AI4media

DW Innovation and ATC iLab

AI Support Needed to Counteract Disinformation

White Paper - October 2022

Executive Summary

The phenomenon of online disinformation has evolved since around 2010. Both scope and impact are expected to increase, also due to advances in Artificial Intelligence (AI). Many different societal stakeholders are engaged in counteracting disinformation through a range of approaches. One of them is the development of AI technologies that support fact checkers and verification experts in their day-to-day work, making it easier and quicker to detect, analyse and understand false, distorted, or misleading content items or narratives.

One of seven use cases in the AI4Media project deals with new AI functions for detecting and analysing disinformation. Apart from testing the new functions provided by technology partners in the project, it conducts research into the challenges and needs of professional users in the European fact checking and verification sector.

This White Paper is intended for AI researchers and technology developers with an interest in providing new research and functions related to counteracting disinformation. It summarises the results of the use case work, describes challenges and end user requirements as well as responses from a survey conducted with European fact checking and verification experts. The paper details the needs and requirements for:

1. Detection of synthetic media items or synthetic elements, and identification of content manipulation,
2. Detection of disinformation narratives in online/social media, including respective content, actors, or networks and
3. AI support functions that are trustworthy and transparent for non-technical users.

AI technologies already play a role today in supporting fact checkers and verification specialists. Although shortcomings were mentioned by some respondents, results from the survey with fact checkers and verification experts show that AI-powered support is much needed and highly valued for the task of fact checking and verification. More than two thirds of respondents had either a high or moderate need for support to help them detect whether a specific media item has been synthetically generated or synthetically manipulated – with high rates of importance applied to all content types, especially Photos, Video and Text. The same high level of need was expressed by over two thirds of respondents for support with detecting and understanding disinformation narratives that occur in online channels, and especially with identifying related actors/networks. Almost all respondents said that they have a high or moderate need for AI functions that have implemented specific trustworthy AI features. Over two thirds stated that they have a high level of need for such features and explainability/transparency was the most important dimension of Trustworthy AI.

Key messages

- Most fact checking and verification specialists regard AI technologies as highly valuable and important to support them in the task of counteracting disinformation, despite shortcomings associated with some existing tools.

- New AI support functions are needed in two main areas of fact checking and verification work:

  1. Detection of synthetic media items or synthetic elements, and identification of content manipulation,

  2. Detection of disinformation narratives in online/ social media, including respective content, actors, or networks.

- The user group of fact checkers and verification specialists has a high need for trustworthy, understandable AI support functions, especially in terms of explainability, transparency, and robustness.

Contents

1

Introduction

Background: Counteracting disinformation

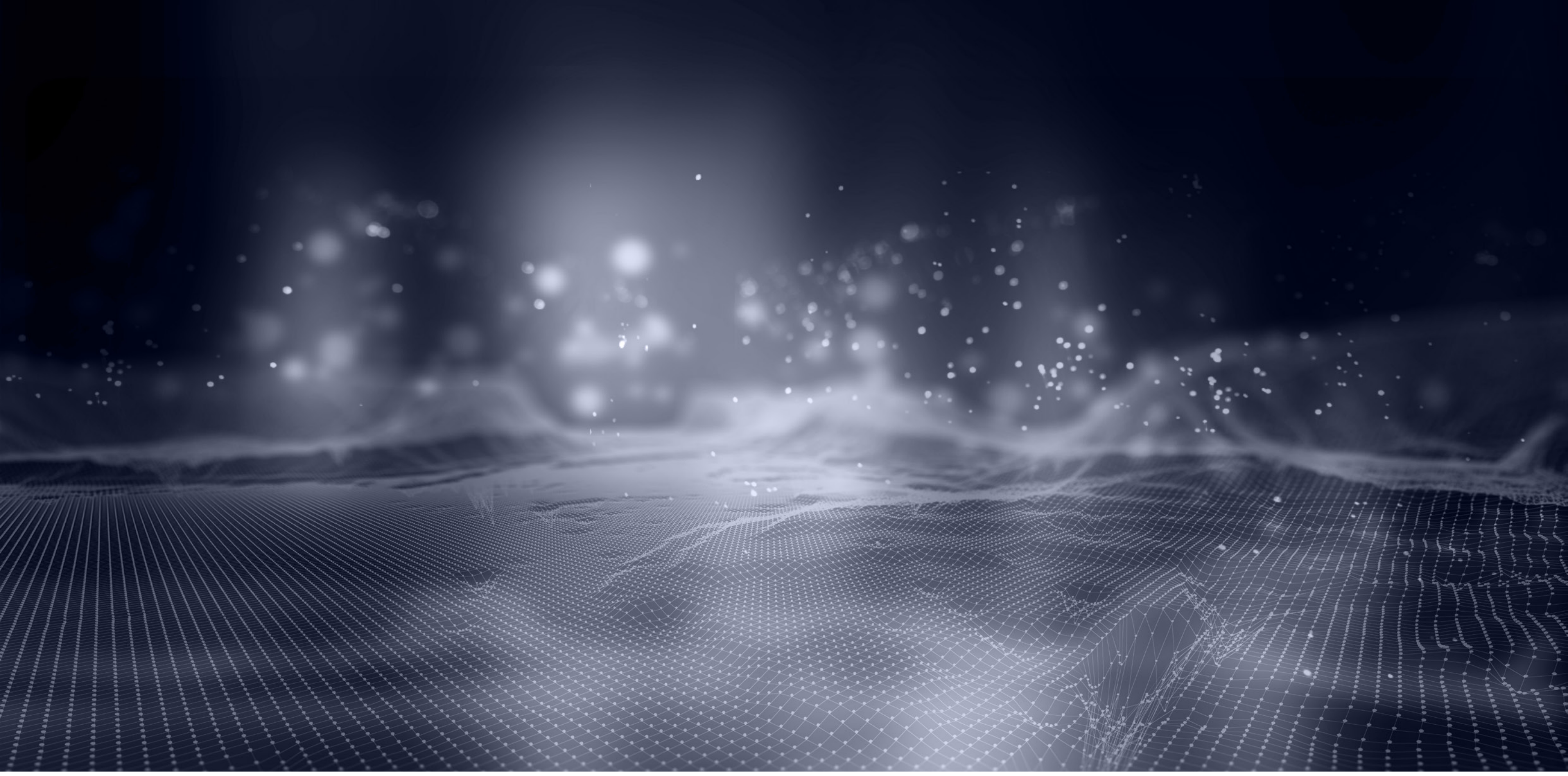
The phenomenon of online disinformation has evolved since around 2010 and refers to false, inaccurate, or misleading information that has intentionally harmful objectives. While the spreading of false or manipulative information has occurred for centuries, the significance and negative impact of this activity has increased with the emergence of social media and digital platforms as well as advances in technology, including Artificial Intelligence (AI).

Although online disinformation has been addressed by verification specialists as well as journalists as part of their work for almost one decade, more recent events such as major elections, the Covid-19 pandemic, and international conflicts have brought the 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 (related) approaches come into use for the purpose of counteracting disinformation:

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

This White Paper focuses on the last of the six points above: counteractive technologies and support tools. This topic is also the focus of one of seven use cases in AI4Media.



Innovation development and research in the AI4Media project

AI4Media is an EU co-funded research initiative with 30 technology and media partners for diverse aspects of AI in the media sector, advanced AI technology research, and development of specific solutions for seven use cases. One of these cases focuses on counteracting disinformation, run by ATC and DW.

The use case defines requirements, deploys, and tests new AI technologies to improve tools used by fact checking and verification experts for disinformation detection/understanding. These new AI functions are provided by technology partners in AI4Media. During the project, they are made available in a prototype demonstrator related to existing support tools: Truly Media (a web-based platform for collaborative verification) and TruthNest (a Twitter analytics and bot detection tool).

Work in the use case covers two main topics, synthetic media detection and identification of narratives related to disinformation, for which a detailed list of requirements from fact checking and verification specialists was developed. Another aspect is the exploration of Trustworthy AI in relation to AI-powered functions that are deployed within media tools.

In the first half of 2022, DW and ATC conducted a survey with 19 European fact-checking and verification practitioners to further explore AI support needs from this European expert community. The large majority of participants in this survey (79%) work for either fact checking, or news media organisations and their job roles directly relate to fact checking and verification tasks. The remaining participants were disinformation experts with a high degree of understanding of the workflows and tasks involved. The majority of respondents (68%) works in a specialised team, which is dedicated to fact checking and verification. The respondents have been selected to reflect the currently specialised, complex nature of this work. It can be expected that this fact checking and verification user group will expand with the increasing availability of suitable, user-centred, and easy-to-use technology support tools.

2

What are the issues and challenges?

Use of AI technologies to counteract disinformation

In addition to manual/human analysis techniques, various technologies play an important role to support fact checking and verification specialists in their work to counteract disinformation. This ranges from basic assistance such as image enlargement or frame-by-frame viewing to several AI solutions that are in use today, such as reverse image search. The survey asked respondents for which fact checking and verification tasks they are currently using AI-powered support technologies. Key application areas mentioned are listed below.

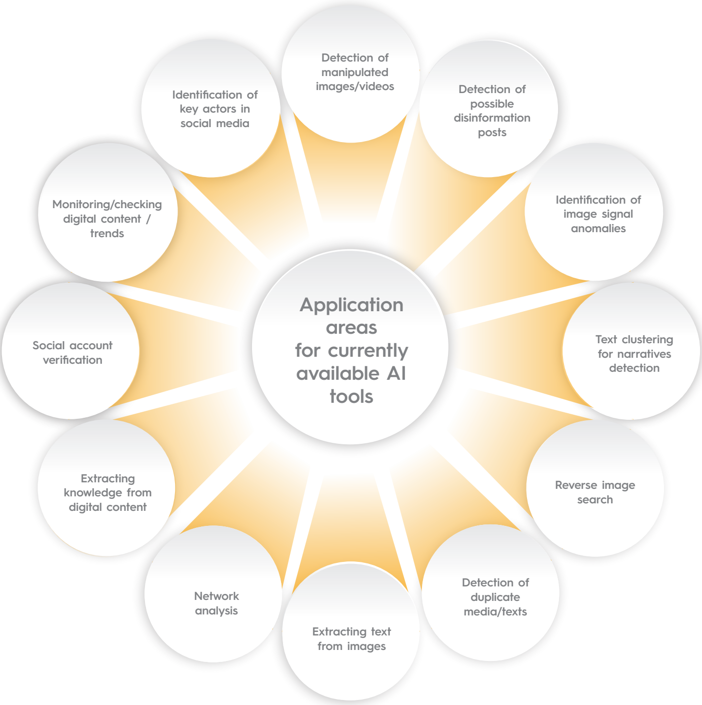

Application areas for currently available AI tools
Detection of manipulated images/videos
Detection of possible disinformation posts
Identification of image signal anomalies
Text clustering for narratives detection
Reverse image search
Detection of duplicate media/texts
Extracting text from images
Network analysis
Extracting knowledge from digital content
Social account verification
Monitoring/checking digital content / trends
Identification of key actors in social media

Figure 1: Counteracting disinformation: application areas for currently available AI tools

Respondents also mentioned several shortcomings and limitations with currently available AI tools. Disinformation analysis tools are often limited to Twitter, and there are technology-related doubts over the accuracy of results, the underlying data sets and AI analysis criteria used. It is felt that certain AI support technologies are not yet mature enough to be relied upon in a practical media/content workflow, as shown by this statement: "I know an AI system that chooses posts in social media that may be disinformation related and presents these for human evaluation, but 85% of these suggestions are not eligible for fact-checking, they are AI noise." Another point for criticism was that many tools are not open access (fee-based) and that some of them are too complex for day-to-day use, due to a lack of user-centred interfaces. At this point in time, there are situations, where AI-powered support may not be necessary or helpful, as noted by this respondent: „The most problematic content that we detect as fact checkers is not very elaborate. We hardly find deepfakes, but we do see many 'cheap fakes'. These manipulations are not very elaborate, so it is not essential to use AI tools to detect them."

As shown above, AI technologies already play a role today in supporting fact checkers and verification specialists, despite existing limitations. Although shortcomings were mentioned by some respondents, results from the survey show that AI-powered support is much needed and highly valued for the task of fact checking and verification. When asked to give their opinion on statements regarding the role of AI in these workflows, the large majority (68%) "strongly agree" or "agree" that AI technologies are an important element today and nearly half of them (48%) agreed that without the help of AI-powered tools they could not achieve many important outcomes. 53% are aware that AI is involved in the functions of their tool. Only a minority (16%) feels that AI is overrated in its importance for this kind of work.

What is your opinion on the following statements, related to *the role of AI in fact-checking and verification workflows?*

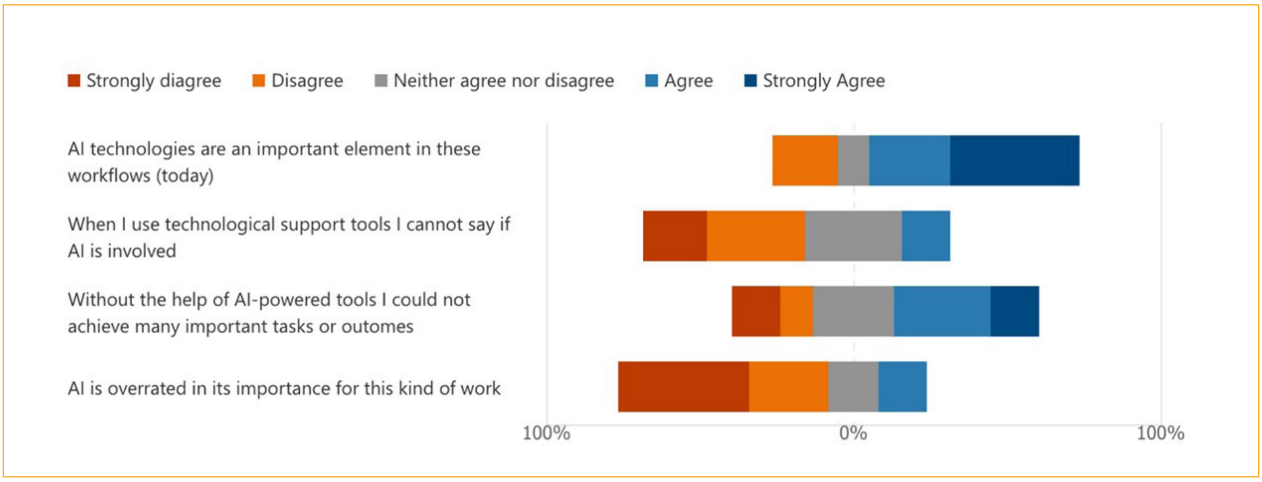


Figure 2: Opinions on the role of AI in fact checking and verification workflows

The need for technological support has recently increased and will be increasing in the future: On one hand, the frequency and scope of disinformation have grown to a level that makes many tasks difficult to handle with only manual approaches. On the other hand, more advanced AI technologies and automation approaches will be used for targeted disinformation narratives, content manipulation or synthetic media production. High-end produced deepfake videos or photos that show people's faces may in many cases not be detectable by humans without in-depth analysis. It is also expected that subtler occurrences of disinformation will occur, especially regarding the involvement of synthetic media. For example, rather than "cheap fakes" and isolated deepfakes seen today, synthetic media could be combined with traditionally produced information in a subtle way, making it more difficult to detect.

User requirements and challenges in relation to new AI functions that support fact checking and content verification tasks are broadly related to four areas, which are described in the following chapters.

1. **Detection of synthetic media items** or synthetic elements, and the identification of **content manipulation.**
2. **Detection of disinformation narratives** in online/social media, including respective **content, actors, or networks.**
3. **AI support functions** that are trustworthy and transparent for non-technical users in this community.
4. **Contextual aspects,** such as user interfaces, workflow integration, and human-AI collaboration.

Detection of synthetic media and content manipulation

While there are many editorial or entertaining applications for synthetic media, it is also being used in the context of disinformation.

The term “synthetic media” refers to artificially generated media where AI technologies are responsible for one or all parts of the production (content generation) task. Media items can be fully synthetic (AI-generated) or contain some synthetic elements. High-end synthetic media items that are used for disinformation purposes are colloquially known as “deepfakes”.

All types of digital media content can be synthetically generated or manipulated: Text, Audio, Spoken Word, Photos, Videos, and Images. Unless synthetic media items are clearly labelled, they increase the volume of distrusted content in social and digital media. It is the task of fact checkers and verification experts to identify, understand and explain the use of synthetic media in the context of disinformation, ranging from deepfake videos/images to (unlabelled) machine-generated text or audio and manipulated photos. In the context of fast-paced publishing workflows and the need to quickly “debunk“ to mitigate the impact of disinformation, this task must be achieved in a short timeframe, and yet with a high degree of accuracy/certainty.

There is much at stake, because media and dedicated fact-checking organisations cannot afford to make mistakes, e.g., declaring a real video as a deepfake, or vice versa.

Detection of disinformation narratives and understanding of content, actors, or networks

This journalistic process is closely related but goes beyond the verification of single content items.

This journalistic process is closely related but goes beyond the verification of single content items. Fact checkers and verification specialists also need to detect and understand strategic disinformation narratives in online/ social media or reports that are based on false/distorted information. Such disinformation related communication patterns can emerge dynamically or persist over longer periods of time. They often spread to other platforms and languages. It is the task of verification specialists to analyse and fact check entire information patterns, single statements from public figures (claims) or topical narratives, including the content, actors, accounts, and networks/communities involved. Further, it is helpful to understand the origin of such narratives and the direction of information spreading. The key challenge here is the sheer volume of online/social information that needs to be covered and analysed, including multiple platforms and languages. Again, results must be delivered in a short period of time, for example during rapidly developing breaking news events.

Trustworthy AI functions and transparency for non-technical users

AI-powered support tools for counteracting disinformation are largely used by specialist staff for fact checking and verification, who are often trained journalists, investigative researchers and may have training in data-related issues.

By the nature of their role, they tend to have curious minds and may be also governed by editorial, media, fact checking and AI guidelines as well as specific codes of practice. For this reason and to trust/use specific AI-powered predictions, they need to better understand the AI support functions presented to them. Current tools lack general transparency (e.g., related to AI methods used or legal compliance) as well as specific trustworthy AI features (e.g., related to explainability, fairness or robustness). None of the respondents in the survey had come across AI-powered support tools fact checking and verification that had dedicated trustworthy or transparent AI features. This presents an issue regarding the acceptance of these tools, also by managers who are responsible for their implementation in the context of corporate AI guidelines.

Contextual aspects: User interfaces, workflows, and human-AI collaboration

While disinformation activities and the volume of related verification work increase, the workflow for verifying digital content and detecting disinformation remains very complex, time-consuming and specialist.

New AI support functions will be limited to the use by dedicated verification specialists/ experts, unless they have suitable interfaces, that are easy-to-use and can be easily integrated into existing workflows. There are also issues with human-AI collaboration, i.e., the high level of human intervention and control required at present. Although the survey focused on AI support functions, many respondents mentioned general contextual issues: the lack of staff for disinformation monitoring, limited automation with human still needing to interpret/judge AI results, the complexity of tools and difficulty to read AI results, a lack of APIs that can be easily integrated and too many tools for each task.

3

What are the needs?

Key areas for AI technology support

The previous chapters have shown that despite the availability of various AI-powered support functions, there are several shortcomings, limitations, and missing elements to ensure long-term success in counteracting disinformation. In addition, there is a need for easy-to-use tools with trusted AI functions, so that a wider group of non-specialist users and other researchers can be involved in content verification and disinformation detection tasks. To ensure acceptance of AI-based tools and their implementation, users need to be able to judge aspects of trustworthiness regarding machine-generated results and predictions.

In summary, there are three key areas where further AI technology support is needed:

1. Detection of synthetic and manipulated media,
2. Analysis of disinformation narratives, actors, and networks and
3. Capability for Trustworthy AI by design.

In addition, there are general needs related to associated user interfaces, workflow integration, and human-AI collaboration.

The following chapters describe the specific user needs, pain points, and requirements that have been collected for the three areas listed above among fact checking and verification experts.

AI support needs for detecting synthetic media and content manipulation

Our survey shows that there is a high need for AI support for the task of detecting synthetic media items and manipulated content. This task is illustrated by the following user story:

> *"As a fact checking or verification expert, I want to detect whether a text, image, audio or video might be synthetically generated or manipulated so that I have additional information for my manual verification process and that I can prove and explain specific disinformation activities".*

84% of the respondents had a "high" or "moderate" need for AI technology support to detect whether a specific media item has been completely synthetically generated. Broadly the same applies to the AI support requirement for detecting whether synthetic content elements have been used to manipulate an otherwise traditional media item (89%). In both cases, over half of the respondents said that they have a "high" need: 57% for synthetic item detection and 63% for manipulation detection.

Key pain points mentioned by users are the difficulty (and time needed) to identify unlabelled synthetic media items, including the concern of not being able to recognise this type of synthetically generated content at all. In the context of synthetic media, there is also a need for support regarding the detection of changes and manipulations that have been applied to media items. For example, post-edited synthetic media, where the remaining flaws have been erased after image generation. AI support is especially needed for detecting synthetic or manipulated media items that occur within a large volume of disinformation related content, rather than in isolation as some deefake examples. When a synthetic media item has been detected, users also appreciate supporting information on where and how fast it has spread in social media. Generally, users point out that these types of detection need to be done within a short time frame. They also like solutions that allow for quick matching of disinformation related content items that have already been debunked by other fact-checkers. Although there are AI-powered solutions available, some users are not satisfied with the accuracy of the results.

Regarding AI technology support for detecting synthetic media items or synthetic Manipulation - please rate

how important this support is for each type of media

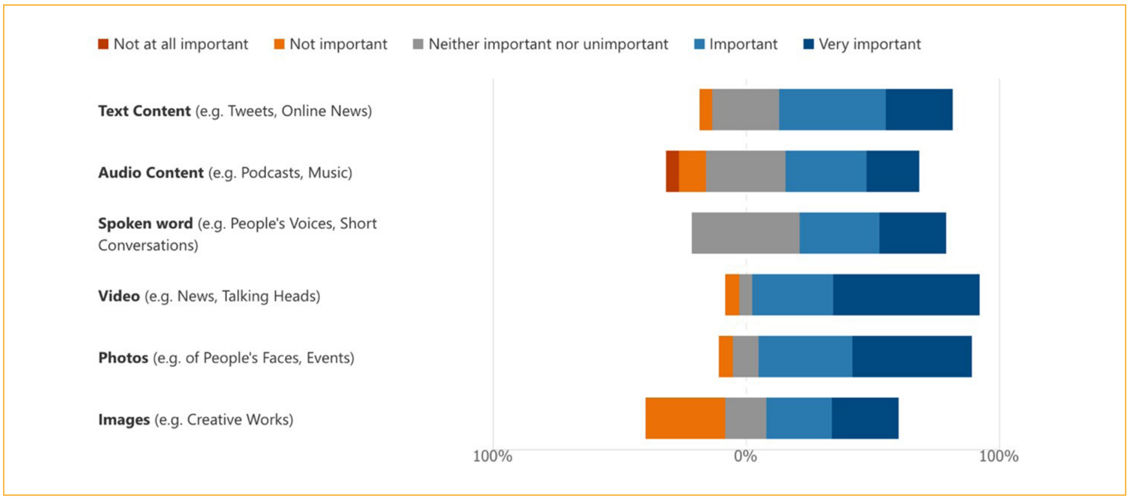


Figure 3: How much is AI technology support needed for each media type

The diagram on the previous page shows results from our user survey regarding the importance of AI support for each media type. AI support is most needed for Video and Photo content (58% rate this as "very important" for Video and 47% for Photos). Generally, there is a significant level of importance associated to all content types, when considering answers that rated AI support as "very important" or "important": Photos (84%), Video (68%), Text (68%), Spoken Word (58%), Audio (53%) and Images (52%).

For synthetic **text content,** the main requirement from fact-checkers is to get support with identifying whether a text has been fully or partially AI-generated (if not labelled, which is usually done by media organisations that publish automatically generated content/services). This applies to online news content as well as social media items, e.g., Tweets. In addition, there is a need to detect subtle text manipulations, using synthetic media techniques. 68% of respondents regard AI support for detecting synthetic text items or synthetic manipulation as "very important" or "important".

Regarding synthetic **image content,** users require AI support to detect synthetically generated portraits, which refers to photos of people in online information or portraits used as profile/avatar pictures in social media accounts. Further, there is a need to detect specific changes that have been applied to an original photo. For example, the addition or deletion of elements and changes made in relation to light, background, season, skin colour, facial expression, glasses, age or pose. 84% regard AI support for detecting synthetic photos or synthetic photo manipulation as "very important" or "important". For images other than photos (e.g., illustrations) the level of importance is lower (52%).

When it comes to synthetic **audio content**, it is again important to have support with detecting synthetically generated audios or synthetic elements and their possible duplication/location in an audio file. Ideally, this detection would be possible in (near) real-time and cover background noises. Regarding natural voices of people (e.g., speaking in videos, at events or in interviews), there is also a need to identify the authenticity of a person's voice (does the voice belong to the speaker or is a voice actor involved, wording, pronunciation and finding the original speech). 53% of respondents regard specific AI support for the purpose of detecting synthetic audio items or synthetic manipulation as "very important" or "important". The level of importance is at 58% broadly similar regarding Spoken Word.

For synthetic **video content,** the user needs are similar to those related to audio and image content. As previously, there is a requirement for support with detecting synthetically generated videos, or elements in a video (including addition, duplication, and deletion). Further, it is useful to detect the authenticity of a speaker as well as changes that might have been applied related to light, background or season, and lip synchronisation matches. Regarding AI solutions, special attention should be paid to low-compression video files on social networks, which currently present difficulties for synthetic media detection. The same applies to efforts aimed at lowering the high rate of false positives found in existing deepfake detection systems. 68% regard AI support for detecting synthetic video items or synthetic manipulation as "very important" or "important".

AI support needs for detecting disinformation narratives

As for the task of synthetic media detection, there is also a high need for AI support regarding the detection of disinformation narratives, including the content, actors or networks involved. This task is illustrated by the following user story:

> ***"As a fact checking and verification expert, I want to analyse disinformation narratives and stories, networks and accounts in digital media related to specific keywords (e.g., Covid-19), so that I can better understand the context, dynamics, and causal relationships within these narratives."***

89% of the respondents had a "high" or "moderate" need for AI technology support for the general task of detecting and understanding disinformation narratives that occur in online information. The same high level of need has also been stated for the specific task to detect actors and networks behind such narratives (89% state a "high" or "moderate" need). Almost two thirds of respondents (68%) said that they have a "high" need for support when it comes to detecting and understanding the actors and networks involved in disinformation narratives.

Specific pain points mentioned by users are the difficulty in recognising strategic or rapidly emerging disinformation narratives and the time taken to identify and analyse associated user networks or accounts. Fact checking and verification experts generally need to get more aggregated insight from diverse data based on a keyword and to see various types of results, such as key actors, narratives, micro-communities, conversations of communities as well as trending posts or hashtags. This is considered more difficult when disinformation narratives are very subtle, occur in multiple languages or when they are mixed with general factual news and information. It is considered time consuming to analyse and distinguish misinformation from disinformation (answering the question whether false/misleading content identified was created purposefully). To avoid doubling up their efforts, users are interested in alert systems that can match new incoming content with content that has already been debunked, which would speed up and organise their processes. Users also point out that there is a lack of AI-powered support tools to detect disinformation in multiple languages, beyond

English. There is also a need for the automatic translation of keywords used and to automatically have an archived version of the disinformation source that is being fact-checked.

Regarding the analysis of **specific social media actors and networks**, a key requirement is to get support with the detection of high-impact influencer accounts/networks as well as understanding their characteristics and development over time. This includes insight into relationships between networks/accounts and their respective target groups. It is also important to identify new or emerging accounts/actors versus existing ones. Further, users are interested to quickly identify the content platforms that spread certain narratives, where a narrative originated (e.g., tracking the path of a message back to its origin) and in which direction it is spreading (e.g., to which platforms/languages and on which platforms the narrative is currently shared). Generally, in the context of disinformation analysis, it is desirable to apply AI-powered natural language analysis solutions from one country/language to another.

As important is the need to get support with a keyword-based analysis of the **content** distributed in social media, including an analysis of content items, context characteristics and recognising similar content/narratives. This includes the detection of multimedia elements and how large their proportion is in comparison to text content. It is also helpful to get support with identifying the combination of certain words and similar expressions. This task refers to content found in longer text forms, such as news articles or comments, but also in short text forms like Tweets or social posts, where the latter is considered more difficult. An analysis of crowdsourcing comments is another area where AI support is considered as useful for differentiating verifiable statements from opinions. Users would find it helpful to receive topic alerts for content narratives/items, especially those that are emerging and not yet well understood. For the analysis of specific accounts, it is key to get support with the identification of Bot-Accounts as well as information related to geographic location, nature of speech or sentiment.

The diagram below shows results from our user survey regarding the importance of AI support for each analysis aspect of disinformation narratives. AI support is most needed for the task of identifying Actors and Networks (68% rate this as "very important" for Actors and 63% for Networks). Generally, there is a significant level of importance associated to all these analysis aspects, when combining answers that rated AI support as "very important" or "important": Actors (89%), Networks (89%), Topics (78%), Accounts (74%), Digital Platforms (69%), Content Elements (68%) and Languages (42%).

Regarding AI technology support for detecting and understanding disinformation narratives - please rate how important this support is for each analysis aspect of disinformation narratives.

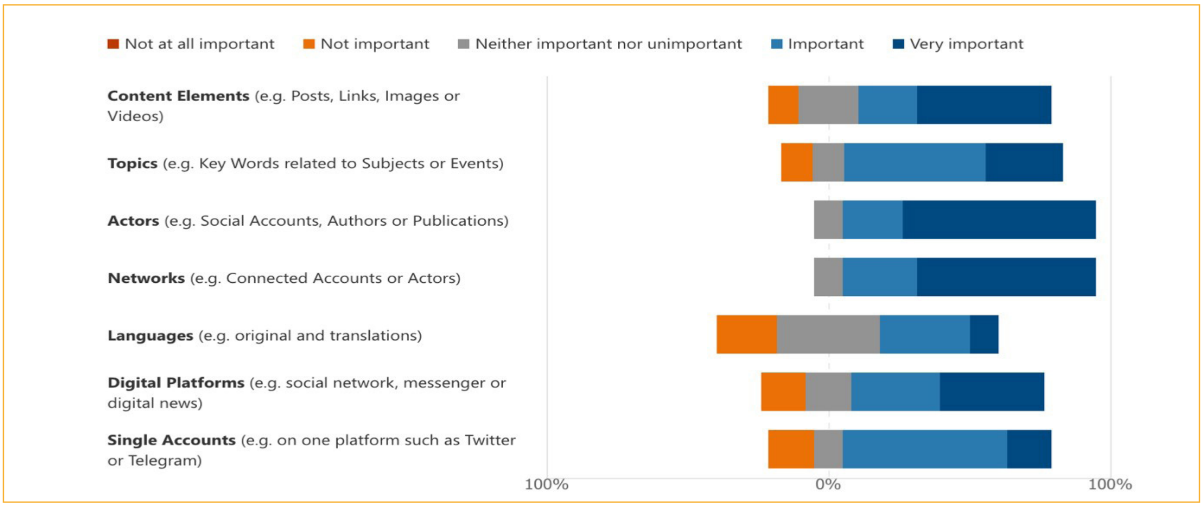
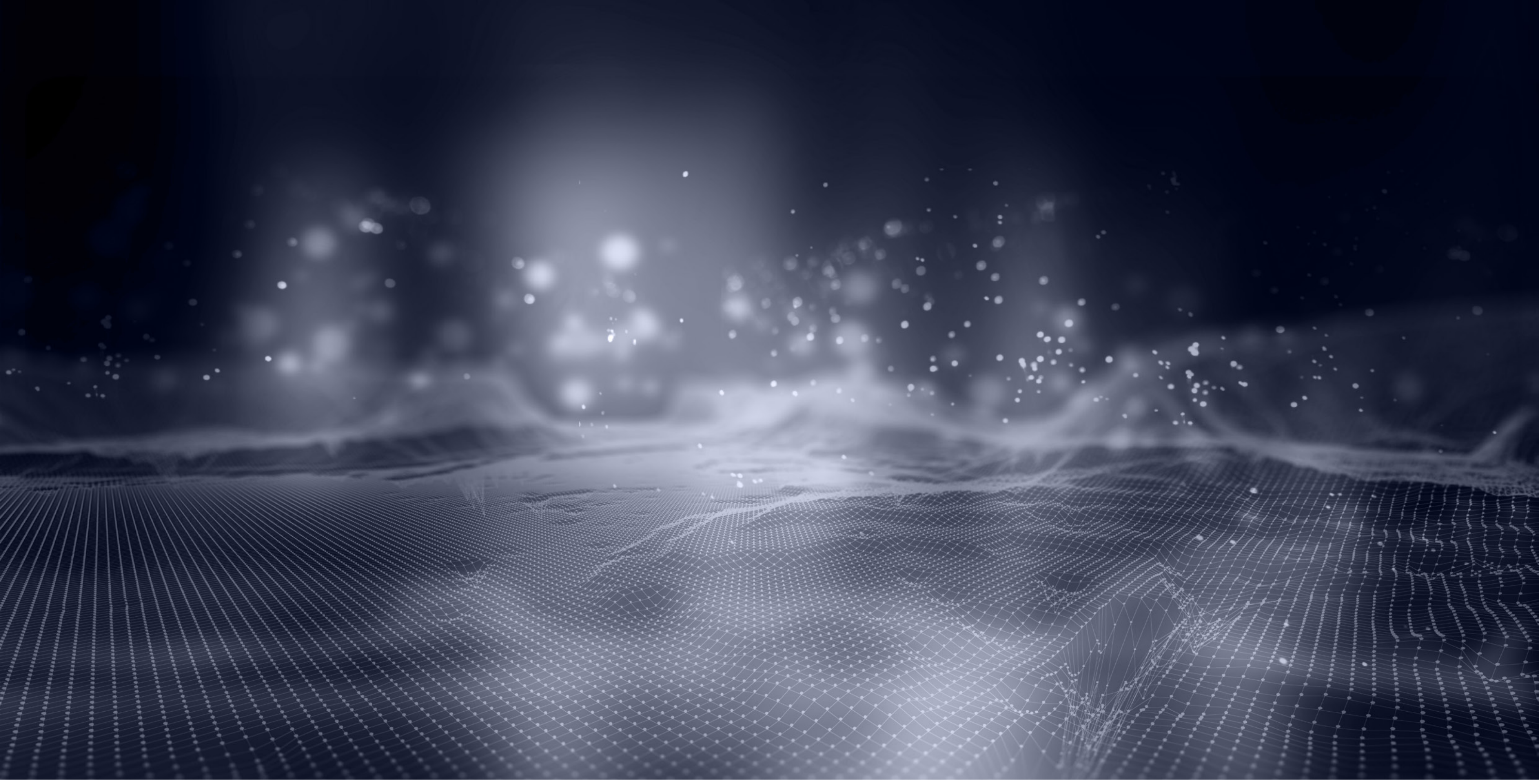


Figure 4: How much is AI technology support needed for each analysis aspect



Capability for Trustworthy AI

In addition to the functional AI support described in the previous chapters, there is a significant level of need from fact checking and verification experts for aspects of Trustworthy AI, such as Explainability, Transparency, Robustness, Fairness or Legal Compliance. The capability for Trustworthy AI is illustrated by the following user story:

"As a fact checking and verification specialist, I want that AI services provided in my tool have addressed trustworthy AI principles "by design" and provide for me understandable information, so that I can trust the results/predictions delivered and use these services without editorial, ethical, legal or security concerns."

90% of the respondents had a "high" or "moderate" need for AI functions that have implemented specific trustworthy AI features. The large majority of this group expressed a "high" need for this (79%). Most respondents (79%) said that they have either "come across" or were "quite familiar" with the concept of Trustworthy AI.

Fact checking and verification experts point out that there is a general lack of insight and understanding regarding the results delivered by AI functions or the AI methods and data sets used. This is also an issue for managers who need to implement tools with AI-functions, as they have to comply with AI legislation as well as internal corporate AI guidelines. The needs from end users and managers are related to information about legal compliance, the data sets used, aspects of bias mitigation and fairness, the level of robustness, how and why algorithms reach specific predictions (explainability) and overall transparency with a view to enable traceability and auditability. It is particularly important that Trustworthy AI features provide their outcomes/ information in a language and format that non-technical end users and managers can understand and are willing to read. Only then it becomes usable by non AI-experts and can also be incorporated into publications and storytelling for transparency, which builds trust in journalistic work.

Another topic of interest in this context is the provision of information related to sustainability, such as the level of energy used by an AI system ("Green AI"). The clear majority of respondents to

Regarding the different Trustworthy AI features that can be implemented for an AI function how important do you rate each one in the context of fact-checking and verification tasks?

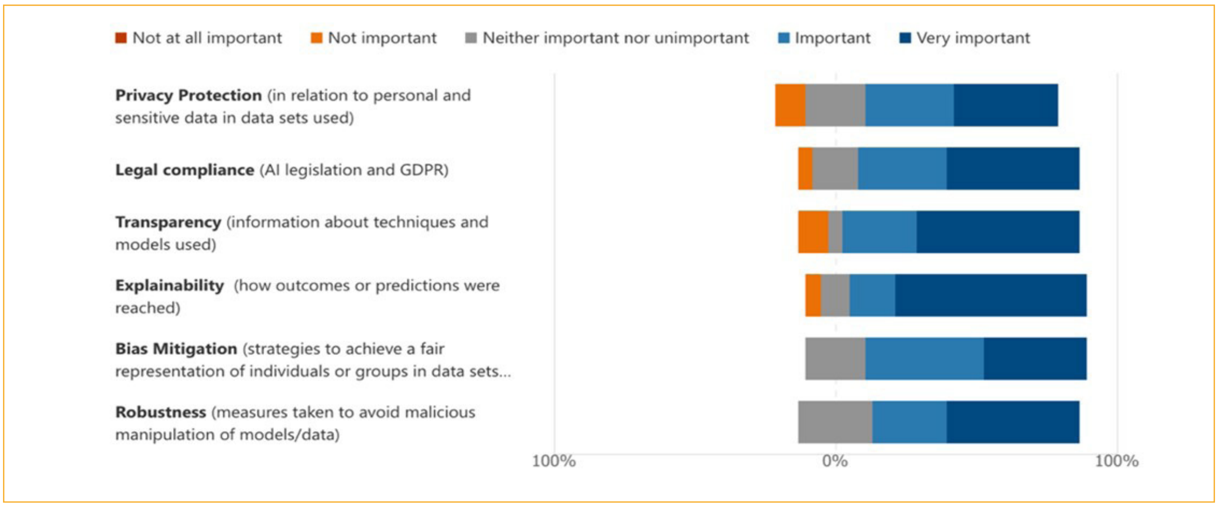


Figure 5: Importance of Trustworthy AI features

our survey (84%) felt that energy efficiency should be considered, but that the accuracy of results is more important if there is a trade-off. Only a minority (10%) thought that energy efficiency should always be a priority, in any case.

Regarding the different dimensions of Trustworthy AI, the lack of transparency/explainability is mentioned by fact checking and verification experts most frequently. It is pointed out that AI functions should generate explanations and provide information about the criteria used by a disinformation detector to reach a decision ("We should avoid the black box effect as much as possible as otherwise the detector will not be trusted"). Examples mentioned were sets of criteria used by AI systems to flag a potential deepfake or disinformation related content. A respondent explained that some users don't trust criteria used by AI systems and would therefore accept results only as an "alert" for further manual investigation. In this context, it was pointed out that criteria used to detect disinformation are widely published, which may allow creators of disinformation to take them into account to avoid detection.

The most needed Trustworthy AI features are Explainability (68%), Transparency (58%) and Robustness (47%), where higher numbers of respondents said that these features are "very important". Generally, there is a significant level of importance associated to all these Trustworthy AI aspects, when combining answers for each feature with ratings of "very important" or "important": Transparency (84%), Explainability (84%), Legal Compliance (79%), Bias Mitigation (79%), Robustness (73%) and Privacy Protection (69%).

4

Conclusion

This White Paper described challenges in the area of counteracting disinformation and summarised AI support needs and opinions, collected from fact checking and verification specialists.

This information was designed to help AI researchers and technology developers to define relevant research topics and provide technical AI support in the key areas required (detection of synthetic media, content manipulation and, identification of disinformation narratives, actors, or networks). It was shown that users regard support from AI technologies as highly valuable and important, including a high need for trustworthy AI, especially with regard to explainability, transparency, and robustness.

AI support is likely to be even more welcome by this user group in coming years, as pointed out by a one of the respondents: “Any improvements to the AI’s ability to detect problematic content will help. It should also be noted that disinformation produced massively by artificial intelligence is not a big problem today, but it may be in the near future, so AI tools designed to detect it may be essential in a very short time“.

The research with users has also shown that AI researchers and technology providers face some steep challenges to match user expectations across multiple dimensions, e.g., in terms of accuracy, trust and the way tools are provided in the market. This is reflected by the following statement from a respondent, who discusses AI support in the context of synthetic media detection: “For the detection of deepfakes, AI tools are currently rarely helpful (low hit rate, not transparent and fee-based). For the foreseeable future, logical thinking and human intuition/tracking will be much more effective and better than tools at detecting deepfakes.“

Although the user needs and areas can be described, it may not be possible to develop solutions for all needs. This is due to possible limitations regarding datasets, data languages or other constraints. One respondent described this as follows: “The language in which disinformation is disseminated is relevant. Even in English, there is a limited volume of fact checks to feed a certain AI tool that has been developed. In my view, it is a misconception to hope for an imminent use of AI to support fact checkers in this task“.

The issues with some existing tools and the AI solutions currently missing show that there is much scope for future research, AI solution development and potential impact in the area of counteracting disinformation. The research has provided a clear message from fact checkers and verification specialists to AI technology providers that new and trustworthy AI-based solutions are highly welcome and needed.

AI4media

Follow us @ai4mediaproject

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

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.

info@ai4media.eu www.ai4media.eu