities are enabled on one hand by major advances in realising Trustworthy AI (accurate, performant technologies with yet trusted and explained outcomes) and on the other hand by widely available, ethically, and legally certified datasets that are needed for AI model training and evaluation for such functions. In combination, this is the basis for and can enable successful human-AI collaboration. Stand-alone AI technologies, functions and services will be integrated into tailored, user-friendly, and accessible support products targeted at information workers, which are widely available as off-the-shelf applications, affordable web-based subscription services or for seamless integration into corporate content management systems and their user interfaces. Specific public subsidy and co-funding programmes are in place to ensure access to these high-end, AI-powered systems for all types of users who need them. Core back-end technologies connect with multiple front ends for different user domains, where front ends are featuring high degrees of multi-faceted personalisation, dashboard views, fine-grained visualisation of AI-predictions and easy-to-grasp (and easy-to-accept) trust-related information, such as AI certifications or explanations for AI actions that can be understood by non-technical users).

On the way towards achieving these ultimate goals around the year 2040, the following milestone points can be defined as interim goals for the years 2025, 2030 and 2035 (Figure 4):

Milestone 5 years: By 2025, apart from continued advances in AI technology, support products and user experience design, the AI functions provided for the purpose of counteracting *disinformation* will be certified in terms of Trustworthy AI and based on tailored, ethically, and legally compliant (certified) datasets.

Milestone 10 years: The developments during the 2020s formed the necessary baseline for achieving more (acceptable) automation in fact-checking and verification workflows as well as

true human-AI collaboration, which is largely in place by 2030. This achievement is also driven by the by then more powerful AI analysis capabilities, advances in datasets, and widely accepted, accessible support tool products with a strong focus on AI-results usability.

Milestone 15 years: By 2035, the progress described above now begins to show real impact, leading to a significant reduction of both *disinformation* itself and its negative effect on media and society. This is driven by a combination of factors: further technical advance and excellence of Trustworthy AI functions and underlying datasets in use, wide availability of user-friendly and accepted AI-support tools (including public subsidies) and the implementation of seamless human-AI collaboration, enabling largely automated workflows if and where chosen.

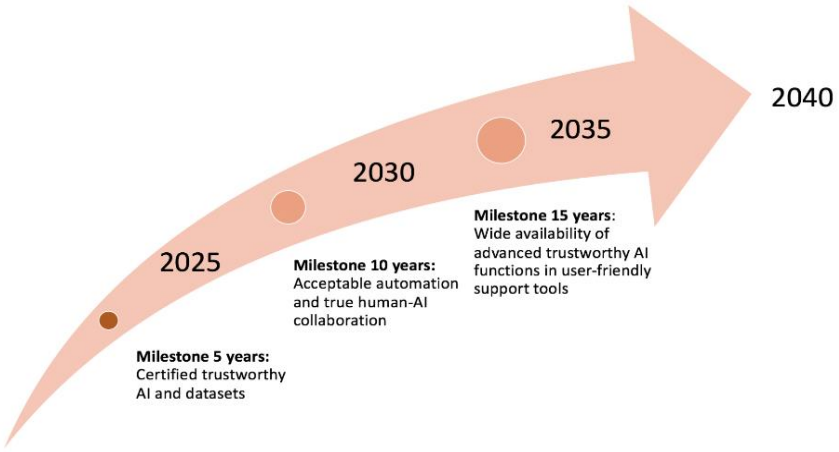

Figure 4: Milestones up to 2040 for counteracting disinformation with AI.

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu