DisAI.eu

D1.3
Scientific webinar series report

Project Title	Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies
Contract Nº.	101079164
Type of Action	HORIZON-CSA
Topic	Disinformation Combating
Project start date	1st Dec 2022
Duration	36 months

Funded by
the European Union

Deliverable title	Scientific webinar series report
Deliverable number	D1.3
Deliverable version	1.0
Contractual date of delivery	30 November 2024
Actual date of delivery	29 November 2024
Nature of deliverable	Report
Dissemination level	Public
Work Package	WP1
Task(s)	T1.2
Partner responsible	CERTH
Author(s)	George Karantaidis (CERTH), Symeon Papadopoulos (CERTH), Nikos Sarris (CERTH), Simon Ostermann (DFKI), Qiwei Peng (UCPH)

Abstract	Within T1.2, we organized a series of four scientific webinars aimed at enhancing research excellence, particularly for KInIT researchers. These webinars cover key topics relevant to the project, including combating disinformation, multilingual AI, trustworthy AI, and multimodal AI. Through expert lectures followed by discussions and knowledge sharing, the webinars provide insights into cutting-edge research, state-of-the-art methodologies, and open challenges in these areas. All webinars have been recorded and are available on the project website.

Keywords	Scientific webinars, disinformation, multilingual AI, trustworthy AI, multimodal AI

Table of Contents

1 Introduction

This deliverable reports on the organisation and execution of the scientific webinar series conducted within the DisAI project under Task 1.2. The primary objective of this series is to enhance the research excellence of KInIT through a structured program of expert-led webinars on key topics relevant to the project's scope. Coordinated by CERTH with support from DFKI and UCPH, the series addresses four major themes: combating disinformation, multimodal AI, multilingual AI, and trustworthy AI.

Four webinars were conducted from M3 to M24, each led by experts who presented the latest research, methodologies, and outcomes from various EU initiatives. These presentations offer a comprehensive overview of the cutting-edge research and methodologies discussed throughout the series. Each session covers state-of-the-art methods relevant to the DisAI project's focus areas, including advancements in disinformation detection, AI-driven multimodal data analysis, multilingual AI models, and trustworthy AI systems. The content also incorporates insights from recent EU-funded research and innovation projects on disinformation, demonstrating real-world applications of AI for combating disinformation.

After each webinar, participants actively engaged in discussions by asking questions directly to the speakers. This interactive format enabled participants to clarify concepts, deepen their understanding of the topics presented, and fostered a dynamic learning environment that enhanced their overall comprehension of the material.

Targeted primarily at senior PhD students, postdoctoral researchers, and junior researchers, these webinars aimed to deliver in-depth, research-oriented content. The focus was on deepening knowledge in emerging areas such as disinformation detection, AI model explainability, multimodal and multilingual AI. By providing a platform for expert presentations followed by interactive discussions, the webinar series contributed to advancing the research capacities of the participants, while also aligning with the broader goals of the DisAI project.

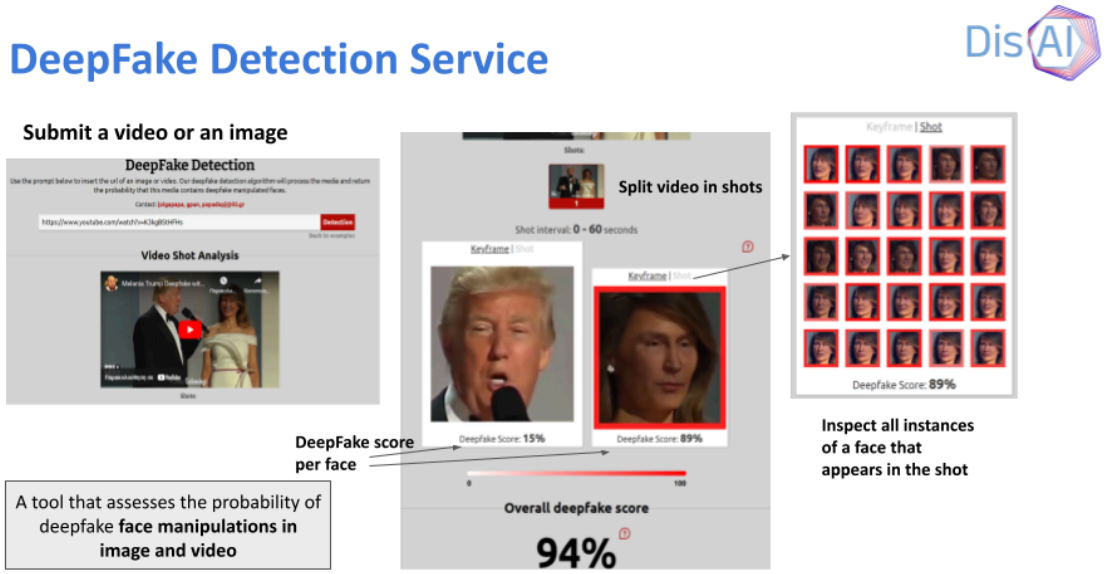
Presentations, recordings, and participants lists of the webinars are available as supplementary material in the Appendix.

2 Scientific Webinar Series

2.1 Webinar #1: Disinformation combatting: Challenges, methods and tools

The fist webinar in the DisAI series, hosted by CERTH and led by Principal Researcher Dr. Symeon Papadopoulos, focused on the topic of disinformation combating. This webinar offered a comprehensive look into the current landscape of disinformation research, focusing on its diverse means of spreading and the growing threat posed by disinformation. The session introduced participants to a range of cutting-edge research, methodologies, datasets, and tools for combatting disinformation in various forms, from text-based disinformation to multimodal content. The webinar also shared valuable lessons learned from CERTH's Media Analysis, Verification, and Retrieval (MeVer) research group, offering practical tools to tackle this global issue.

The session began with an overview of the disinformation crisis, highlighting its widespread presence across platforms and formats, including text, visuals, and multimodal forms. Disinformation can mislead the public, damage reputations, manipulate elections, and harm public health by spreading dangerous misinformation. Key elements, such as fake news, misleading narratives, and out-of-context information, were discussed for their impact on public perception. The webinar examined rumour propagation, text-based fake news detection, and automated fact-checking, noting challenges in detecting multilingual and cross-platform content. It stressed the need for real-time learning systems and better datasets, as well as the gap between academic research and the tools used by journalists and fact-checkers, calling for more user-friendly interfaces. The focus then shifted to visual disinformation, particularly image forgery detection, covering forensic methods and fusion techniques for accuracy. DeepFakes, AI-generated false images and videos, were explored, along with detection methods like neural networks and physiological feature analysis. The DeepFake detection service developed by CERTH is depicted in Figure 1. Despite advancements, challenges in robustness and generalisation persist. Finally, multimodal disinformation, combining text and images to create deceptive narratives, was addressed, showcasing how joint analysis can improve detection but faces hurdles like out-of-context and biassed content.



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).

Figure 1. Demonstration of the DeepFake Detection Tool Developed by the MeVer Group (Presentation Snapshot).

A significant portion of the discussion focused on collective responses from the research community. Media literacy emerged as a key solution, emphasising the need to educate the public on how to critically evaluate the information they encounter online. Transparency measures, such as disclosing the sponsors of online ads, were also recommended to increase accountability in the digital space. Legal and regulatory frameworks were seen as essential tools in holding both individuals and platforms accountable for spreading false information. Transparency in AI models and datasets was a key theme of the webinar, with several initiatives highlighted to improve accountability. Microsoft's "Datasheets for Datasets" framework promotes transparency by documenting data creation and limitations, while Google's "Model Cards" provide details on AI models' intended use, performance, and constraints. IBM's "AI FactSheets 360" further offers templates for enhancing transparency in AI systems. Ethical AI principles, including respect for human autonomy, prevention of harm, fairness, and explicability, were discussed, alongside seven key requirements for trustworthy AI: human oversight, technical robustness, privacy, transparency, diversity, societal wellbeing, and accountability. These principles are essential for creating reliable AI systems to combat disinformation.

The webinar concluded with a call for continued research and collaboration to tackle the evolving challenges posed by disinformation. While tools and methods for detecting disinformation are becoming more sophisticated, much work remains in addressing the psychological, ethical, and technical complexities of this issue. By integrating media literacy, transparency measures, and ethical AI principles, the fight against disinformation can be more effectively coordinated across academia, industry, and regulatory bodies.

MeVer tools for combating disinformation. During the webinar, the MeVer group's tools were highlighted as key resources in the fight against disinformation. These tools are designed to address the complex challenge of detecting and analysing disinformation, focusing on three critical aspects: content, context, and propagation. Drawing from their experience in AI for multimedia and social network analysis, MeVer tools target a range of visual disinformation forms, including manipulated images, out-of-context visuals, false graphs, and memes. Below is an overview of the tools discussed in the webinar, categorized by their specific functions:

- Content:
  - Image Verification Assistant: Detects tampered areas in images, identifying manipulation types like splicing, inpainting, copy-move, and cropping. It integrates 12 forensic algorithms and provides metadata inspection and reverse image search.
  - DeepFake Detection Service: Assesses the likelihood of deepfake face manipulations in images and videos.
  - Visual Location Estimation: Estimates the geographic location of an image based solely on visual cues, detecting when images are out of context.
- Context:
  - Near-Duplicate Detection: Identifies near-duplicate visual content, detecting similarities between images and videos across different media formats.
  - Context Aggregation and Analysis: Analyses the broader context surrounding images or videos to understand how they fit within larger media narratives.
- Propagation:
  - Network Analysis and Visualization: Offers insights into how disinformation spreads through social media conversations, revealing interaction-based communities and the flow of false information

The MAAM (Media Asset Annotation & Management) platform was also discussed in the webinar as a powerful tool for organising and analysing media assets. It automatically generates short descriptions of images by identifying objects and actions within them, adding these as tags for easy retrieval. Users can manually adjust or add tags as needed, and the platform classifies assets into categories such as disturbing content, NSFW, or memes. This automation simplifies the process of managing large volumes of media, making it easier to track and assess potentially harmful or misleading content.

Q&As and lessons learned. Following Webinar #1, the presenter engaged with participants, addressing key challenges in disinformation combatting. The discussion began with the need to bridge the gap between academia and industry professionals through participatory design and collaboration with journalists and fact-checkers. This approach aims to enhance tool usability and identify areas for improvement. Questions about dataset availability for the DisAI project were answered affirmatively, confirming access to several datasets, particularly for deepfake detection and image forensics. The importance

of task-specific model evaluations was emphasized, noting that performance often hinges more on datasets than on model architecture. The conversation also tackled the challenge of cross-platform datasets, suggesting a focus on text snippet-level analysis for effective claim detection. Participants inquired about CERTH's visual location estimation tool, which combines classification and retrieval methods for improved accuracy. Interest in deepfake detection techniques was expressed, particularly regarding the need for multimodal and multilingual approaches. KInIT's ongoing dataset collection from various platforms was highlighted as a valuable resource. Finally, the potential for integrating explainability into deepfake detection models was discussed, acknowledging the challenges posed by the adversarial nature of the technology. The need for transparent identification of specific generative models was suggested to enhance analyst understanding. Overall, the discussion underscored collaborative opportunities within the DisAI project and the necessity for innovative strategies to combat disinformation effectively.

2.2 Webinar #2: Multilingual NLP: Overview and Low-Resource Multilingual Processing

The second scientific webinar on the topic of Multilingual NLP gave an Overview of the field and focused on Low-Resource Multilingual Processing. The webinar was hosted by the German Research Center for AI (DFKI) and led by Dr. Simon Ostermann (Senior Researcher at DFKI) and Cristina España-Bonet (Senior Researcher at DFKI).

In the first part of the talk, the presenters gave an overview of common approaches and techniques in multilingual and cross-lingual natural language processing, with a specific focus on multilingual representations and encodings, as well as non-societal biases in multilingual models. They started out with basics of multilingual language models, multilingual word, and sentence representations. An important distinction needs to be made between monolingual (one language), cross-lingual (language 1 as input and language 2 as output), and multilingual (any language in, any language out) models. Both multilingual and cross-lingual models are of high relevance especially for low resource languages, which are to date spoken by a majority of people world-wide. Multilingual models are often just trained by mixing diverse languages into the training data of large language models. Based on such multilingual models, sentence representation models can be trained, a technique that is especially relevant for the research in DisAI, which looks at text retrieval models for claim matching. Current approaches use contrastive fine-tuning objectives on parallel data, or distillation approaches, which aim to distil multilingual sentence representation capabilities from large models with multiple objectives Finally, the first talk touched briefly on current research topics, like reducing the so-called curse of multilinguality, and the question on when it is better to use a multilingual model or train a highly specialised monolingual one. An application example for grammaticality detection is depicted in Figure 2.

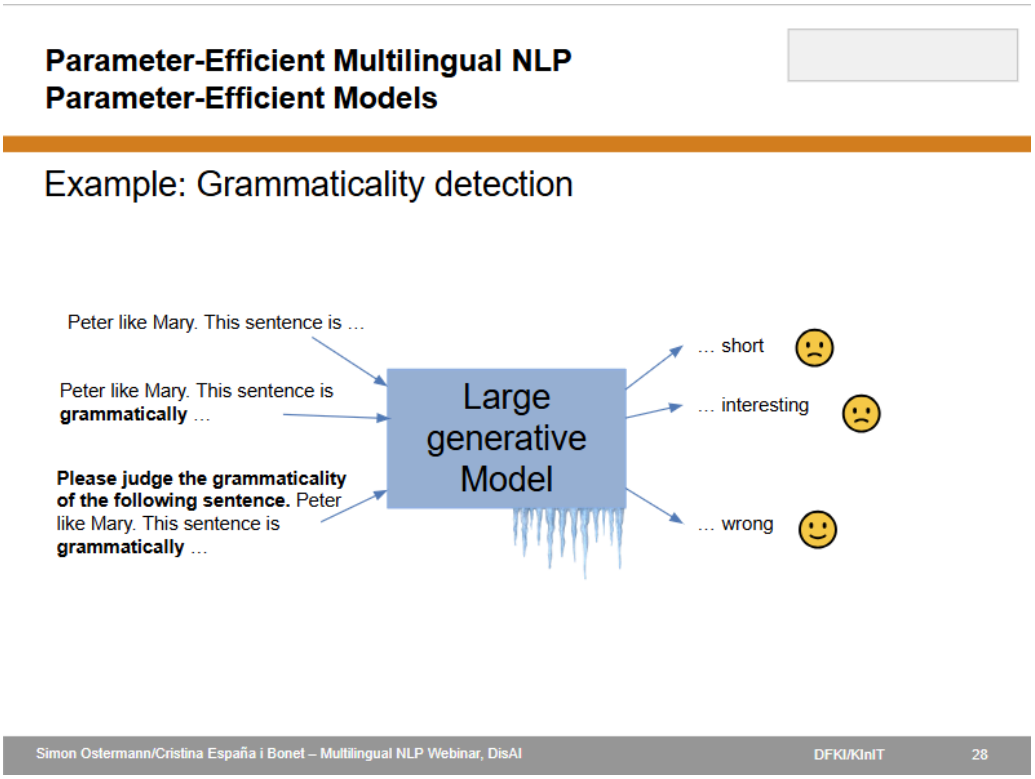


Figure 2. Example of AI model for grammaticality detection (Presentation snapshot).

Current research topics: Biases in multilingual models, self-supervised machine translation, and parameter-efficient techniques. The second part focused on low-resource multilingual processing and covered biases in multilingual models, and self-supervised machine translation as an example of data-efficient processing, as well as adapters and prompt tuning as examples for parameter-efficient techniques, both of which are important building blocks of low-resource NLP.

In the second part of the talk, the presenters gave an overview of current research at DFKI on biases in multilingual models. Biases of a social and non-social kind exist in multilingual models, as they just reflect the same biases that exist in their training data. The new CA-WEAT dataset that was introduced in this talk is the first attempt to measure cultural biases in language models that go beyond the Western/US American culture.

The third part of the talk focused on self-supervised machine translation (SNMT). SNMT is an extremely data-efficient training technique, in which parallel data extraction is used as an auxiliary task to enable NMT training, and NMT training is used as an auxiliary task to enhance parallel sentence extraction. Based on quasi-parallel data, a pair of models is trained to learn from each other and improve vice versa. The authors were able to show that using such training, no fully parallel data are required to get a good performing machine translation system, and the resulting systems are able to learn semantically meaningful representations.

The fourth and last part of the talk gave an overview of parameter-efficient learning techniques for low-resource multilingual NLP. After a quick summary on established methods for improving performance in low resource settings, e.g. through transfer

learning or data augmentation, the talk introduced the concept of Adapters, small sets of parameters that are inserted inside of transformer layers, which allow for a very efficient adaptation of large models to new languages. An alternative was presented in the form of soft prompts, additional sets of parameters added to the input, which offer an even more efficient but less powerful alternative to Adapters.

Q&As and lessons learned. After the webinar conclusion, the presenters Cristina España i Bonet and Simon Ostermann answered questions from the audience and participated in discussions, further analysing the specific issues raised by the participants. One question was raised on why fine-tuning the model is sometimes worse than other approaches and whether this signals that the fine-tuning is done incorrectly. The lecturers and participants discussed that fine-tuning a full model requires adjusting a large amount of parameters that the information is spread over. Therefore it is difficult for the model to determine the best information flow. Tuning soft prompts is less flexible because the model remains frozen so it is forced to concentrate on important parts of the data.

A participant suggested that strong regularisation could mimic that effect in fine-tuning. It was discussed that the difference in how much data is needed for fine-tuning and prompt tuning a model to obtain the same results is studied in previous work. The authors show empirically that fine-tuning requires much more data but do not give any explanation as to why that is the case. Also, it is questionable how well their measure generalises across datasets.

The webinar highlighted both foundational concepts and advanced techniques in multilingual NLP, with a focus on the unique challenges of low-resource languages. From bias measurement to parameter-efficient learning, the event underscored the importance of addressing societal challenges and optimising resources for better NLP systems. Researchers are advancing in areas like bias mitigation, data-efficient processing, and parameter-efficient learning, especially for low-resource languages, which could transform the application of NLP across the globe.

2.3 Webinar #3: Trustworthy AI: Trustworthy AI: Transparency, Fairness and Privacy

The third scientific webinar put focus on trustworthy AI and its transparency, fairness and privacy. This webinar was presented by Prof. Anders Søgaard (University of Copenhagen) and gave a high-level overview of the fundamental problems in trustworthy AI. The talk covered the discussion on two distinct flavours on transparency, inference and training transparency. The classical problem of inference transparency and why it is a hard problem were emphasised. It also looked at how this problem affects the epistemic value of applications of AI to the sciences, as well as at possible ways out. The presentation also addressed some common fallacies when addressing the trustworthiness of AI systems, in general.

The webinar began by introducing basic components of trust in AI, consisting of accountability (who's to blame if something goes wrong), transparency (what's going on in AI models), fairness (AI models are not to the advantage of very specific demographics)

and privacy (e.g., redistribution of copyrighted material and sensitive information). Starting from functions and neural networks are made of a series of functions, the talk looked into transparency and asked the question – are networks opaque? Two types (inference-opacity and training-opacity) of opacities were elaborated and discussed with examples. A presentation snapshot regarding trustworthy AI is shown in Figure 3.

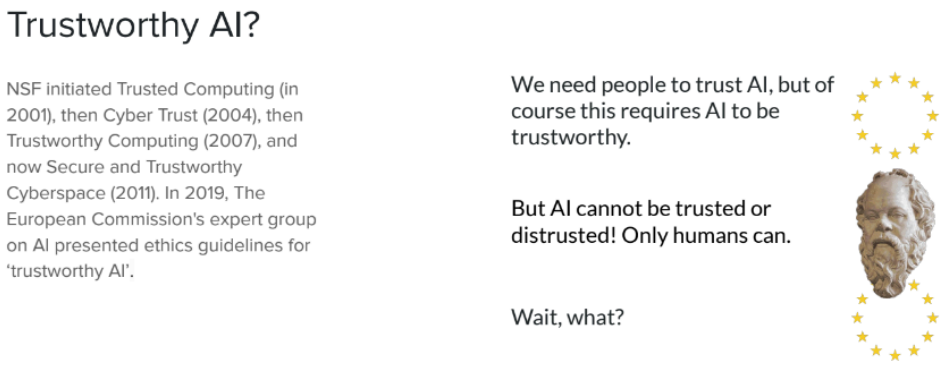


Figure 3. Importance of trustworthy AI in modern AI development (Presentation snapshot).

The argument that neural networks are not black boxes but white boxes and explanations (though being super long) was emphasised. The talk next introduced classical techniques and methods in explainable AI (XAI), including Occlusion and LIME. The evaluation strategies of different XAI methods were also traced and discussed, covering heuristic evaluation (input reduction, hot flip, etc.), human annotation (e.g., ERASER), human-in-the-loop (forward prediction, reverse turing test), and real-life experiments (QuizBowl, EkstraBladet).

Scientific stances of XAI were further explored in detail, including the physical stance, design stance, and intentional stance based on varying levels of abstraction. The fundamental trade-off between fathomability and faithfulness in physical stance, where most XAI methods are working with, were also discussed. The talk further touched some limitations of XAI methods such as low recall and linear approximation. The talk further talked about viewing XAI methods from a philosophy science perspective. Two types of philosophy theories were introduced in XAI, where transition theories (mostly physical stance theories/methods) explored transitions of a system under causal subsumption, while property theories concerned more about specific properties of a system that could explain the result. In the end of this section, the talk proposed three options to think about XAI methods. First, one could simply embrace the complexity. Second, one could embrace the approximate nature of these explanations, and be aware of their

unfaithfulness. Third, one could go with property theories. The talk then delved into some examples (world knowledge) of property theories with large language models.

The next part of the webinar put focus on fairness and privacy. Examples of inequality in different languages were shown to urge the necessity to ensure fairness. Copyrighted materials were demonstrated to be memorised by large language models, causing serious privacy issues. The presenter further emphasised the importance of considering the performance evaluation in different dimensions (e.g., multilinguality, fairness and bias, efficiency, and interpretability) and their intersections. Biases toward certain research areas (e.g., multilinguality, machine translation), languages (mainly English), performance metric (e.g., F1) were observed in recent ACL papers, and most of such papers were not multidimensional.

In the next part, the presenter brought up the discussion around trustworthiness and whether it would apply to machines. Two widespread definitions of trustworthy AI were discussed and robustness was said to be a crucial part to trustworthiness. In the final part of the talk, anthropomorphism and some known proper function fallacy were briefly discussed.

Q&As and lessons learned. Following the presentation, the presenter, Anders Søgaard, answered questions from the audience and engaged in discussions on different aspects of trustworthy and explainable AI. One participant raised a valuable point that the notion that any AI that is not designed with malicious intent is trustworthy is somehow questionable. There are many examples or tools when there was a good will and good intent, but there was an absence of knowledge or even ignorance when creating a tool. The presenter highlighted that the notion aforementioned was a biassed definition of trustworthiness provided by Karen Jones. The argument was correct and malicious intent was not necessary for things to go wrong. However, it provided a much better bridge to the ethics literature. The main takeaway would be that we wanted to build such bridges to ethics, philosophy of science, philosophy of mind and the legal literature. Another question focused on what a real black box algorithm would be like. The presenter talked about his view on black box algorithms that there are not black box algorithms that there is stuff we cannot see. He further elaborated on two types of black box concepts – the things unknown, and the things put on hold on purpose. The discussion further delved into the explainability and understandability of neural networks for humans. The presenter mentioned an interesting term, fathomability, which referred to the ability to wrap the head around something overwhelmingly complex. He expressed his opinion that the questions concerning understandability to humans was more of a pragmatic issue and the existence of the trade-off between faithfulness and fathomability. There was a lot of stuff like psychology without actually understanding the mechanics, without being able to take the physical stance. Still, there was science people could do. One participant further emphasised the distinction between explainability and reliability, and suggested that explaining deep learning systems in a very faithful and understandable way was almost a contradiction in terms. This perspective was highly agreed by the presenter and related to a common illusion that humans are explainable. The presenter shared his opinion on the fact that humans would not be considered as explainable by virtue of fMRI scans. A

discussion on how trust was established in humans and other things was further elaborated. Trust can be decomposed into quality and robustness, and it is important to estimate the robustness of the system. The ability to do that while blackboxing the system (not looking at the internal dynamics of the system), comes down to how much representative data can be gathered for inferring how model performance is. So, if in scenarios where the condition is stable, the trustworthiness of the system can be inferred from blackboxing it and just looking at the system behaviour. A crucial problem remains that such robustness is not there yet. In such cases, looking at the internal dynamics could be helpful. Lastly, the presenter and participants discussed the urgent work to use all tools in the toolbox including XAI methods taking physical stances, property theories and adversarial evaluation that can be used to estimate robustness.

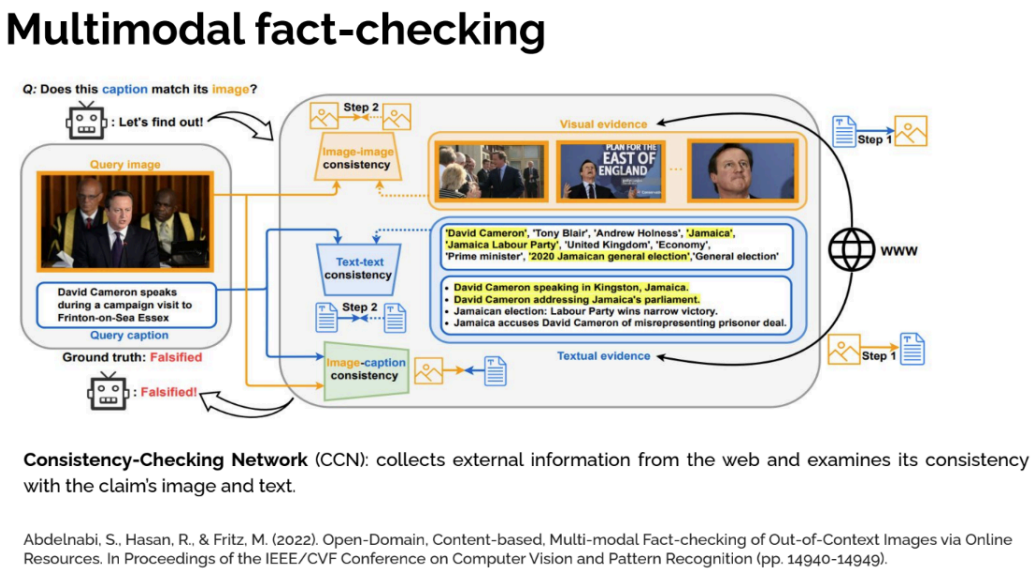
This webinar in DisAI series highlighted different components in trust and how they were applied to trustworthy AI. Beginning with transparency, the webinar introduced different types of opacity, classical XAI methods, and their evaluation strategies as well as limitations. The fathomability-faithfulness trade-off was specifically emphasised. To better understand the pros and cons of XAI methods, the webinar had a discussion in great detail from a philosophy science perspective. Fairness and privacy were also briefly discussed. This webinar underscored the importance of trustworthiness in AI, and pointed out interesting directions to follow in combating the trade-off between fathomability and faithfulness. The crucial role of robustness was brought up. In regards, the urgent work is to utilise all tools in the toolbox including XAI methods taking physical stances, property theories and adversarial evaluation that can be used to estimate robustness.

2.4 Webinar #4: Multimodal AI: Vision-Language, Multimodal Models & Applications

The fourth webinar in the DisAI series, hosted by CERTH and led by Principal Researcher Dr. Symeon Papadopoulos, focused on the topic of Multimodal AI and its applications. It provided an in-depth exploration of multimodal AI and its applications in media and journalism, emphasizing the challenges and advancements in the detection and prevention of disinformation. The presentation covered a broad range of topics, from fundamental concepts in modality and multimodality to advanced models like CLIP and Vision Transformers, as well as the role of AI in media and journalism in the era of deepfakes and disinformation.

The webinar began by introducing different modalities, including natural language, visual (image and video), auditory, haptic, and even electrocardiogram (ECG) data. Insights were provided on how these modalities are used individually and in combination, emphasizing the concept of multimodality. It was illustrated that combining two or more modalities can reduce uncertainty in tasks, as each modality offers complementary information. For instance, the integration of text and image data can provide richer insights than either modality alone. This section also addressed the increasing importance of multimodal systems in understanding complex inputs, particularly in the context of disinformation detection. An example of multimodal fact-checking is depicted in Figure 4.

The evolution of deep neural networks (DNNs), particularly in the field of computer vision, was traced, covering notable architectures such as DenseNet, ResNet, and VGG-16, and explaining their impact on visual recognition tasks. The discussion then delved into the workings of encoder-decoder architectures, which are central to many AI models. The functions of these architectures, including denoising, compression of information, and feature extraction, were explored in detail, providing participants with a foundation for understanding more complex models later in the presentation.



Consistency-Checking Network (CCN): collects external information from the web and examines its consistency with the claim's image and text.

Abdelnabi, S., Hasan, R., & Fritz, M. (2022). Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14940-14949).

Figure 4. Importance of multimodal fact-checking on verifying information across multiple data sources (Presentation snapshot).

A significant portion of the webinar was devoted to transformers, specifically vision transformers (ViT), and their role in multimodal AI. Self-attention and cross-attention mechanisms were compared, explaining how these techniques allow models to process different types of input, such as images and text. The concept of vision-language models was also introduced, which combine information from both visual and linguistic modalities to perform tasks such as image captioning, question answering, and sentiment analysis. Architectural components, including image encoders, text encoders, and multimodal fusion layers, were discussed, along with paradigms like pretraining, transfer learning, and distillation for fine-tuning models on specific tasks.

The webinar then focused on CLIP, one of the most advanced multimodal AI models developed by OpenAI. An explanation was provided on how CLIP functions, describing its training on a large dataset of image-text pairs and the process by which it learns to associate visual and textual information. This enables CLIP to perform zero-shot classification, where it can recognize objects in images without being explicitly trained on specific categories. The session highlighted the versatility and power of CLIP in understanding and processing multimodal data, making it a valuable tool in combating disinformation.

Applications of multimodal AI in media and journalism. In the final part of the webinar, the practical applications of multimodal AI in the fields of media and journalism were discussed. Several typical tasks where AI plays a crucial role were outlined:

- Monitoring large volumes of posts, articles, images, and videos on a daily basis.
- Parsing text, inspecting images, and skimming videos to gather information.
- Understanding key elements such as identifying who, what, and where.
- Searching for additional evidence using tools like reverse image search.
- Verifying the authenticity of images and videos to combat disinformation.

Specific applications, including Multimodal AI against Disinformation, Internet Meme Analysis, Misinformation Detection, and DeepFake Detection, were also addressed. Various types of mis- and disinformation, such as out-of-context misinformation, were discussed, and it was demonstrated how state-of-the-art approaches in multimodal deepfake detection can be applied to uncover manipulated content.

Q&As and lessons learned. Following the presentation, a Q&A session was held, providing an opportunity for participants to engage directly with the speaker. During the Q&A session of the webinar, several insightful comments and questions regarding the CLIP model and its relevance in the context of modern visual-language models were addressed. One participant raised a valid point about the robustness of CLIP to domain shifts, arguing that the results presented in the original CLIP paper may not adequately demonstrate this robustness. They noted that unlike networks trained on datasets like ImageNet, which encounter truly out-of-domain images (e.g., sketches of bananas), CLIP's training dataset includes a broader range of images. This perspective was acknowledged as a meaningful critique. Another question focused on the continued utility of CLIP, especially given the popularity of large language models (LLMs). It was explained that while CLIP may be considered outdated, its simplicity and effectiveness make it valuable for specific tasks, particularly in multimodal retrieval. The speaker highlighted that leveraging CLIP's powerful vision and text encoders can often yield better results for tasks like multimodal fact-checking, without the resource intensity associated with using newer, larger models. Concerns were also raised about CLIP's limitations in detecting fine-grained details in images, as it uses a frozen visual encoder. Suggestions were made regarding the potential for fine-tuning the visual encoder to enhance its capability to capture these details. The speaker agreed, noting that current trends in the field are moving towards embedding images and text in a common space, utilizing a single transformer to process tokens generated from both modalities. This Q&A session provided valuable insights into the ongoing evolution of multimodal AI and the challenges and opportunities presented by both older and newer models in addressing complex tasks in this domain.

The fourth webinar in the DisAI series provided several valuable lessons regarding the use of multimodal AI in combating disinformation. One of the key takeaways was the significant advancements made in the integration of multiple modalities—such as vision, language, and auditory data—which have proven to enhance the detection and analysis of disinformation. Models like CLIP exemplify the strong potential for real-world applications, particularly in fields like journalism, where the need for accurate and efficient disinformation detection is critical. Additionally, the session underscored the practical

relevance of multimodal AI to common journalistic tasks such as monitoring content, verifying sources, and conducting searches for evidence. However, despite these promising applications, the deployment of these models still faces challenges, particularly when it comes to scaling them for everyday use without compromising on accuracy or performance. Another important lesson centered on the issue of trustworthiness in AI systems. Participants raised concerns about bias, false positives, and the necessity for AI systems to offer explainable outputs. Building trust in these technologies, especially in sensitive areas like disinformation detection, remains a complex and ongoing challenge. The webinar emphasized the need for further research in this area to ensure that AI systems are both reliable and transparent. Overall, while the rapid advancements in multimodal AI offer exciting possibilities, the webinar highlighted the importance of continuing to address the open challenges, particularly around trustworthiness and the effective representation of multimodal data.

3 Conclusions

The DisAI webinar series provided valuable insights into key areas of AI research, focusing on disinformation detection, multilingual NLP, trustworthy AI, and multimodal AI. Each session highlighted significant advancements in AI while addressing the technical, ethical, and societal challenges associated with its application. The discussions on multilingual NLP emphasised the importance of creating inclusive AI systems, particularly for low-resource languages. Techniques such as self-supervised machine translation and parameter-efficient learning were presented as solutions to improve multilingual processing. However, biases in multilingual models remain a pressing issue, with researchers stressing the need to address cultural biases. The session on trustworthy AI focused on essential factors such as transparency, fairness, and privacy. The complexity of making AI systems more explainable was explored, particularly the fathomability-faithfulness trade-off, which illustrates the challenges of creating transparent models without compromising accuracy. The discussions highlighted the need for cross-disciplinary collaboration, integrating insights from philosophy, ethics, and law to build more accountable AI systems. The exploration of multimodal AI demonstrated its potential in disinformation detection, particularly through models like CLIP and Vision Transformers. These technologies show promise in integrating multiple data modalities, such as text and images, to detect and verify manipulated content. However, challenges related to scalability, accuracy, and trustworthiness persist, and further research is needed to improve the reliability of these systems in critical applications, such as media and journalism. In conclusion, the DisAI webinar series showcased the significant progress being made in AI research, with promising developments in multilingual NLP, trustworthy AI, and multimodal systems. Nevertheless, achieving ethical and transparent AI remains a challenge, requiring sustained efforts from both the technical and ethical perspectives to ensure AI's responsible use in addressing global issues like disinformation.

Appendix: Supplementary Material

Webinar #1: Disinformation combatting: Challenges, methods and tools

Slides: Presentation (Google drive)

Video:
DisAI Webinar #1: Disinformation combatting Challenges, methods and tools (youtube.com)

Participants list (24 participants): List (Google docs)

Webinar #2: Multilingual NLP: Overview and Low-Resource Multilingual Processing

Slides: Presentation (Google drive)

Video:
DisAI Webinar #2: Multilingual NLP Overview and Low Resource Multilingual Processing (youtube.com)

Participants list (20 participants): List (Google docs)

Webinar #3: Trustworthy AI: Trustworthy AI: Transparency, Fairness and Privacy

Slides: Presentation (Google drive)

Video:
DisAI Webinar #3: Trustworthy AI (youtube.com)

Participants list (30 participants): List (Google docs)

Webinar #4: Multimodal AI: Vision-Language, Multimodal Models & Applications

Slides: Presentation (Google drive)

Video:
DisAI Webinar #4: Multimodal AI (youtube.com)

Participants list (34 participants): List (Google docs)