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

www.ai4trust.eu

AI4TRUST

AI4TRUST

D4.2 EXPLAINABILITY AND TRANSPARENCY REPORT AND AI TOOLS

PARTNERS

FONDAZIONE BRUNO KESSLER
CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS
UNIVERSITÀ DI TRENTO
NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS"
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
GDI Global Disinformation Index
sky
SAHER EUROPE
DEMAGOG
MALDITA.ES
ASTIKI MI KERDOSKOPIKI ETAIRIA KENTRO KATAPOLEMISIS TIS PARAPLIROFORISIS / CIVIL NON-PROFIT COMPANY KENTRO KATAPOLEMISIS TIS PARAPLIROFORISIS
EURACTIV
ASOCIATIA DIGITAL BRIDGE
EUROPEJSKIE MEDIA SP ZOO
FINCONS GROUP
UNIVERSITY OF CAMBRIDGE

Title	D4.2 Explainability and Transparency Report and AI Tools
Editor	UCAM
Main author(s):	Gina Neff, Stefanie Felsberger, Hugo Leal
Dissemination level	☐ CO: Confidential, only for members of the consortium (including the Commission Services) ☐ RE: Restricted to a group specified by the consortium (including the Commission Services) ☐ PP: Restricted to other programme participants (including the Commission Services) ☒ PU: Public
Reviewers	SAHER
Status	☐ Draft ☐ WP Manager accepted ☒ Coordinator accepted
Due date	30/04/2024
Delivery date	31/05/2024
Work Package:	4
Lead partner for this deliverable:	UCAM
Partner(s) contributing:	CERTH, CNRS, UPB, UNITN
Contributor(s):	Evlampios Apostolidis, Symeon Papadopoulos and Vasileios Mezaris (CERTH), Emmanuel Lazega, Paola Tubaro, LaCamille Roth, and Yasmine Houri (CNRS), Thomas Louf (FBK), Gina Neff, Hugo Leal, and Stefanie Felsberger (UCAM), Elena Pavan, Matteo Scianna (UNITN), Horia Cucu, Octavian Pascu and Dan Oneață (UPB),

Summary of modifications

VERSION	DATE	AUTHOR(S)	SUMMARY OF MAIN CHANGES
0.1	8/03/2024	Stefanie Felsberger (UCAM)	Table of Contents
2	15/03/2023	Evlampios Apostolidis (CERTH), Emmanuel Lazega, Paola Tubaro, LaCamille Roth, and Yasmine Houri (CNRS), Thomas Louf (FBK), Gina Neff, Hugo Leal, and Stefanie Felsberger (UCAM), Elena Pavan, Matteo Scianna (UNITN), Horia Cucu, Octavian Pascu (UPB)	First Outline
3	14/04/2023	Evlampios Apostolidis (CERTH), Emmanuel Lazega, Paola Tubaro, LaCamille Roth, and Yasmine Houri (CNRS), Thomas Louf (FBK), Gina Neff, Hugo Leal, and Stefanie Felsberger (UCAM), Elena Pavan, Matteo Scianna (UNITN), Horia Cucu, Octavian Pascu (UPB)	First Incomplete Draft
4	15/05/2024	Evlampios Apostolidis, Symeon Papadopoulos, Vasileios Mezaris (CERTH), Emmanuel Lazega, Paola Tubaro, LaCamille Roth, and Yasmine Houri (CNRS), Thomas Louf (FBK) Gina Neff, Hugo Leal, and Stefanie Felsberger (UCAM), Elena Pavan, Matteo Scianna	Complete Draft

		(UNITN), Horia Cucu, Octavian Pascu (UPB)	
5	20/05/2024	Evlampios Apostolidis, Symeon Papadopoulos, Vasileios Mezaris (CERTH), Emmanuel Lazega, Paola Tubaro, LaCamille Roth, and Yasmine Houri (CNRS), Thomas Louf (FBK) Gina Neff, Hugo Leal, and Stefanie Felsberger (UCAM), Elena Pavan, Matteo Scianna (UNITN), Horia Cucu, Octavian Pascu (UPB)	Final Draft
6	21/05/2024	Evlampios Apostolidis, Symeon Papadopoulos, Vasileios Mezaris (CERTH), Emmanuel Lazega, Paola Tubaro, LaCamille Roth, and Yasmine Houri (CNRS), Thomas Louf (FBK) Gina Neff, Hugo Leal, and Stefanie Felsberger (UCAM), Elena Pavan, Matteo Scianna (UNITN), Horia Cucu, Octavian Pascu (UPB)	WP Manager Approved
7	30/05/2025		Checked by Saher

Table of contents

List of abbreviations

ABBREVIATION	MEANING
AI	Artificial Intelligence
ACM	Association for Computing Machinery
AI HLEG	High-Level Expert Group on Artificial Intelligence
ChatGPT	Chat Generative Pre-trained Transformer
CNN	Convolutional Neural Networks
DF	DeepFakes
DFDC	DeepFake Detection Challenge

EERs	Equal Error Rates
EU	European Union
F2F	Face2Face
FS	FaceSwap
GAN-loss	Generative Adversarial Network loss
ICMR	International Conference on Multimedia Retrieval
MAD	Multimedia AI against Disinformation
MBConv	Inverted Bottleneck convolutions
NES	Natural Evolution Strategies
NT	NeuralTextures
LIME	Local Interpretable Model-Agnostic Explanations
LRP	Layer-Wise Relevance Propagation
RISE	Randomised Input Sampling for Explanation
SHAP	SHapley Additive exPlanations
SMP	Social Media Platform

SNA	Social Network Analysis
SLIC	Simple Linear Iterative Clustering
SVM	Support Vector Machine
URL	Uniform Resource Locator
xDNN	Explainable Deep Neural Network

1. Introduction

This report sets out to accomplish three things. First, it defines trustworthy and explainable AI in terms for the development of the AI4Trust platform. The report does this by translating regulations, norms and values that make up users' assessments of trustworthy and explainable into design choices that the AI4Trust team will follow. Second, the report covers socio-linguistic, user contextual, and social network factors of the design of the AI4Trust platform by focusing on how social network analysis can provide useful signals for detecting misinformation and disinformation across platforms. Third, this report evaluates the algorithms in development by the AI4Trust team for being trustworthy and explainable. These include the algorithms for audio and video deepfake detection.

This report concludes with key points for the development of the AI4Trust platform. First, the report argues for using contextual indicators to support claims about disinformation operations. Second, the report argues that mitigation measures cannot be equally deployed across all online platforms and services and suggests strategies for handling this. Third, asymmetries of information between platforms and the public create challenges for accountability and regulation. The AI4Trust project can partially address these challenges. Fourth, AI4Trust privileges human-centred AI technologies and research methods to both map and counter information disorders, keeping with EU values and regulations.

The report concludes by looking forward to future mitigation strategies. The AI4Trust platform rebalances the current state of information asymmetry through the principle and practice of platform observability. Data access will continue to be a challenge for a public-side monitoring system. The AI4Trust platform enlarges the spectrum of mitigation beyond content moderation by using a range of contextual and content-based measures.

2. Trustworthy and Explainable AI

The aim of the AI4TRUST project is to provide trustworthy AI solutions to help different stakeholders in countering and detecting mis/dis/mal information. Many different frameworks and approaches exist that define trust in AI and how it can be operationalised or implemented in the development of AI systems. The first section of this report aims to provide a common understanding of trustworthiness and explainability of AI for the AI4TRUST project by drawing on the EU Ethics Guidelines for Trustworthy AI (2019). Many of the existing frameworks around trust in AI are not tailored at the specific needs and particularities of different sectors. The second section of this report discusses why a framework on trust and explainability is important for the AI4TRUST consortium and how specifically the stakeholders who are envisioned to use the AI4TRUST tool understand and assess trustworthiness of AI technologies in their work. In short, it outlines what trust in AI means for the AI4TRUST project.

These insights are informed by fieldwork conducted by the University of Cambridge with stakeholders in media and journalism, policy, fact checkers, academia and education, and other civil society organisations active in the field of disinformation. This report discusses the different factors stakeholders considered when deciding whether they could trust a new AI tool: values, reputation, and bias of the organisation developing a tool, accuracy and reproducibility, transparency and explainability, privacy and data governance, accountability and human oversight, longevity and reliability of the tool, access and accessibility, bias, and stakeholder engagement. Finally, this report calls for a comprehensive engagement with the question of trust and explainability of the overall AI4TRUST consortium and the establishment of a work plan to determine how the different WPs can contribute to developing trustworthy AI solutions to tackle disinformation.

2.1. Frameworks & Definitions

As AI systems govern more and more aspects of society, the question of how to develop ethical AI systems has become more urgent. Many public institutions, governments, and companies have issues and developed frameworks or principles with the aim of making AI systems more ethical, safe, transparent, just, trustworthy, accountable, sustainable, and/or fair (Jobin, Ienca, and Vayena 2019). The AI4TRUST project, which is based in Europe and founded on European values, takes the European Union's Ethics Guidelines for Trustworthy AI (AI HLEG 2019) as foundational document to think about what trust in AI technologies to detect and counter mis/dis/malinformation means.

The EU's approach has several benefits that align with the aims of the AI4TRUST project. First, the approach focuses on the full life cycle of AI systems while other approaches often target AI development or auditing AI systems after their deployment. Given the fast developing context of mis/dis/malinformation the AI4TRUST platform seeks to intervene, an approach that tackles the development, deployment, and use of AI systems is appropriate. Second, the document takes a socio-technical approach to trust in AI systems.

> Striving towards Trustworthy AI hence concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle. (AI HLEG 2019, 5)

Civil society organisations have long called for a move towards 'socio-technical' approaches to AI ethics and trust, since algorithmic biases are often "inextricably linked to power asymmetries and structural inequity" (Kak and Myers West 2023), which cannot be fixed through technical solutions alone. The challenges of mis/dis/malinformation are socio-technical in nature and therefore require socio-technical approaches to solutions, although neither social nor technical solutions alone are sufficient. AI-generated visual content cannot be detected without technical tools for verification. Technical solutions—detection tools, labelling, or watermarking of AI-generated content—do not suffice in countering disinformation. The EU approach encompasses many existing frameworks such as explainability or transparency under the umbrella of trust. The EU framework builds on three components: (1) AI systems need to comply with existing regulations and requirements for data protection, data governance and human rights, (2) AI systems need to be ethical (going beyond legal compliance), (3) technically robust, meaning safe against cyberattacks and misuse. All these components align with the aim and requirements of the AI4TRUST project.

Overall, the EU framework is built on four ethical principles and seven requirements, which are explained below. The ethical principles demand

1. AI systems should foster **human autonomy**, support human decision making, rather than "unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans" (AI HLEG 2019, 12-13).
2. AI systems should **do no harm**: not cause or make worse existing harm, be safe and secure technically and not open to misuse.
3. AI systems should be **fair**: procedurally, this entails procedures to seek redress against decisions made by AI systems and humans operating them as well as clarity on who is accountable. Substantively, this entails elimination of bias, discrimination or stigmatisation, equal access to technologies, knowledge, services and benefits from AI systems
4. AI systems should be **explainable**: processes, decisions and purposes of AI systems must be clearly and understandably laid out. (AI HLEG 2019, 12-13).

These principles are translated into seven different requirements which inform how AI systems should be developed and deployed:

1. **Human agency and oversight**: respect for fundamental rights, fostering human agency and human oversight
2. **Technical robustness and safety**: resilience against attacks, security of AI systems, safety, accuracy, reproducibility
3. **Privacy and data governance:** including respect for privacy, quality and integrity of data sets, and access to data
4. **Transparency:** traceability, explainability and communication in understandable manner
5. **Diversity, non-discrimination and fairness:** avoiding unfair bias, accessibility and universal design, and meaningful stakeholder participation
6. **Societal and environmental wellbeing:** sustainability and environmental concerns, social impact, AI systems that uphold democracy
7. **Accountability:** auditability of AI systems, minimisation and reporting of negative impact, trade-offs and procedures for seeking redress

2.2. Specific user needs and trust requirements in the field of mis/dis/malinformation

This section addresses why trust and explainability in AI matter in particular to the imagined end users who work in the field of mis/dis/malinformation. This section is informed by fieldwork conducted for the AI4TRUST project with potential end users of the AI4TRUST platform. A range of different qualitative methods including interviews, focus groups, and ethnography were employed.

Five focus groups were conducted with policy professionals, experts from different civil society and human rights organisations, AI start-ups, fact checkers, media professionals, public servants, and educators. The aim of the focus groups was to determine the main priorities and needs of potential users of a tool to detect and counter online mis/dis/malinformation and assess how these users might engage with AI tools to do their work. Ethnographic fieldwork was conducted at Maldita, a member of the consortium, to gain contextual understanding of what tools and strategies fact checking organisations employ in their work against mis/dis/malinformation. This fieldwork consisted of participant observation and eight interviews with professionals working in all different areas of the organisation. A round of expert interviews with stakeholders in the field of disinformation was conducted with a specific focus on the role of trust in tools stakeholders employ to fight mis/dis/malinformation. In conversations with research participants, a range of topics were discussed that fed the analysis below. Often participants described the kinds of tools that they use in their current work. They explained how and why they were both useful and trusted. They described their process of assessing any new tool or platform they encountered whether it was trustworthy or whether they would use it, and often the two were closely interlinked.

Participants worked in journalism, policy or politics, academia, media literacy, or a range of civil society organisations concerned with online safety and trust in information in different European countries. Participants were all experts in their field, but their areas of expertise differed extensively: while some demonstrated extensive technical knowledge of AI systems, others had little technical expertise but deep understanding of either relevant policy or educational approaches to media literacy. Every participant already had a specific understanding or perception of the role of AI algorithms in society—for better or worse—and this perception also influenced the ways in which they discussed trust in AI systems. Some policy makers stressed they could never fully trust a fully automated AI system to provide them insights into the dynamics of (dis)information flows. Journalists for example experienced the disruption generative AI tools such as ChatGPT presented in their work place where they were perceived as creating more work for them and were at the same time in need of support but also sceptical of the possibilities AI systems had to alleviate work pressures.

This range of background knowledge and expertise and attitudes to AI is crucial to keep in mind when it comes to understanding how and why different participants trust an AI tool. Overall, participants prioritised different factors in their assessments of whether a tool could be trusted—

often these were also contradictory. For example, some participants needed to understand step-by-step how an AI system worked, while others found it only important to be able to verify and reproduce its accuracy. It is crucial to keep these nuances in mind as the AI4TRUST moves to design its educational materials, the outputs of its platform, and the overall interface.

In general, the potential end-users' assessment of whether a tool or platform was trustworthy was deeply informed by their area of expertise and the context in which participants worked. For many participants, a lot is at stake when they assess whether or not to integrate a new tool into their existing toolbox. First, they are constricted by capacity and time. Many participants work in fast-paced environments or in organisations that have funding or staffing constraints. They often must prioritise or decide quickly about what works, what does not, and what is not worth their time. A tool they cannot trust to make better decisions or to give them accurate insights and results is quickly dismissed. Second, many participants stressed that in the disinformation field, trust is hard won but easily lost. Fact checking organisations, research organisations and civil society organisations work hard and diligently to establish the trust of the public and different stakeholders. They might be at risk from legal repercussions[1] or targets of bad actors who may use their mistakes made to undermine the organisation's legitimacy. Stakeholders who put their own names on the line may be hesitant to rely on a new tool without being able to fully understand who made it, how accurate it is, where it can be wrong, and how it stores data. This is because by trusting an external tool, they also put their own reputation on the line. Third, many different verification or analysis tools are already used by different stakeholders and new tools are entering the market. Many participants draw from a range of different tools in order to do their work. Any new tool they put their trust in has to offer something in addition to their existing toolkit. Nevertheless, many participants saw a great need for better AI tools to tackle disinformation. Finally, participants also expressed concerns that existing or new AI tools could be counterproductive, especially by falsely labelling truthful information as "fake" or by providing difficult to interpret results which can increase confusion around whether or not a piece of information is false or not.

2.2.1.Organisation, Reputation & Uptake

Often the first step in how participants evaluated whether they would trust an AI tool had nothing to do with the tool itself but rather the **organisation** behind the tool. As an initial step, many first assessed who developed the tool: many asked what was the organisation's expertise and background, where did their funding come from, what kind of bias could be detected, and what was the tool's intended purpose? One interviewee from the education sector who lacked technical expertise to assess for themselves how any given tool worked, explained that they would have to trust the expertise and reputation of the people or organisation behind a tool, assuming that a reputable organisation would have done their due diligence in developing a trustworthy tool.

[1] For example the UK-based Centre for Countering Digital Hate was sued by Elon Musk and X (formerly Twitter) for allegedly causing X tens of millions of loss in advertising revenue after publishing a report on misinformation and hate speech on the platform (CCDH 2024).

Reputation overall played an additional factor in people's trust: one policy maker mentioned that tools or platforms built by the government of one state rather than an international organisation, for example, were much less trustworthy. Additionally, word of mouth was mentioned by one stakeholder: she explained that if well-established fact-checkers and analysts were using or recommending a particular tool, this would increase her own confidence in a tool's trustworthiness. This was closely interlinked with a **tool's uptake**: the more widely used, the more likely to trust a tool this participant was. Others explained that they would look at a tool's longer term impacts and the extent to which they deliver their promises over time, before trusting a tool.

Questions about the Organisation:

1. Who is behind the tool? Is it a reputable organisation?
2. What is their expertise and background?
3. How is it funded?
4. What is its intended purpose?
5. What are the organisation's biases and values?

2.2.2.Accuracy & Reproducibility

The question of accuracy and its reproducibility was often the next factor participants discussed. First and foremost, all participants required any tool they trusted to be accurate. This in particular refers to verification tools that detect whether content is AI generated. All participants without fail stressed a need for better and more accurate AI detection tools in images, videos, sound and text.

A tool's **accuracy** in detecting AI-generated content was a main factor in how participants assessed trust, but depending on participants' backgrounds they assessed accuracy differently. First, as mentioned above, those who could not test the tools technical capacities used trust in the organisation as a proxy (see Reputation). Second, many participants stressed that seeing how the tool communicated its accuracy & limitations was key to them trusting it (this is further explained below in the section on Explainability and Transparency). Third, most participants needed to test and verify a tool's accuracy themselves on real and fake examples to assess its reliability and reproducibility which allows them to build confidence in the tool.

There were several layers to the way participants tested the tool's **reproducibility**. For some, the reproducibility of results was also a necessity for any tool they employed. For those who publish their work—for example debunks, research papers, journalistic articles, or reports—their readers and audience need to be able to verify these results themselves. Often the audience includes other stakeholders and experts or those seeking to expose flaws or faults, which makes it even more important that stakeholders can trust both a tools' accuracy and its reproducibility. For most participants being able to test and verify the tool's accuracy for themselves was the most crucial way to establish trust in the tool. Some even stated that they did not need to know a step-by-step breakdown of how the tool worked, but were only interested in its consistent accuracy.

Questions about Accuracy & Reproducibility:

1. Can the accuracy of the tool be tested and verified?
2. Can others access the tool and verify its accuracy for themselves?

2.2.3. Transparency & Explainability

For a subset of participants, the fact that the tool was accurate and they could verify its accuracy, was not sufficient. They needed a tool to transparently explain accuracy, processes, and methodology. Transparency and explainability are therefore closely interlinked for some potential users, even though these concepts are different. The EU framework refers to transparency as a requirement to make "elements relevant to an AI system: the data, the system and the business models" transparent to the user (AI HLEG 2019, 18). Explainability is more closely linked to the ability of the user to understand what is made transparent to them: the document defines explainability as the ability of the AI system to provide explanations and reasons for its individual decisions or predictions in a way that humans can understand (AI HLEG 2019, 18). This requires a different approach meaning that explanations of how the AI system makes decisions need to be transparent but also explained in a way that different users with different expertise can understand these explanations. The potential end users of the AI4TRUST tool are all experts in mis/dis/malinformation, but their technical understanding and expertise of AI systems varies drastically. Explainability as a goal needs to take these differences seriously in order to explain the AI4TRUST tool to end users in an understandable way—without this many end users will not trust the tool and opt for one of the many competitors instead.

With regards to the transparency and explainability of the tool's **processes and methodology**, participants' needs and approaches diverged. To some only the results and outcomes of the tool mattered (how accurate, error rates), others trusted a tool implicitly when they could assess that there was a scientific basis without necessarily going through a step-by-step assessment of that scientific basis and its implementation. But to other users, in order to establish trust they needed to be able to pick apart the inner workings of a tool: from the underlying data sets, to the methodology or AI algorithms used, the rationale behind specific outcomes in order to understand how decisions were made and results produced. For the AI4TRUST platform, this means that transparency and explainability need to both satisfy users who need to establish a scientific basis at a quick glance and those who need to understand the system in detail.

Transparency and explainability of AI tools was mentioned as one of the most important factors by almost all participants. How the tool's **accuracy rates** and its track record of false positives were made transparent and explained was crucial to participants' ability to assess the tool. Especially with regards to tools detecting AI generation, participants stressed time and again needing to know the tool's **strengths, weaknesses and error rates** in order to be able to assess any limitations. One participant mentioned their concern about false positives caused by recompression and re-uploading of manipulated images or videos and expressed wanting to know any tool's ability to accurately detect them. Fact checkers and journalists especially expressed frustration or even exasperation at existing AI detection tools. While being aware of the tools limitations, the tensions erupt because both fact checkers and journalists need to be able to determine with 100% accuracy whether something is fake or not, true or false, AI generated or an actual photo. One participant shared that they struggled to integrate tools in their work that would provide them with, for example, a 46% probability rate that an image is AI generated. As fact checkers, that result does not help them make assessments but rather compounds confusion. This point was compounded by other participants who pointed out that AI tools that provide users with difficult to interpret results could be counterproductive and further erode fact checkers or journalists ability to determine facts. Therefore the explainability of the AI4TRUST tool one step further by assisting its end users in interpreting the results and scores the toolkit produces.

Questions about Transparency:

1. *Does the tool explain how it makes decisions and produces verdicts or results?*
2. *Can this process be independently verified by a user?*
3. *What data sets and algorithms are employed in the AI system?*
4. *What is the scientific basis and methodology behind the tool?*

Questions about the Explainability:

1. *Does the tool explain its limitations? What can it do and what can it not do?*
2. *What are the tools' error rates?*
3. *Do I understand the information provided by the tool?*
4. *Does the tool help the user to interpret results rather than just present results?*

2.2.4. Privacy & Data Governance

Several participants argued that what was most important to them was that any AI tool they used followed and upheld existing rights to (data) privacy and basic human rights. Stakeholders in policy and civil society raised concerns over tools that needlessly invaded public users’ privacy and stated they would not use such tools, as they misaligned with their organisations’ values. Especially, policy makers at EU levels stressed that there was institutional hesitancy to trust external tools to analyse data where there is a lack of clarity over data storage and data streams. Clear explanations and transparency of a tool’s intended purpose, what data it uses, how data is stored and used further were required by interviewees.

Questions about the Privacy & Data Governance:

1. Who is *What data is used and how?*
2. *How is data stored? Can users protect the privacy of information they share with the tool?*
3. *As fact checkers, does the tool own the data I input?*
4. *Does the tool respect human rights?*

2.2.5. Access & Accessibility

Another concern that was crucial for participants were questions around access and accessibility. As explained above, when organisations or stakeholders use verification tools in their work and publications, others need to be able to verify the accuracy of the tool used to inform organisations' publications, debunks, or recommendations. For many participants, this was an intrinsic aspect of any trustworthy tool in their toolbox.

By extension, participants also stressed that the tool needed to be accessible for the different users and their level of expertise and understanding. This factor is closely tied up with explainability.

Questions about Access & Accessibility:

1. Is the tool accessible to other organisations, like civil society?
2. Is the tool designed in an accessible way?
3. Can it be understood by a range of users with different experiences?

2.2.6. Accountability & Human Oversight

Other participants emphasised the importance of human oversight over the outputs of AI systems as well as the existence of accountability processes. Especially fact checkers cautioned against fully automated tools for fact checking without human oversight and expressed deep trust of such AI systems. Policy makers equally stressed mistrust on AI systems without human oversight. The AI4TRUST platform should therefore clearly explain how and where humans have shaped the development of the AI system and where humans assess and evaluate the AI systems decision making.

In relation, participants often mentioned accountability as an important factor in establishing trust. Accountability also included the clear communication of data privacy and data governance principles, the transparency of the AI system and its explainability.

Questions about Accountability & Human Oversight:

1. Is there human oversight of decisions made by the AI system?
2. How does this oversight work?

2.2.7. Other Factors: Longevity, Bias & Stakeholder Engagement

Finally, individual participants mentioned important aspects that foster trust in AI tools which were important but mentioned by less interviewees. One participant mentioned the sustainability or longevity of a tool as a crucial aspect in fostering trust. They argued that any AI tool needs to be maintained and kept up to date in order to maintain its functionality and accuracy. They argued that many new tools enter the market and disappear. For a tool to be trustworthy, the participant argued he needed to rely on the tool's availability over time.

One stakeholder argued that they often assessed trustworthiness of a verification tool by asking whether relevant stakeholders were involved in the development of a tool. He referenced AI companies not meaningfully consulting the expertise of those on the frontlines working against disinformation when developing solutions. Others mentioned the tool should avoid politicised or biassed language.

Other Questions:

1. Does the tool use neutral and objective language?
2. Is the tool regularly updated and maintained over time?
3. Have the tool developers meaningfully engaged with stakeholders who work in the field of mis/dis/malinformation?

2.3.Trustworthiness assessment "list" of potential end-users

Based on our research, we recommend the following list of questions as ones that inform end-users' assessment of trustworthiness in AI tools:

1. Organisation
    1. Who is behind the tool? Is it a reputable organisation?
    2. What is their expertise and background?
    3. How is it funded?
    4. What is its intended purpose?
    5. What are the organisation's biases?
2. Accuracy & track record of accuracy
    1. How accurate is the tool? What are its error rates?
3. Reproducibility
    1. Can the accuracy of the tool be tested and verified?
    2. Can others access the tool and verify its accuracy for themselves?
4. Transparency
    1. Does the tool explain how it makes decisions and produces verdicts or results?
    2. Can this process be independently verified by a user?
    3. What data sets and algorithms are employed in the AI system?

    4. What is the scientific basis and methodology behind the tool?
5. Explainability
    1. Does the tool explain its limitations?
    2. What can it do and what can it not do?
    3. What are the tools' error rates?
    4. How does the AI system make decisions?
    5. Does the tool help the user to interpret results rather than just present results?
6. Privacy & Data Governance
    1. What data is used and how?
    2. How is data stored? Can users protect the privacy of information they share with the tool?
    3. As fact checkers, does the tool own the data I input?
    4. Does the tool respect human rights?
7. Accountability & Human Oversight
    1. Is there human oversight of decisions the AI system makes? How does this oversight work?
8. Bias
    1. Does the tool use neutral and objective language?
9. Sustainability
    1. Is the tool updated and maintained over time?
10. Reputation & Uptake of the Tool
    1. Is the tool developed by a reputable organisation?
    2. Are other actors in the field using and recommending the tool?
11. Accessibility & Values
    1. Is the tool accessible to other organisations, like civil society?
    2. Is the tool designed in an accessible way and can it be understood by a range of users with different expertise?
12. Stakeholder Engagement
    1. Does the tool engage with stakeholders who work in the field of disinformation?

2.4. Conclusion

This section discussed the importance of trust and explainability for the development of the AI4TRUST platform. According to fieldwork conducted for the AI4TRUST project with stakeholders and experts in the field of disinformation, potential end users of the AI4TRUST platform assess trust according to a range of different factors: values, reputation, and bias of the organisation developing a tool, accuracy and reproducibility, transparency and explainability, privacy and data governance, accountability and human oversight, longevity and reliability of the tool, access and accessibility, bias, and stakeholder engagement.

Two conclusions can be drawn from the fieldwork for the future work of the AI4TRUST consortium.

1. **End-users require additional explanations to assist in interpreting results:** Many stakeholders raised challenges in understanding and interpreting results presented by tools that detect AI-generated content. Explanations of AI detection tools therefore require additional information to assist stakeholders in interpreting results such as probability scores. This information should include information on (1) processes: what algorithm has been used, what has it been trained to detect; (2) results: such as probability scores, heat maps, etc.; (3) limitations: what has this tool not been able to detect. This information could be provided in a "report card" end-users can download and cite. The challenge remains to determine how much detail can add to the existing challenge in interpreting results and processing information on the side of the end-user, and how much detail is not enough.

2. **Trust and explainability in AI solutions require a socio-technical approach:** Trust is placed both in the AI systems and the humans developing and operating AI systems, as evidenced in participants' questions about the organisation behind an AI tool. Many participants stressed a need to understand the step-by-step decision making of any AI tool they integrate in their work: both verification tools and analysis tools. Therefore, the question of trust and explainability is one the overall AI4TRUST consortium needs to tackle. End-users of the AI4TRUST tool need to be able to understand how each of the AI tools in the toolkit makes decisions and works in order to integrate them in their work—this includes verification tools, the disinformation early warning system and any other tools that feed into the early warning system. This section calls for a comprehensive engagement with the question of trust and explainability of the overall AI4TRUST consortium and the establishment of a work plan to determine how the different Work Packages can contribute to developing trustworthy AI solutions to tackle disinformation.

3. Socio-linguistic, user contextual, and social network factors of design for disinformation and mitigating misinformation

This section of the deliverable defines how Social Network Analysis (SNA) can be employed as part of the pre-processing steps in the AI4TRUST platform. As part of the pre-processing stage, SNA can support the identification of instances of dis/misinformation by helping to disentangle the complexities of the massive generation of online content and directing users' attention towards specific content. Future developments in both tasks 4.1 and 4.2 will help show how SNA can contribute to detecting patterns of dis/misinformation.

3.1. Signals of inappropriate information on social media platforms with SNA

3.1.1. SNA as a preprocessing tool

As part of a pre-processing stage in the AI4TRUST platform, **SNA is the preferred modelling technique** to render a variety of interconnected social media users' behaviours—from content posting to content sharing, commenting, and reciprocal interactions. These behaviours may **signal and/or sustain the diffusion of mis/disinformation**. At platform scale, the daily volumes of digital content production and circulation are so massive that they require the use of analytical techniques to reduce overall levels of complexity. At the same time, content creation, dissemination processes, and user interactions vary based on the unique characteristics of different platforms. We label these characteristics **platform ontologies**. Platform ontologies facilitate a range of semantic and social dynamics that require to be investigated in depth and in relation to specific platforms.

Thus, **commitments to perform cross-platform analysis face severe challenges** (see D2.1). The context for mis/disinformation is inherently heterogeneous. The specific tasks of finding signals for possible inappropriate information creation and circulation and detecting mis/disinformation super-spreaders can be dealt with in practice by adopting a **multitiered relational approach.** This means different platforms are traced and investigated simultaneously, but in parallel, and the presence/circulation of similar content is used to create links between different platforms.

The simultaneous and autonomous mapping of multiple platforms allows tracing networks that are **situated and data-driven**, insofar as they stem from the processing of data that are typical of and made publicly available by the specific platform under observation. These networks are **human-centred**, as they **indicate social media users' behaviours** (even malicious ones). These networks develop within particular platform ontologies and allow researchers to **cope with the heterogeneity** of the current hypermediated online information environment. Importantly, because these networks are traced starting from data that refer to the basic functionalities allowed by each platform under observation (e.g., video posting and commenting in YouTube, message forwarding in Telegram) they are meant to be **replicable over time**. This is the engineering pipeline that we recommend to design the AI4Trust platform on, because it directs users' attention towards specific pieces of content/users/channels.

Highly contextual networks can be analysed according to a common set of tools both with respect to their structure and the contents that flow along their ties. From a structural point of view, **platform-based networks can be:**

a) **Examined at the micro-level in search for particularly prominent contents/users**, which can be further checked to address their active involvement in mis/disinformation flows;

b) **Partitioned** into smaller "areas of interest" that can be further prioritised based on SNA and other statistical measures, and invite **further semantic inspection,** which can be performed with a variety of techniques and tools, from qualitative content analysis and thematic analysis to natural language processing (NLP).

Leveraging publicly platform-available data and processed with techniques that are transparent and explainable, SNA contributes to the overall AI4TRUST endeavour by **offering to the AI4TRUST platform users a set of humanly understandable signals that are also ethics-compliant** which can be an asset to the detection of dis/misinformation items and, therefore, support the process of making judgement and adopt effective strategies to mitigate these phenomena.

Summary Box - SNA as a pre-processing tool for AI4TRUST platform

SNA pre-processing in AI4TRUST

- **Multi-Tiered approach that:**
  - Traces and analyses simultaneously single platform-based networks
  - Links networks based on common pieces of contents/texts
- **Each network is:**
  - Situated and data-driven
  - Human-centred and indicative of a social media user behaviours
  - Replicable over time
- **Processing steps:**
  - Network mapping based on publicly available data
  - Network partition in areas of interest
  - Prioritisation of areas of interest based on SNA measures
  - Semantic analysis of areas with multiple possible techniques

In the following section, we display how we have applied our approach to two platforms: Telegram (section 3.2.1) and YouTube (section 3.2.2).

3.2. Extracting data-driven networks from distinct platform ontologies: from content publication to contextualised user interaction

In this section, we adopt a centrality-based approach to find signals of mis/disinformation presence/circulation and detect super-spreaders. The idea of centrality in human communication was introduced by Bavelas (1948), and applied to social networks by Freeman (1979). The intuition behind measures of structural centrality in social networks is that a few nodes in the graph concentrate the largest share of potential communications. By seeding and targeting these key structural actors, also referred as "influentials" or "central actors" in a network, communication campaigns can indirectly reach a broader audience. By definition, degree centrality (or simply degree) is the number of connections that a node has to others. Taking into account also the weight coefficient of these connections, we can define the strength of a node i is thus as:

$$S(i) = \Sigma_j e_{i,j} w_{i,j} ,$$

where $e_{i,j}$ is a dummy variable for the existence of an edge between nodes i and j, and $w_{i,j}$ is the weight of that edge. In a directed network, we distinguish between in-strength and out-strength, which respectively account for the total incoming and outgoing links for each node. In general, within communication networks like those we are mapping and analysing in this deliverable, higher in-degree scores belong to nodes that receive a lot of attention and/or are targeted by many other nodes in the network through their contents (e.g., as they are mentioned by or receive a reply from other nodes). Conversely, higher out-degree scores point to rather "active" nodes, which initiate communications and exchanges with a higher number of nodes in the same network. Depending on the specific platform under observation, centrality studies are based on simpler strength measures or leverage the in-strength/out-strength distinction. Additionally, always depending on the platforms considered, centrality studies are complemented by different statistical analyses that help refine the precious indications they convey about nodes that receive or send a lot of information/attention. Detailed accounts and justifications of methodological choices made with respect to selected platforms are provided in the subsections below.

Another common feature of the preliminary studies presented below consists of the choice of reducing overall complexity levels that inevitably characterise large-scale networks by identifying dense areas of information/attention sharing. We do so by partitioning larger network structures within "areas of interest" that result from systematic and denser interactions between a portion of nodes. Indeed, actors in a social network tend to organise in groups, also called *communities*, *clusters*, *cohesive subgroups* or *modules* (Tang et al., 2010). Individual actors within each group tend to interact more frequently with one another than with actors outside the group. Identifying these groups is a standard task in social network analysis which is usually referred to as *community detection*. In particular, we will focus on the Louvain community detection method, developed by Blondel et al. from the University of Louvain in 2008. This partition is based on a simple algorithm that extracts non-overlapping community structures from larger networks by iteratively maximising the modularity resulting from assigning a given node to a given community. At the time of publication, this algorithm had been shown to outperform other existing partitioning methods in terms of computing time and accuracy (Blondel et al., 2008).

3.2.1. Telegram

The following subsection summarises the overall approach through which content-flow networks were mapped starting from publicly-available Telegram data and mis/disinformation spreaders which can be detected in this platform. An overview of the different steps that flow into our Telegram Pipeline is provided in the Summary Box below.

Summary Box - Telegram Pipeline

Telegram Social Network Analysis

- **Directed Weighted Network Design:** a directed link from channel A to channel B indicates that at least one message containing a URL was forwarded from channel A to channel B
- **1st signal - Attention levels:** Computation of out-degree (information sent) scores for channels
- **2nd signal - Closeness levels:** Computation of closeness centrality for channels
- **Community detection**
- **Community Ranking:** prioritisation of communities based on their combined attention and closeness levels
- **Community Semantic Inspection**

Data and methods

We used the open source Telethon Python library to conduct a massive data collection from channels suspected to engage in disinformation sharing. In line with the AI4Trust bag-of-words approach to detect potentially contentious content, we selected the keywords related to climate change collaboratively identified by partners[2] to sample a seed of channels containing these words in their title or description. From this seed, we snowballed to connected channels using the "forwarded from" message feature, and at a later stage, combining this with the "similar channels" feature that was added to Telegram as of November 2023[3]. Data collection is still ongoing as this report is being produced[4].

To model collective patterns of information diffusion on Telegram, we built yearly networks where each node is a Telegram channel and a directed edge exists between channel A and channel B if and only if a message containing a hyperlink was forwarded from channel A to channel B[5]. The motivation behind this approach is twofold. Firstly, extensive literature on social network analysis on Telegram effectively uses the "forward" feature to define network links (Bovet et al., Peeters et al., 2022, Zehrig et al., 2023). Secondly, in early stages of analysis, restricting our sample to messages containing hyperlinks presents the opportunity to identify shared domains that are known to present a higher disinformation risk. It is also the opportunity to initiate cross-platform analysis by tracking messages which point to another social media platform.

This report includes preliminary results from a case study on the 2019 network of forwarded hyperlinks, built from a total of 167,905 messages exchanged between 5,304 channels. The resulting network is composed of 5,304 nodes, and 12,314 weighted edges. In this network, we normalise weights by the in-degree of the destination node. Hence, considering an edge of weight

[2] See D2.1 Annex 1

[3] https://telegram.org/blog/similar-channels/fr?setln=en

[4] For a detailed description of Telegram data collection, please refer to section 3.5 in deliverable 2.2. In this report, we use the entire dataset as was available on May 10th, 2024.

[5] Note that if channel B then forwards the same message to a third channel C then the message will appear as forwarded from A to C with no trace of B's intermediation, which significantly limits our knowledge of transmission chains.

w between channel A and channel B, *w* is the share of incoming ties to channel B originating from channel A, as expressed in the following formula:

$$w_{A,B} \ = \frac{e_{A,B}{}^{\square}}{\Sigma_i e_{i,B}{}^{\square}}$$

where $w_{A,B}$ is the weight of the directed edge from A to B, $e_{A,B}{}^{\square}$ (resp. $e_{i,B}$) is the total number of edges from A (resp. i) to B. Note that the denominator is simply the in-degree of node B.

Hereafter, the network described in this paragraph is referred to as the “Telegram network”.

Centrality

Freeman (1979) identifies three key measures of centrality, based respectively on the location of a node at the centre of a network, giving it maximum possible degrees (**degree centrality**), whether it falls in the shortest possible paths between many other nodes (**betweenness centrality**), and its distance to other nodes (**closeness centrality**). We tackle influence in the spreading of online disinformation through the prism of these three measures.

Strength centrality

Firstly, we state that Telegram channels with the highest out-strength centrality are key producers of information in the network of collected channels. Our measure of strength centrality is immediately adapted from Freeman’s definition of degree centrality, with the addition that we take weights into account as described above. In the Telegram network, we choose to focus on out-strength centrality, because channels that most forwards originate from can be assimilated to intermediate sources of information, whereas channels that receive most forwards from others but do not carry on sharing content are dead-ends on the receiving end and do not play a prominent role in the initial stages of the diffusion process.

In the Telegram network, it appears that very few channels concentrate the highest number of outgoing links. Indeed, 60% of Telegram channels only count 1 outgoing link. Although, on the one hand, 8% of channels have no outgoing link at all, on the other hand, the top 10% (n=569) of channels in the distribution of strength centrality count between 5 and 64 outgoing links.

Hereafter, for the sake of interpretability and comparability, we will work with min-max normalised out-strengths, using the following normalisation formula:

$$\delta_N(i) = \frac{S(i) \ - min\, S(G)}{max\, S(G) \ - min\, S(G)}$$

where $\delta_N(i)$ is the min-max normalised strength of node i, $S(i)$ is the raw strength of node i, G is the graph built from the Telegram network, and $S(G)$ is the list of strengths for all nodes in graph G. It is interesting to note that the subgraph of central channels is rather tightly connected, with an average strength of approximately 10, signalling that information is often shared between central channels.

High out-degree centrality will affect the diffusion of information in the Telegram network in many ways. Firstly, central channels are key sources of information outbreak and spread. Centrality is a measure of popularity and influence which might point towards “opinion leaders”: any news that they adopt and post in their discussion thread has the potential to be shared to a large number of other Telegram users who might endorse the same opinion as the “opinion leaders”. This news will reach other Telegram users only after being pre-contextualised by the central spreaders, which might influence their understanding. In the Telegram network, 476 channels are in the top 10th decile of the out-degree centrality distribution (i.e. at least three forwards originated from them).

Furthermore, in-strength can be interpreted as a measure of channel engagement in information sharing on the receiving end. The higher the in-strength, the more messages were forwarded to a given channel, and the more active it is in engaging with information sharing on Telegram. This measure, combined with out-strength centrality, allows to identify active areas in the network, where information circulates between influential hubs of information publishing, and receivers who frequently engage with them. In the Telegram network, 90% of channels are not a recipient of any forwarded message, and only 945 have received at least one forward. This observation indicates that the distribution of the mode of engagement in the Telegram network is skewed towards senders of information rather than receivers.

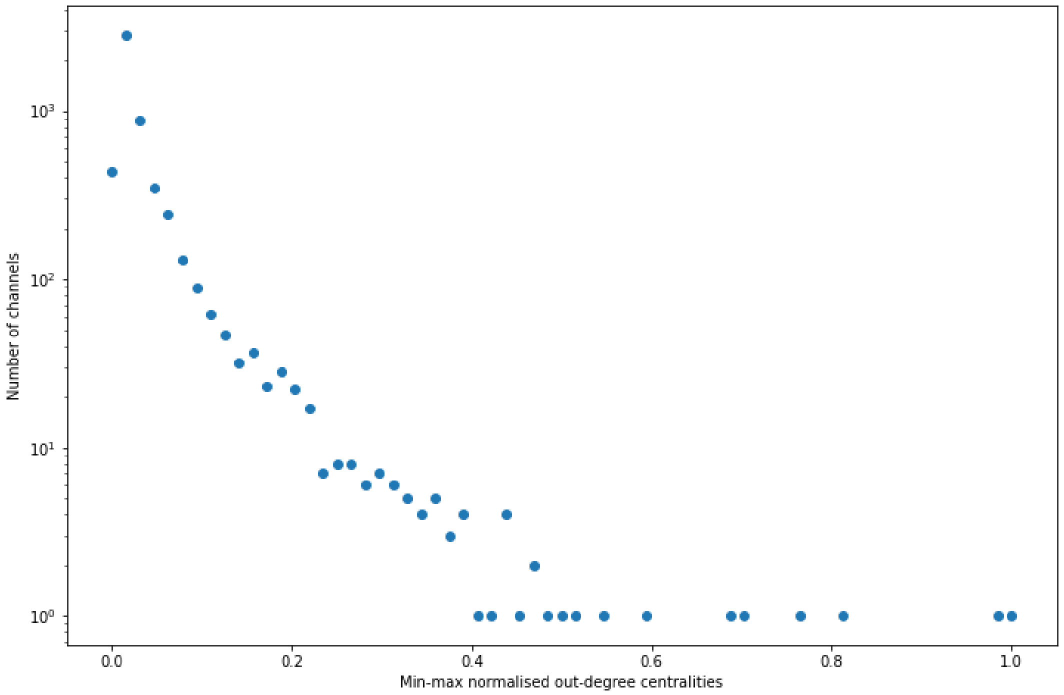
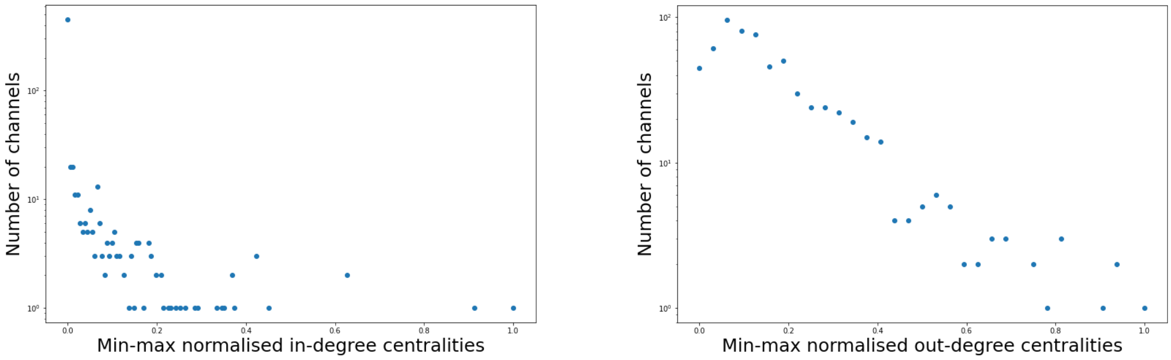


Figure 1: Log-scaled distribution of min-max normalised out-degree centralities in the Telegram network

Closeness centrality

We then investigate which channels have the potential to reach others most directly. Closeness centrality is the average of the shortest path between a given node and any other node in the network. This measure accounts for how close a given node is to others. It does not take into account edge direction and isolated nodes which have no connection to others and would therefore result in infinite distances.

Closeness centrality is very unevenly distributed across the Telegram network. The 10th decile of this distribution is made up of 647 channels with a closeness centrality ranging from 0.19 to 0.25. The degree distributions of these channels are represented in Figure 1 and 2a,b, and they indicate that they are important producers of information rather than receivers.



a. In-degree

b. Out-degree

Figure 2: Log-scaled degree distributions of channels with high closeness centrality

Betweenness centrality

Betweenness centrality measures the intermediary power of a given node to connect two sub-parts of a network that would not otherwise be connected in its absence. This concept is illustrated in Figure 3 on the Telegram network: in the absence of the red node, many paths in this component would be completely cut.

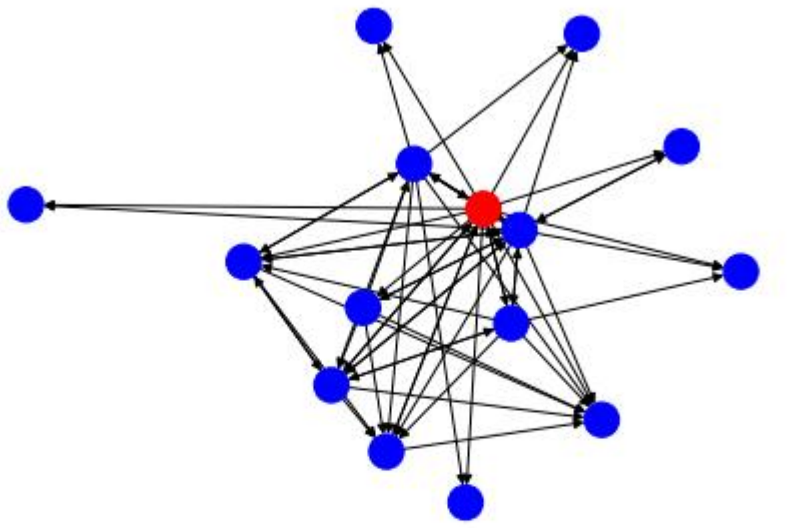


Figure 3: Illustrative example of relatively high closeness centrality in the Telegram network

Due to the structure of our data, measures of betweenness centrality in the Telegram network are extremely close to zero. Indeed, as explained above, publicly available Telegram data does not keep track of all intermediaries in the forward chain, rendering any betweenness analysis irrelevant to the reality of the network.

Channels of interest

Ultimately, we identify spaces of interest in the network as areas of existing interaction between channels with the highest out-strength and closeness centrality (largest issuers of information). This amounts to a total of 948 channels. We expect these groups of actors to be key players in the diffusion of disinformation on Telegram, as they are either central channels from where the information streams down to their subsequent environment, or active relayers of information who consume the information shared by their central counterparts.

Network mapping

Having identified channels of interest as important "individual" (node-level) actors in the Telegram network, we seek to characterise the groups of channels that they more closely interact with.

Connected components

Component subgraphs are a specific type of group in a network that are disconnected from each other. In a directed graph, a connected component is a maximal subset of nodes where there is at least one path between any given pair of nodes in at least one direction (Clark et al., 1991).

The number of (weakly) connected components by size is represented in Figure 4. A single vertex not connected to itself is considered a *trivial* connected component of size 1, and is not included in the plot below. We observe that the largest connected component in the Telegram network contains 4,662 nodes. By definition, in this component, any given pair of channels has known at least one direct or indirect interaction. This suggests that the majority of them share common attributes: it could be that they discuss similar topics, or that the sets of users involved in these channels are intersected. In the absence of a unique identifier for channel members, it is not possible to conduct analysis on the former. Concerning the latter, it would be interesting to carry specific text analysis to evaluate the level of endorsement at the interaction level, to distinguish between links of approval and disapproval.

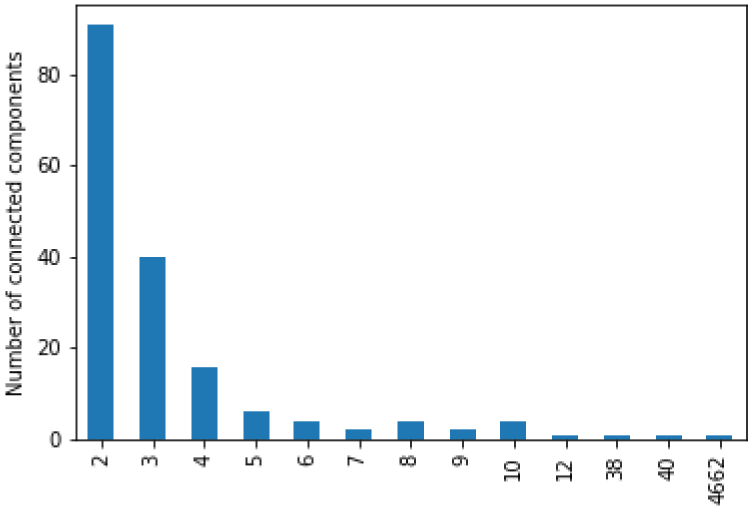


Figure 4: Number of connected components by size in the Telegram network

Communities

To locate the actors of collective elaboration of information on Telegram and characterise their social relations, we partitioned the network into communities using the Louvain method (Blondel et al., 2008). This heuristic method of network partitioning is based on modularity optimisation: at each iteration stage, a given node is added to a community if and only if this addition causes the maximum gain in graph modularity. Thus, nodes are divided into groups on the sole basis of network structure, without any requirements on node attributes. This is particularly convenient for the Telegram network since it does not include any qualitative channel attributes. In addition, the Louvain method is implemented in an existing ready-to-use Python module[6], which makes it easily applicable to our case study.

Using this method, we find around 300 distinct communities. Note that the Louvain method is stochastic, such that two consecutive runs can give slightly different results (Betzel, 2023).

[6] https://github.com/taynaud/python-louvain

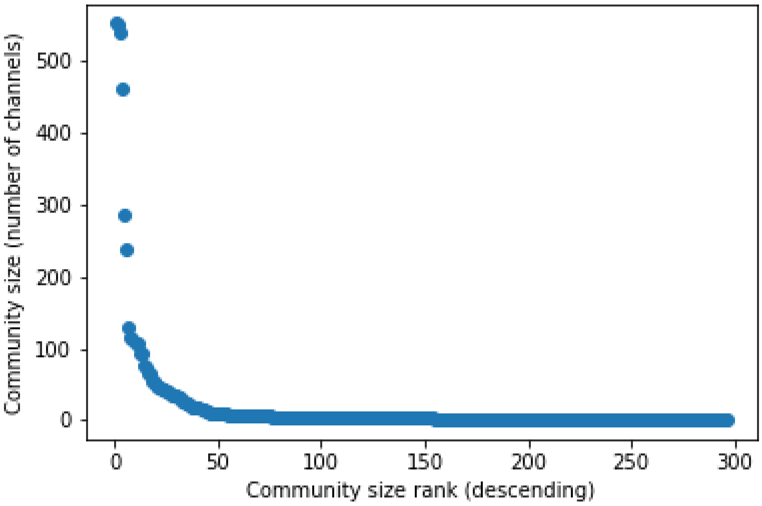


Figure 5: Size of Louvain communities in the Telegram network (by community index)

Consistently throughout different runs of the community detection algorithm, we find that most communities are made up of less than 100 channels (see Figure 5). The largest communities count up to around 550 channels. A visualisation of the largest Louvain communities is represented in Figure 6 below. In this figure, we choose to only draw the largest connected component to make the plot readable, but the communities were identified on the entire network.

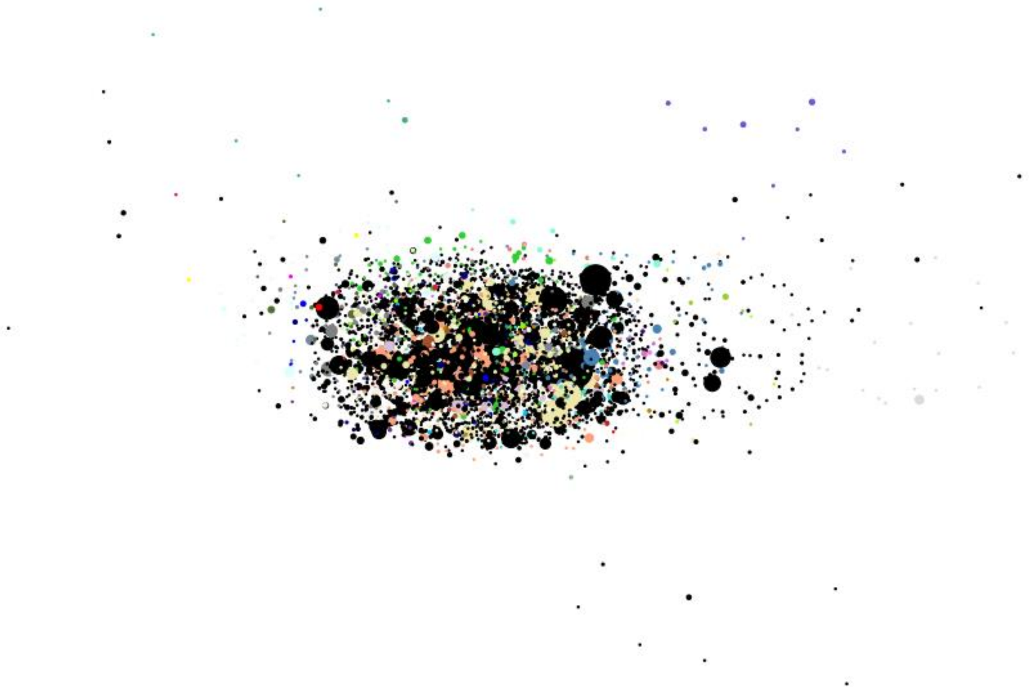


Figure 6: Graphical representation of Louvain communities in the largest connected component in the Telegram network

Note: node size is proportional to node degree, node colour uniquely identifies community attribution for communities with more than 5 nodes. Respectively, nodes that belong to a community of less than 5 nodes are in black.

Communities of interest

Finally, we rank communities according to their relative concentration of channels of interest. We will focus on communities that concentrate at the intersection of the channels in the top 10th decile of the closeness centrality distribution, and the channels in the top 10th decile of the out-strength distribution. This amounts to an intersection of size 948, as described above.

It appears that more than 90% of communities have a zero concentration power of channels of interest. The remaining communities have a concentration power ranging from 0.008 to 0.57, with the higher scores being largely explained by the small size of these communities. We will now zoom closely onto a qualitative analysis of these particular channels.

Topic analysis

The endeavour to qualitatively qualify these communities calls for specific observation of the content they share and a semantic analysis of the contextualisation they make of it. In this section, we present preliminary results of manual topic exploration conducted on the Telegram network at the community level, to get an idea of the main topics in each cluster and qualitatively evaluate thematic coherence in the Louvain partition.

Summary Box - Topic modelling on the Telegram network

Telegram topic modelling

- **Text pre-processing:** we group messages shared within each Louvain community and clean the text
- **Qualitative analysis:** we manually explore messages and produce bag-of-words representations for each community
- **Classification:** we manually attempt to label language and topic for each community

Method

In line with our centrality-based approach, we focus our topic modelling on channels with the highest concentration power of channels of interest. We consider that these communities are important hubs of information sharing, which might need to be further investigated through a sociological lens.

We analyse topics at the community level by first grouping all messages that were shared within a given Louvain community of interest. To do so, we first use the list of all channels that belong to each community, and group the messages that were forwarded from them or to them at least once.

This approach creates potentially intersected corpora, as one message can be forwarded from a community to another.

To clean the text, we first remove URLs, delete empty messages, decode special characters[7], and convert all letters to lowercase. Each community-level corpus is then turned into a "bag of words" as all messages are tokenized at the sentence level using `nltk`'s `word_tokenize`[8]. From this bag of words, we remove stopwords from 10 languages that are within the AI4Trust scope or likely to be found in the data despite not being included in the AI4Trust list: English, French, German, Greek, Italian, Polish, Romanian, Spanish, Russian, Ukrainian[9], and finally remove words longer than 10 characters, to avoid including channels IDs and other noisy elements in the descriptive analysis.

At the current stage of analysis, we manually inspect the bags of sentences for each community to get a pragmatic idea of community interests on the one hand, and text features on the other. In later stages, we expect to conduct quantitative topic modelling using large language models.

Results

After running this pipeline on communities of interest, we find a number of interesting results.

First, the Louvain communities seem to be linguistically homogeneous and usually involve messages written in the same language. Telegram users in this network tend to interact with those who speak the same language as they do online. Linguistic homogeneity structures interactions as people are geared toward peers who share a number of real-life cultural references. We also find that when communities are tightly organised around a given language, they also tend to discuss national politics, as is suggested in table 1 below.

The largest communities of interest do not exclusively discuss climate change or even discuss it at all. Their most commonly spoken languages do not always fall within the AI4Trust scope. Telegram has users on all continents, and research shows that it is particularly popular in countries known for censorship and surveillance outside of Europe. Telegram users find privacy for their encrypted communications, which cannot be traced by state surveillance (Akbari et al., 2019). The topic "climate change" may be one theme among many others discussed in Telegram communities by users interested in broader national or international politics.

We mapped the five largest communities of interest and their associated languages and themes (Table 1). Figure 7 shows the most frequent words used in the Spanish language community of interest.

[7] https://pypi.org/project/Unidecode/

[8] https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html

[9] https://github.com/stopwords-iso/stopwords-iso

Member count	Language (most frequent)	Theme
552	German	Diverse right-wing content
551	Russian	Diverse Russia-centred content
538	Spanish	Diverse Spain-centred political content
462	English	URLs, no specific topic identified in messages body
286	English, Italian	Diverse political content

Table 1: Description of the five largest communities of interest in the Telegram network

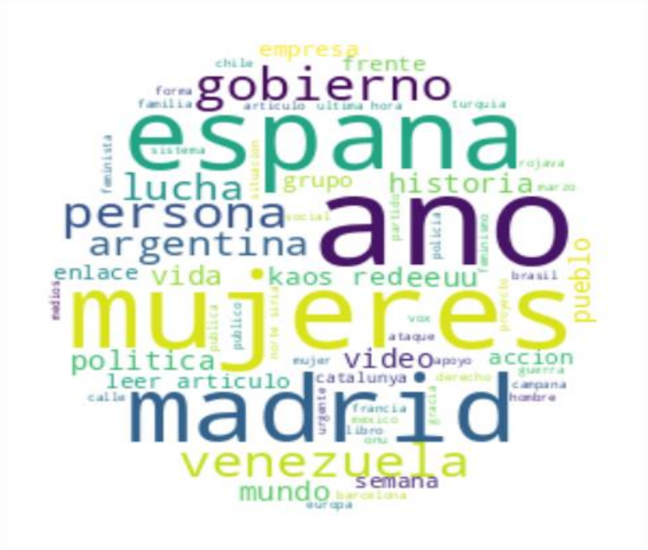


Figure 7: Word cloud visualisation of the vocabulary in one of the largest Spanish-speaking Louvain communities of interest

3.2.2. YouTube

The following subsection summarises the overall approach through which attention-flow networks were mapped starting from publicly-available YouTube data and signals that may point to mis/disinformation contents can be detected in this platform. An overview of the different steps that flow into our YouTube Pipeline is provided in the Summary Box below.

Summary Box -YouTube Pipeline

YouTube Social Network Analysis

- **Bipartite Network Design:** collected videos are linked to the users that comment them
- **Signal Detection:**
    - 1st signal - Attention levels: Calculation of degree scores for videos (attention received) and user (attention given)
    - 2nd signal - Activity concentration: Calculation of Gini index for videos
- **Network partition** in areas of interest via community detection algorithm
- **Community Ranking:** prioritisation of areas of interest based on their combined attention and activity concentration levels
- **Community Semantic Inspection:**
    - Identification of “high rhythm” contents
    - Qualitative semantic analysis

Data and Methods

The YouTube Data API was exploited in order to conduct a massive keywords-based data collection, extracting all videos that contain specific climate change related trigger words identified by partners. Together with information about the videos themselves (e.g., title, description and publication date), comments to each video and relative metadata were also extracted (cfr. D2.2).

In what follows, we present the procedures and the main results of a test analysis for a specific subsample of video extracted through the keyword “climate change hoax”.

The figures below illustrate two sources of variability that characterise the selected dataset. On the one hand, trends of comments differ, sometimes significantly, between videos (figure 8a). On the other, the majority of comments in the dataset were posted between 2019 and 2023 (figure 8b).

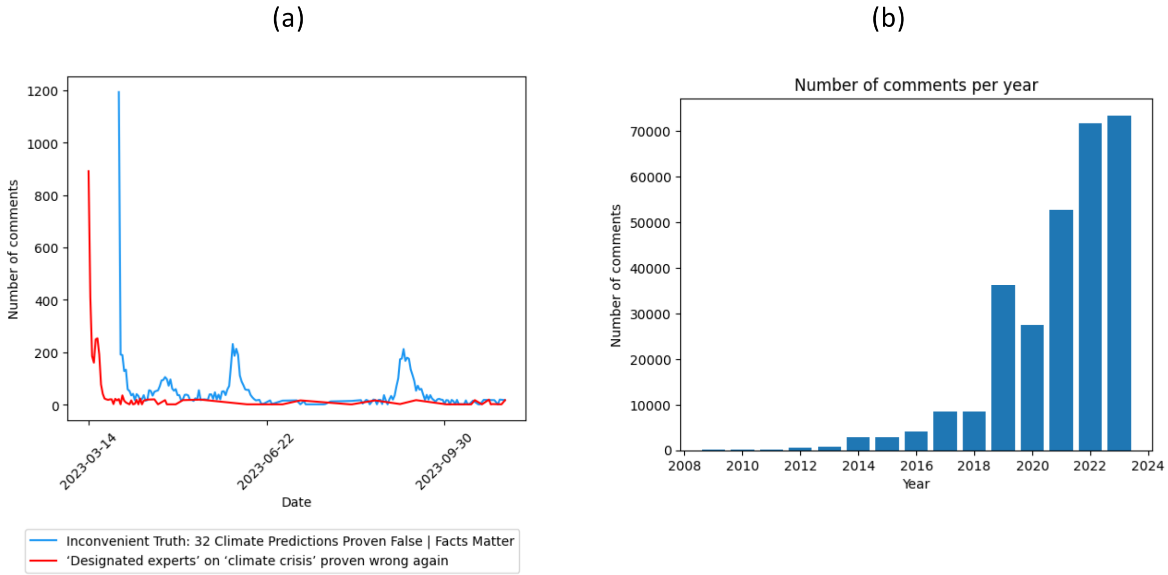


Figure 8: (a) Comments' trends for two random videos in the dataset. (b) Comments distribution over time.

By simply defining edges between videos and users who commented on the video, we built a bipartite network containing 338 videos, 239,332 commenting users and 292,295 edges. We Further excluded from this network all videos with only 1 comment (degree=1). The motivation behind this choice is that videos that attract only one comment in such a long time frame do not configure immediately as troublesome contents, given that they do not trigger on the selected platform any relevant user activity. Through this initial data cleaning we obtain a network that gathers 280 videos, 239,283 commenting users and 292,147 edges. Overall, the network density—that is, the proportion of existing edges on all possible edges - is $1.02 \cdot 10^{-5}$signalling that at least a handful of users commented on more than one video.

On this network, we perform a set of operations in order to identify a set of areas of interest that we then rank so as to prioritise the urgency with which they require further semantic exploration.

Step 1 - Video and user measures

By measuring degree centrality on both layers of the network (i.e., video and users), we ranked videos based on how much attention they gathered (i.e., how many commenting users they attracted). Overall, the average number of comments received by each video (i.e., video average degree) is 1,043.34 with a standard deviation of 2,345.35 (cfr. table 2). In this specific case, high degree videos point to specific pieces of content that elicit increasing levels of public attention. Given the necessarily public nature of phenomena such mis/disinformation, **video degree scores**

therefore could provide a first signal to pay attention to in order to check the presence of malicious content.

Looking at the commenting nodes, we could highlight a general tendency of users to comment mainly on one video (users' average degree score=1.22). Additionally, the average edge weight in the network is 1.5 (SD = 15.04)—further proof of the episodic and targeted nature of video commenting actions.

Metric	Videos	Commenters
Number of Nodes	280	239,283
Minimum Degree	2	1
Maximum Degree	20,916	44
Average Degree	1,043,38	1.22
SD Degree	2,345,35	0.83

Table 2: - Basic video and user measures

Step 2 - Gini Index

While degree could be a relevant indicator of the context for mis/disinformation, there are many reasons why a piece of content triggers public attention. In light of this consideration, we leverage the availability of metadata that refer to the time in which a comment was posted to devise a statistical complement to degree centrality which allows to better specify the modes in which one video triggers public attention.

This complement grounds on a renowned statistical measure of dispersion - i.e., the Gini index. In the context of descriptive statistics, the Gini Index represents a measure of concentration of a given distribution. For a discrete sample x_i ; i=1,...,n, the correspondent Gini Index can be calculated as

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left|x_i - x_j\right|}{2n \sum_{i=1}^{n} x_i}$$

Hence, if a distribution is perfectly evenly distributed ($x_i = x_j \; \forall i$) the Gini index will be equal to 0, while the index takes higher values as the heterogeneity of the distribution increases until reaching a value equal to one if the distribution is fully concentrated on a single element of the sample. In our specific case, we aggregate all comments received by a video depending on the day in which they were posted, and calculate the Gini index to reveal how concentrated the public attention around a video is. The higher the value of the Gini index, the more concentrated the comments of a video are in a particular moment in time. In this sense, higher levels of the index point to the existence of spikes of commenting activities. As mis/disinformation contents tend to spread virally and quickly, a **concentration of comments in a specific moment could provide an additional signal for the presence of malicious content**.

As shown in table 3, videos in the network have an average value of 0,35, with a standard deviation of 0,31. These values reveal that, on the one hand, commenting activities that structure our network tend to be mildly concentrated. However, there is a good amount of variation (indeed, index values in some cases rise up to 0,9).

Gini index	Value
Average	0,35
SD	0,31
Min.	0
Max.	0,9

Table 3: Gini index across the YouTube network - overall descriptives

Step 3 - Community detection and measurements

In order to reduce our network into smaller areas of interest and proceed towards a more in depth analysis, we performed a community detection algorithm on the overall structure under analysis, following the modularity-based Louvain method. The stochastic nature of the chosen model does not allow us to obtain a unique partition of the graph, but through multiple iterations we identified ~90 different communities, with an average modularity index of 0.81, pointing to a fairly clustered structure of the overall network.

As shown in table 4, on average, communities in the network contain 3 videos and more than 2,6k of comments thus representing rather confined areas of activity. The smaller communities gather one video and a couple of comments, but larger communities are far more lively areas, with 27 videos on average and more than 19k comments.

Parameter	Overall	Videos	Comments
Avg. Number of Nodes	2,637.56	3,07	2,629.48
SD Number of Nodes	3,461.21	4,42	3,459.92
Min Number of Nodes	3	1	2
Max Number of Nodes	19,507	27	19,505

Table 4: Communities across the YouTube network - overall descriptives

Each community we identify at this step is assigned two values. On the one hand, a community degree score, consisting of the average value of all videos it contains. This value provides us with an indication of average levels of attention and activity within the boundaries of a certain community. On the other hand, a community Gini index score, consisting of the average value of the Gini index of all the videos it contains. This second value provides us with an indication of how much commenting activities targeting the video(s) in a community are concentrated in a specific moment in time.

Table 5 illustrates the general features of the network we are analysing with respect to these two aspects. As it shows, on average, communities in our network tend to gather videos that receive almost 1.5k comments, which also tend to be posted in a rather limited timeframe (average Gini index=0.31). However, also in this case, there is a good amount of variation: some communities in the network represent livelier areas of activities with up to 13k comments; and not all videos trigger a sudden burst of attention, as in some cases comments are posted quite evenly across time.

Parameter	Community Degree	Community Gini index
Average	1,458.62	0.31
SD	2,544.44	0.26
Min	2	0
Max	13.031	0.78

Table 5: Average degree and Gini index across communities

Step 4 - Joint classification of communities

The final step of our pipeline consists in ranking the communities by crossing the two metrics (Degree and Gini) defined above. Hence, we classify each community i for each metric m as one of three possible categories in the following way:

- Low: if $m_i < avg_m$
- Medium: if $avg_m \leq m_i \leq avg_m + sd_m$
- High: if $avg_m + sd_m \leq m_i$

Where avg_m and sd_m are the average and the standard deviation of the metric m computed at network level (cfr. from Table 1 and Table 2).
The resultant ranking of the communities can be observed in the heatmap below (Figure 9).

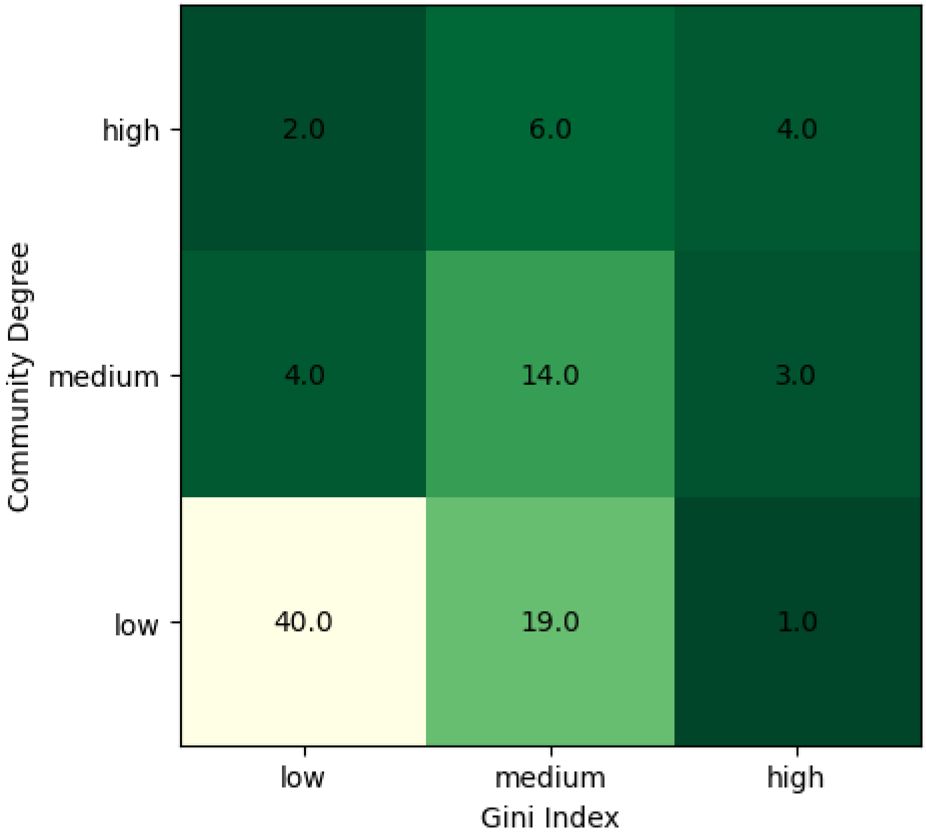


Figure 9: Ranking of network communities according to the joint evaluation of community degree and community Gini index

As displayed in fig. 9, the majority of the communities (40%) fall in the "Low" category in both parameters—i.e., there are videos that receive few comments that also tend to be distributed over time. These communities constitute low interest areas for analysis, as attention levels tend to be scant and quite regular. Conversely, a handful of communities (4), rank "High" on both parameters. These are areas with higher levels of concentrated comments, and analysts should prioritised these areas for semantic inspection. In between low and high interest areas, remaining communities can be prioritised depending on specific interests. A logic that privileges attention-driven mechanisms would prioritise communities that exhibit medium degree scores, while a logic focused on activity-driven mechanisms would prioritise communities with medium Gini index scores.

Communities inspection

Prioritising semantic inspection based on video rhythm

Considering the videos labelled "high-high" that received larger shares of attention in a short amount of time, we tested a mode to further prioritise semantic inspection based on the "rhythm" of each video's reception. We started by evaluating whether or not levels of attention tend to spike in specific moments, which can signal dis/misinformation. We can analyse whether videos experience relevant attention spikes over a short timeframe and, if so, consider them as displaying "irregular rhythms". We trace the rhythm of a video in the following way:

- We define as $n_{i,j}$ the number of comments of the video i on the day j
- We assign to each date within the active lifetime of a video (i.e., the period over which it receives comments) a coefficient $\Delta_{i,j}$ defined as:

$$\Delta_{i,j} = \frac{n_{i,j} - n_{i,j-1}}{\sum_{j=1}^{m} \square\, n_{i,j}}$$

  Generally speaking, this coefficient can range from 0 to 1 and summarises the percentage of comments that a video gains between two consecutive days. For example, if a video counts a total amount of comments equal to 100, passing from 20 to 60 comments from day 9 to day 10, the $\Delta_{i,10}$ coefficient will be:

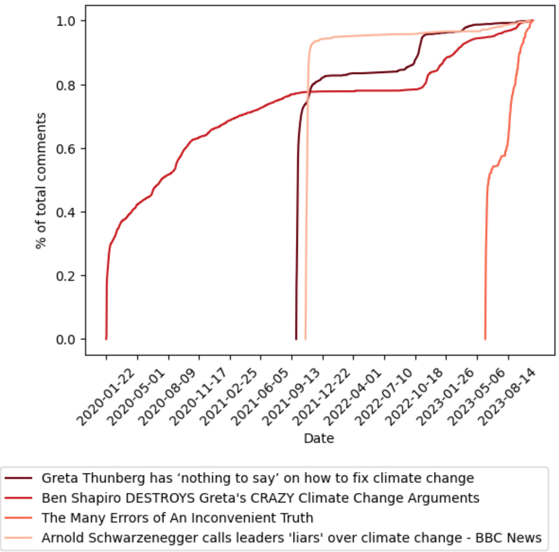
$$\Delta_{i,10} = \frac{n_{i,10} - n_{i,9}}{\sum_{j=1}^{m} \square n_{i,j}} = \frac{60-20}{100} = 0.4\square$$

- Given all $\Delta_{i,j}$ for a specific video, we focus on the $max_j(\Delta_{i,j})$, that is, the maximum increment it experiments during its active lifetime;

- We rank all $max_j(\Delta_{i,j})$ from higher to lower and start inspecting videos based on this ranking. In case two videos display the same values of $max_j(\Delta_{i,j})$, we consider their second higher $\Delta_{i,j}$.

In Figure 10 we present the evolution of the rhythm of the top 4 videos in the communities ranked as "high-high" according to our typology. Colours from darker to lighter (panel a) refer to the corresponding value of $max_j(\Delta_{i,j})$ (panel b).

(a)



(b)

Video	$max_j(\Delta_{i,j})$
Greta Thunberg has ‘nothing to say’…	0.171
Ben Shapiro DESTROYS…	0.16
The Many Errors…	0.156
Arnold Schwarzenegger calls…	0.149

Figure 10: Rhythm evolution for the top-4 videos in the “high-high” communities

Prioritised videos’ inspection

A closer qualitative examination of the above-mentioned four videos which have been prioritised according to our pipeline display some common features. First, all of them were published by largely visible channels (i.e., YouTube channels with a high number of subscribers). Two of them were published by media accounts (Sky Australia and the BBC) and two of them within channels of popular individuals (Ben Shapiro, a prominent US writer, columnist, and opinionist in the US and Simon Clarke, a renown science communicator). Second, they all contain, albeit in different fashions, several problematic pieces of information about the climate crisis. At a very immediate level, this result confirms the need to account for prominent accounts, whether belonging to individual or organisational entities, as potential super-spreaders of problematic information.

At the same time, these four videos reveal that climate change dis/misinformation can be channelled according to different repertoires. The first, well exemplified by the videos by Sky Australia and Ben Shapiro, is an aggressive repertoire that pivots around the smearing of individuals, in this case Greta Thunberg who is, indeed, explicitly targeted as a prominent environmental activism and delegitimised both as a person and with respect to the claim she advances. This emerges in a particularly evident way from results depicted in the upper row of the tag clouds in figure 11. The name of the activist features prominently in both tag clouds. In the cloud on the left, it is accompanied by a “blah blah”—an expression Thunberg used during the World Congress for Climate Justice in Milan in 2021 to criticise the ineffectiveness of institutions in responding to climate change (another prominent word). In this video from 2021, Australian Tv host Andrew Bolt who is known for his conservative and climate-change negationist opinions, used the expression to mock Thunberg, arguing that she is not advancing any practical solution to the problem but is behaving like an annoying child. The second video presents this even more strongly.

In the tag cloud on the right of figure 11, *Greta* is accompanied by other two prominent words, child and people, reflecting a discourse that Thunberg is nothing more than an "angry child" imposing her tantrums on people. Ultimately, in both cases, the sexist and ageist smearing of the activist online serves the purpose of stripping legitimacy from any environmental claim that she makes, and thus advancing climate negationist discourses.

The second repertoire is more subtle. It builds on a proactive approach to climate change as something that exists but tightly intertwined with politics, especially at the institutional level. The first video summarised in the tag cloud on the bottom-left in figure 11 is a scientific dismissal made by Simon Clarke about *An Inconvenient Truth*, a 2003 documentary promoted by Al Gore and later found by a UK court as "inaccurate". Building on the court decision, Clarke dismantles nine errors based on scientific articles and testimonies, eventually arguing that Gore co-opted the topic of climate change in support of his political career. The last video is a three-minutes interview with Arnold Schwarzenegger on how, as a governor of California, he actively engaged against pollution without harming California's economy. These two videos communicate a common idea that the climate crisis is something that can be handled and that climate change is not a pressing issue. Both videos present politicians as using the issue of climate change instrumentally to either avoid thinking about it, or, in Al Gore's case, to sensationalise it to serve a political purpose.

These two repertoires trigger similar patterns of engagement by commenting users (similar $max_j(\Delta_{i,j})$). A slight difference in the rhythm of commenting exists between them, with the aggressive repertoire triggering quicker bursts of interest than the more subtle repertoire that is based on contesting the science.

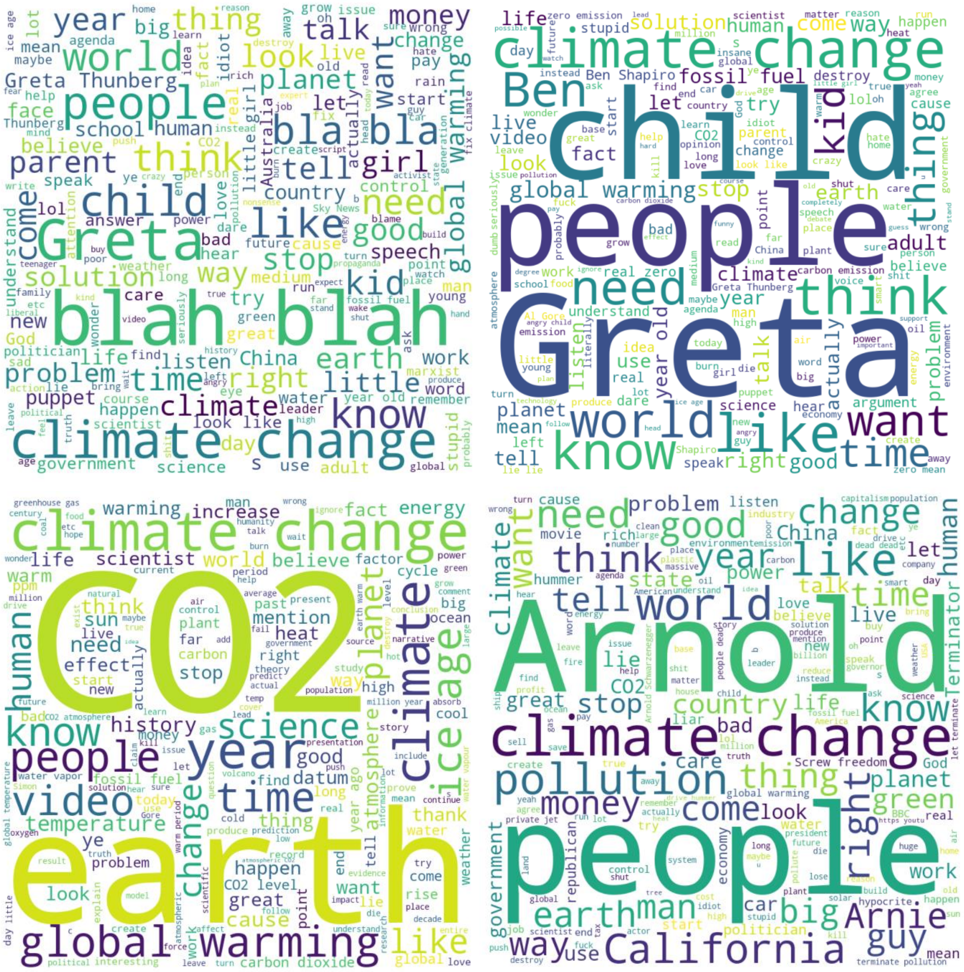


Figure 11: Word cloud visualisation of the vocabulary in the four prioritised videos -Top Left: "Greta Thunberg has nothing to change to fix climate change" (2021); Top Right: "Ben Shapiro DESTROYS Greta CRAZY Climate Change Arguments" (2020); Bottom Left: "The Many Errors of an Inconvenient Truth"(2023); Bottom Right: "Arnold Schwarzenegger called leaders 'liars' over climate change - BBC News" (2021).

3.3. Signalling risk of inappropriateness using insights from network analysis

The following section summarises how social network analysis can contribute to targeting areas in online social networks where inappropriate information is more likely to circulate. Our approach overall is to use quantitative and structural measures of engagement with actors or pieces of content on either platform, and then to consider the combination of these measures to outline areas of dense interaction. We think this makes a good strategy for identifying areas to focus on for curbing the diffusion of online disinformation for two main reasons. First, these dense areas mean that mis/disinformation could potentially reach a broader audience in a reduced amount of time.

Second, high activity in these dense areas could signal the presence and circulation of malicious content in and of itself.

Below we consider that the most straightforward signal of a risk of inappropriateness is the attention given to the unit of analysis on both platforms. Our definition of platform ontology is the set of material aspects of the platforms, and this leads us to slightly different approaches to analysis.

On Telegram we choose to measure this in practice with out-degree centrality. This is immediately linked to the structure of our network. We have for analysis a network of forwarded messages, and then the channels that gather the most attention are those that most messages were forwarded from. This translates as the attention that other channels in the network give to these central actors. We can imagine that relayers of information pay close attention to the feed of central channels. The more content from a central channel is shared, the more we can assume that opinions and topics on these channels are endorsed by a large part of the network, which translates to an indicator of this channel’s influence.

On YouTube we leverage degree centrality, which helps us grasp levels of attention towards a video in terms of the number of commenting users it attracts. Our choice is consistent with YouTube’s platform ontology of a content-based community (Kaplan and Haenlein 2010), where static videos located at a unique URL receive comments from users “travelling” to them. The number of commenting users does not overlap with the number of views a video receives, but it can indicate systematic attention, engagement that goes beyond viewing, and, therefore, of the potential stronger grip of the content on users.

Below we show how combining measures of attention with another measure of importance helps us strengthen our analysis of the risk of inappropriate content, but these measures should vary depending on the platform studied.

On Telegram we stick with a centrality-based approach. We investigated two usual measures of centrality, but one of them, betweenness, was irrelevant to our network structure. The second signal of inappropriateness risk that we consider is closeness centrality. This measure points to key channels in the network which can reach other actors through a limited number of intermediaries. Actors who consume information coming from these channels are at a higher risk of exposure to inappropriate information, because it is less likely that intermediaries control or give nuance to this information. We realise that the use centrality is limited given the platform ontology of Telegram. However, we found relevant closeness scores, and think it warrants an approach to go forward with development.

On YouTube we integrated the study of video centrality with an evaluation of levels of concentration of comments over time. This builds on the assumption that dis/misinformation contents tend to catalyse attention quite quickly. We therefore focus on those videos that received comments from a wider audience (higher degree) and over a short amount of period (higher Gini index). We tested our pipeline on a rather limited dataset so we also added a further measure to prioritise among the target videos. We think the priority targets should be videos that are characterised by more irregular rhythms, i.e., they receive many comments in a short amount of time and comments come in bursts.

4. First version of explainable AI algorithms

4.1. Explaining the social and organisational dynamics of information diffusion

An important contribution to the transparency of results of the AI4Trust platform will be explanations of what can be determined from publicly available social media data, including how connections between actors were established and how network attributes were computed. Existing social network tools can easily be tuned to the AI4TRUST platform to provide users with a better understanding of information flows by (i) **displaying** the constructed networks and interactively visualising the flow of pieces of content along existing paths, **quantitatively synthesising** dynamics of information diffusion by informing on the mathematical calculations behind network measures and drawing, and (ii) visualising the **distribution of topics** across the networks.

Network visualisation

Network visualisation, also known as graph visualisation or link analysis, is the graphical representation of networks whereby actors, or *nodes*, and the links, or *edges*, between them are drawn from a mathematical representation to form an informative plot. These representations can be an edge list, an adjacency list, or matrix representation, among others. The art of network mapping can entail visualising results. However, several display parameters rely on mathematical foundations that are inherent to network science. These parameters can and should be explained to lay users so that they can use these parameters to help them understand how to mitigate online flows of disinformation. Visualisation itself can help explainability. Visualisation of social networks allows end-users to observe social network properties through a cognitively accessible lens, a picture, of the underlying network.

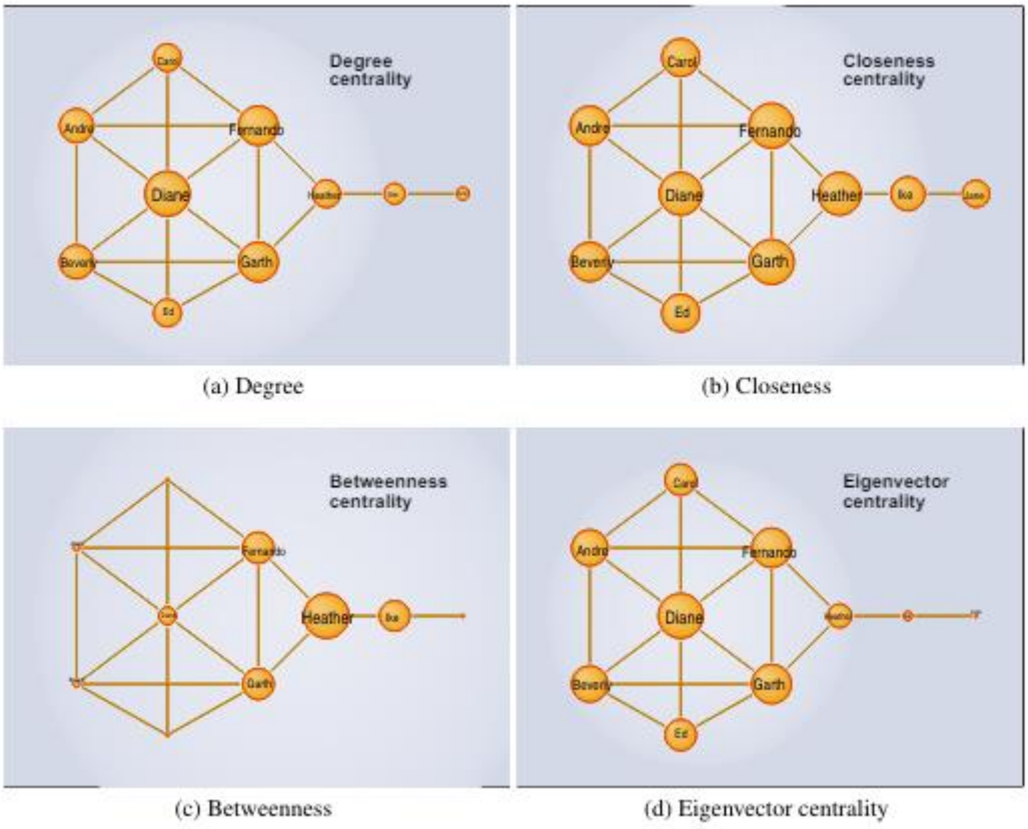
The layout of social network networks is a fundamental choice that determines how nodes' positions in the picture can reveal their social or organisational position. In network visualisations, unlike geographical maps, distance is defined by functional references. This means that visual proximity translates as spheres of influence, potential scopes of action, and contexts of entities, and that are mutually significant (Krempel, 2011). The most popular layout in Python and in other representation tools is the spring layout,[10] which largely relies on force-directed representations. This means that the position of nodes is optimised to minimise the crossing of edges and aim for approximately equal edge length across the network to guarantee esthetically pleasing

[10] https://networkx.org/documentation/stable/reference/generated/networkx.drawing.layout.spring_layout.html

representations. The explainability of these layouts can be provided by straightforwardly exposing the mathematical foundations behind the chosen formula, or by explaining the intuition behind the display, either directly on the platform in an infobox, or inside publications for more complex mathematical theory.

In addition, as we have seen in section 3 above, several visualisation parameters are often adjusted to reflect network measures. Node size might be proportional to node centrality, as illustrated in Figure 12 below, edge width might be proportional to edge strength, and node colour might be positioned on a qualitative scale to reflect existing attributes. These measures can easily be explained by providing the mathematical formula that defines them, which is fairly intuitive, as is the case in section 3 above.

Hence, a large panel of existing tools for static network visualisation that rely on the spatial embeddings, can easily be tuned to map the online informational space and display the spread of given pieces of content in specific networks on specific platforms. We offer to include a network visualisation component to the AI4TRUST platform, whereby the end-user of the platform could specifically visualise the position of groups of actors in the social media network and the proximity between them, as well as the importance/influence of these actors.



(a) Degree

(b) Closeness

(c) Betweenness

(d) Eigenvector centrality

Figure 12: Using node size to display different centralities (Source: Krempel, 2011)

We also plan to use interactive tools to display animated network components that would allow the end-user to trace pieces of content as they travel along paths in the network, going from the source tie through relayers of information (via forwards) and to the last channel that received the information.

In all visualisation cases, the preferred tools are the very popular Gephi[11], or ready-to-use force-directed graph drawing Python tools such as graphviz[12].

Eventually, increased interest in longitudinal social network data has increased the popularity of dynamic network visualisation. This science revolves around understanding how networks develop and change across time. If the consortium were to have access to longitudinal data, it is of interest to explore the use of such dynamic representations to observe how network structures evolve in time with the diffusion of content flows and the variation in attention. We could easily imagine rapid and important evolutions around peaks of scandal or publication of viral contents, and explore if and how these changes stick and affect network structure in the long run.

It is important to note that these methods would mostly be applicable to discussion based platforms such as Telegram, where we can trace transmission chains, but less so on content-based platforms such as YouTube where videos do not flow between users but rather the reciprocal way.

Heatmap projections of topic embeddings

(i) Projecting topics onto users/videos

As we explained in section 3 above, social network analysis for AI4TRUST will include an investigation of topics discussed within networks. To provide the explainable component for this topic analysis, we intend to use the network visualisations produced above. Once we have drawn the networks in a two-dimensional space as described in the previous paragraph, we would colour nodes (i.e. Telegram channels or YouTube videos) according to the most prevalent topics in the textual components of these units (i.e. messages shared on Telegram channels, comments on YouTue videos, etc), thus producing a heatmap-like representation of topics.

(ii) Projecting users/videos onto topics

Alternatively, and quite similarly, we intend to explore projecting units (i.e. Telegram channels or YouTube videos) onto the latent semantic space. To do so, we could proceed as follows:

(1) Map the semantic space of the network using state-of-the-art topic embeddings at the sentence level, such as SentenceBert[13](Reimers et al., 2019)
(2) Project the embeddings onto a 2D space using dimensionality reduction methods such as Uniform Manifold Approximation and Projection (UMAP)[14], thus producing an intelligible representation of the network semantic space

[11] https://gephi.org/

[12] https://pypi.org/project/graphviz/

[13] https://sbert.net/

[14] https://umap-learn.readthedocs.io/en/latest/

(3) Position units of analysis (i.e. Telegram channels or YouTube videos) on this semantic space according to the most prevalent tokens in the textual components of these units.

Thus, the outcome of this pipeline ran on social media data would be a two-dimensional space of topics where units are positioned according to their characteristic topic of interest.

4.2. Explainable audio deepfake detection

4.2.1. Problem statement

Audio deepfake detection is crucial due to the significant risks posed by the misuse of synthetic audio in various aspects of society. Deepfake audio, which uses artificial intelligence to generate highly realistic speech that mimics the voice of any individual, can be employed for malicious purposes such as fraud, identity theft, political misinformation, and reputational damage. For instance, criminals can use audio deepfakes to impersonate executives and authorise fraudulent financial transactions, while malicious actors might spread false statements attributed to public figures, thereby influencing public opinion and undermining trust in institutions. Detecting these fabrications is essential to maintaining the integrity of communication, protecting individuals and organisations from deceit, and preserving the authenticity of information in an era increasingly dominated by digital interactions.

As presented in deliverable D3.1, the audio deepfake detection algorithms we proposed are also AI-based methods. This makes them prone to errors and somehow less trustworthy in the eyes of the end user, i.e. the fact-checker that needs to use these methods to identify AI-based synthetic media. Consequently, offering explanatory mechanisms for the decisions taken by deepfake detection methods significantly improve the users trust in these technologies. In addition, these explanations might allow the user to obtain insights regarding the models or techniques applied to create the audio deepfakes.

In deliverable D3.1 (Section 3.2) we introduced an audio deepfake detection method that estimates the likelihood that an audio is synthetically generated. However, in the case of long audios, such as interviews (which are common in practice), it is difficult to interpret a single score for the entire audio—in these situations it may be that only a short segment is fake. To this end, in this deliverable we investigate how to produce explanations regarding the temporal extent of the manipulations. Concretely, we want to say: "*the given audio is 80% fake because there is a segment from 34.3s to 41.2s that is 90% fake*". The fact checker can then use this information to pinpoint the problem and see if they agree. Moreover, as described in Section 3.2.4 of deliverable D3.1, the output scores (e.g., 80% or 90%) can be interpreted as probabilities and are more calibrated than existing methods.

4.2.2. Related work

Explainability in audio deepfake detection is still relatively unexplored, and the recent survey of Yi et al. (2023) identified the lack of interpretability as one of the main limitations of current models and an important future direction. Most existing work uses explainability techniques developed in the context of general machine learning and computer vision. The idea is to extract the spectrogram of the audio signal and treat it as an image (with frequency on one axis and time on the other); then one can apply standard computer vision techniques. Specifically, Chettri et al. (2018) use Local Interpretable Model-agnostic Explanations (LIME; Ribeiro et al., 2016); Lim et al. (2022) use Integrated Gradients (Sundararajan et al., 2017) and Taylor-based decomposition (Montavon et al., 2017); Ge et al. (2022) use Shapely additive explanations (Lundberg et al., 2017); Müller et al. (2023) use SmoothGrad (Smilkov et al., 2017). By working on the spectrogram these approaches provide explanations regarding which frequency and time locations contribute to the predictions.

A different direction to explainability is to analyse the separability of fake versus real audio in terms of interpretable features. For example, Xue et al. (2022) investigate the fundamental speech frequency F0, while Tak et al. (2020) analyse multiple frequency subbands. Differently, Yadav et al. (2023) attempt to find a disentangled subspace that ensures separability between the two classes, but it still remains unclear whether the resulting dimensions are interpretable to humans.

In terms of producing temporal explanations with dedicated fully supervised localisation models (as we do in the second half of Section 3.2.3), there is an important body of very recent work (Zhang et al., 2022; Zhang et al., 2023; Xie et al., 2024; *inter alia*). For a detailed discussion, please see the *Related work* section in deliverable D3.1 (Section 3.2.2).

4.2.3. Proposed methods

We investigate two methods for producing temporal explanations: (i) an approach that extracts temporal predictions from a pretrained (full audio) detection model, and (ii) a dedicated localisation model. These two methods differ in their training supervision: the first is trained on full utterances (so it uses only weak supervision), while the second is trained on partially manipulated audio (so it uses full supervision).

Weak supervision: Extracting temporal predictions from a detection model. We start from the model described in deliverable D3.1, Section 3.2. This model is trained on full utterances and predicts a single fakeness score for the entire utterance. The model extracts audio features ϕ_t for segments of 20 ms using the wav2vec 2.0 XLS-R 2B feature extractor (Babu et al., 2021). These features are then averaged and used to predict a fakeness score s using a linear classifier w (the logistic regression in our case). Mathematically, we have:

$$s = \langle w, \phi \rangle = \langle w, \frac{1}{T}\sum_{t=1}^{T} \phi_t \rangle$$

To be able to extract temporal scores, we use the linearity of the model. Specifically, we observe that we can “push” the scoring operation (the inner product) inside the averaging operation:

$$s = \frac{1}{T}\sum_{t=1}^{T} \langle w, \phi_t \rangle = \frac{1}{T}\sum_{t=1}^{T} s_t$$

This means that we can decompose the per utterance score s as an average of local per frame scores s_t. These local scores can then be used to provide the fact checkers with temporal explanations.

Full supervision: A dedicated localisation model. As a topline, we consider a model that is explicitly trained to perform localisation. This assumes a fully supervised model that has access to partially manipulated audios and their labels (that is, a list of binary labels, one for each window in the audio). As our localisation models, we used the architectures described in deliverable D3.1, Section 5.2. Specifically, we consider a linear model (which has the same architecture as the one used for weak supervision, but is trained at segment level) and a nonlinear convolutional model (which also extracts wav2vec 2.0 XLS-R 2B features, but uses a more flexible decoder: four convolutional layers interspersed with ReLU activation functions). Note that in deliverable D3.1 Section 5.2 apart from these two architectures we have also experimented with gated multilayer perceptrons (Liu et al., 2021) and transformers (Vasawani et al., 2017), but these performed worse than the convolutional model.

4.2.4. Experimental results

To quantify the quality of our temporal explanations we evaluate our predictions in a localisation setting. Specifically, we evaluate on the PartialSpoof dataset (Zhang et al., 2022), which contains audio with two types of local manipulations: fake segments and splicing points. The annotations come in different resolutions; we use a fine-grained (20 ms) and a coarse (160 ms) resolution. The methods produce scores for each window in an audio, and we evaluate their performance using the equal error rate metric (the lower the better).

4.2.4.1. Quantitative results

	supervision	network	train		test window (ms)	
			dataset	window (ms)	20	160
1	weak	linear	ASVspoof'19	utterance	35.0	30.7
2	weak	linear	PartialSpoof	utterance	42.0	42.7
3	full	linear	PartialSpoof	160	22.3	12.8
4	full	linear	PartialSpoof	20	14.8	28.3
5	full	conv	PartialSpoof	20	**8.1**	**6.0**

Table 6: Equal error rates (%) on the PartialSpoof database for the temporal explanations produced by either the two types of models considered: (i) the weakly supervised approach that extracts temporal predictions from a detection model (first two rows), and (ii) the fully supervised approach trained on partially manipulated audios and their labels (last three rows).

Table 6 shows the results for the two types of methods introduced in Section 3.2.3 and their variants. For the weakly supervised model, we consider two variants: one trained on the ASVspoof'19 dataset (row 1) and one trained on the PartialSpoof dataset (row 2). We observe that the performance of both variants is low in absolute terms: over 30% EER. The model trained on PartialSpoof is worse than its ASVspoof counterpart, presumably because the annotations are noisier: a fake label for an utterance in PartialSpoof refers to only a partially manipulated audio.

For the fully supervised models, we have three variants that differ in terms of (i) the decoder network (either linear or convolutional), and (ii) the temporal resolution at train time (either 20 ms or 160 ms). We see that the linear variants trained with full supervision (rows 3–4) perform much better than their weakly-trained counterparts (rows 1–2). This suggests that full supervision is needed to produce accurate temporal explanations. Moreover, the presence of a strong nonlinear decoder (row 5) further reduces the error to 8.1% and 6.0% EER, respectively.

4.2.4.2. Qualitative results

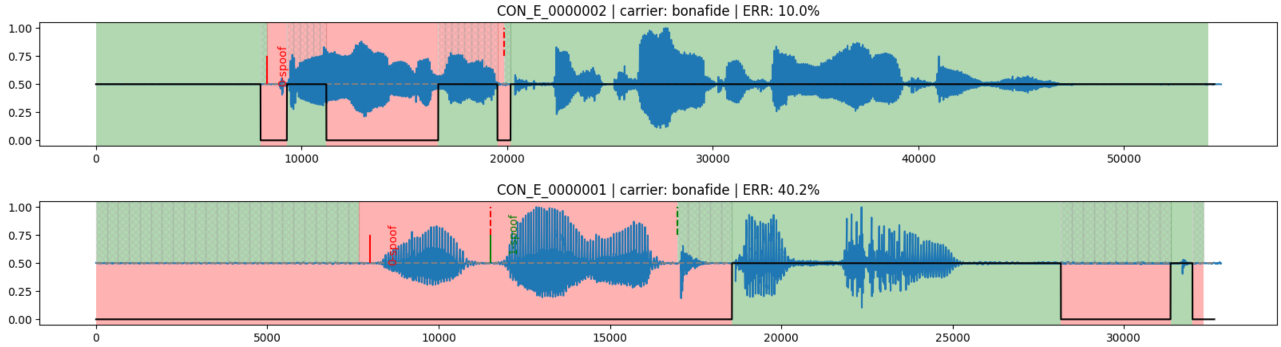


Figure 13: Temporal explanations produced by the fully supervised model for two partially manipulated audios from the PartialSpoof database. Red denotes fake, green denotes real signal. The regions below 0.5 are the groundtruth; the regions above 0.5 are the predictions. Grey hatch areas denote prediction mistakes.

Finally, we provide visualisations of the temporal explanations produced by our method (the fully supervised approach) in Figure 13. For the first example in the figure, the model accurately identifies the fake audio segments around timestamps 9s, 11–17s and 20s, but erroneously extends these areas to include the bonafide audio around timestamps 10s and 19s. Consequently, the file overall error is 10%.

In the second example the error is much higher, 40%, but it is mainly generated because silence labelled as fake in the test dataset is actually predicted by the model as being real. This happens for the audio segments 0–7.5s, around 18s and around 30s.

In conclusion, we observe that the predictions are generally accurate and provide a good interpretation of the results. Even when the results are quantitatively poor (second example, with an error of 40.2%), we notice that the performance is affected by unrelated factors: artefacts of the database, as silence is sometimes labelled as fake.

4.2.5. Conclusions and future work

We have presented a methodology for augmenting binary (fake vs. real) classification scores with temporal information regarding the location of the fake segments. While the first approach of extracting the temporal explanation from a pretrained full audio detection model was not accurate enough, the counterpart of relying on a dedicated fully supervised method yielded promising predictions.

This is only one facet of additional information that can be presented to the fact checker. We plan to enrich the outputs by also considering spectrogram explanations (which regions in the frequency domain triggered a decision?) and model attribution (which text-to-speech or voice conversion systems were used to produce synthetic audio?). All this information can potentially be presented as natural language by running it through a large language model.

4.3. Explainable deepfake video detection

4.3.1. Problem statement

The recent advances in the field of Generative AI have led to new and more sophisticated ways of image and video manipulation, and the creation of a new type of visual disinformation that is often referred to as deepfakes. Deepfakes are AI manipulated media in which a person's face or body is digitally altered in an existing image or video to make them appear as someone else or to reenact them. The ongoing improvement of Generative AI technologies enables the creation of deepfakes that are increasingly difficult to detect. The latter observation, combined with the use of deepfakes for spreading disinformation, necessitates the development of effective solutions for deepfake detection. Moreover, enhancing deepfake detection methods with explanatory mechanisms would significantly improve the users' trust in these technologies and allow obtaining insights about the applied image/video manipulation procedures for creating the detected deepfake.

Despite the growing interest in building increasingly more powerful models for deepfake detection, the provision of trustworthy explanations for the output of these models has not been studied extensively. Most works on explainable deepfake detection, investigate the use of various methods that create visual explanations (usually in the form of 2D heatmaps), but evaluate the performance of methods based only on the basis of qualitative analysis over a limited set of examples (Aghasanli et al., 2023; Jayakumar et al., 2022; Malolan et al., 2020; Silva et al., 2022; Xu et al., 2022). Only a recent work has attempted to assess the performance of various explanation methods on two CNN-based deepfake detection models using a quantitative evaluation framework (Gowrisankar et al., 2024). Nevertheless, their proposed framework uses explanations produced from correctly classified pristine (non-manipulated) images, in order to compare the performance of various explanation approaches. In contrast, we argue that the opposite use-case of explanations - i.e., when the model detects a deepfake - is both more meaningful and useful to the user. Moreover, their framework requires access to pairs of real-fake images, thus being non-applicable on datasets that contain only fake examples, e.g., the WildDeepfake dataset (Zi et al., 2020).

In AI4TRUST, we designed a new evaluation framework that is simpler and more widely-applicable than the one in (Gowrisankar et al., 2024). This framework takes into account the produced visual explanation for the deepfake detector's decision after correctly classifying a fake image, without requiring any access to its original counterpart. Based on this new framework, we evaluated the performance of five explanation methods from the literature on a state-of-the-art model for deepfake detection. The conducted work was reported in (Tsigos et al., 2024) that has been accepted for publication at the 3rd ACM International Workshop on Multimedia AI against Disinformation (MAD '24), which will be made in conjunction with the ACM International Conference on Multimedia Retrieval (ICMR '24).

4.3.2. Related work

Over the last years, there is an increasing interest in the development and training of advanced network architectures for deepfake detection. However, the explanation of the decisions of these networks has been poorly investigated. In an early work, Malolan et al. (2020), trained a variant of the XceptionNet (Chollet, 2017) using a subset of the FaceForensics++ dataset (Rossler et al., 2019) and examined the use of the LIME (Ribeiro et al., 2016) and LRP (Bach et al., 2015) methods for producing visual explanations about the outcomes of the trained model. However, the evaluation of these methods was based on a few samples and mainly focused on the robustness of the produced explanations against various affine transformations or Gaussian blurring of the input image. Xu et al. (2022), utilised the representations of EfficientNet-B0 (Tan et al., 2019) and a supervised contrastive learning methodology to train a linear deepfake detector to discriminate the real from the manipulated images of the FaceForensics++ dataset (Rossler et al., 2019). In terms of explainability, Xu et al. investigated the use of the learned features only for explaining the observed detection performance, using heatmap visualisations and uniform manifold approximation and projection (UMAP). Silva et al. (2022), proposed the use of an ensemble of CNNs (XceptionNet (Chollet, 2017), EfficientNet-B3 (Tan et al., 2019)) and attention-based models for deepfake detection. They provided explanations about the regions of the images that influence the most the decision of the detector, using the Grad-CAM method (Selvaraju et al., 2017) and focusing on the computed gradients for the attention map. Nevertheless, the produced explanations were evaluated only in a qualitative manner by taking into account only a few image samples. Jayakumar et al. (2022), trained a deepfake detection model that utilises the EfficientNet-B0 (Tan et al., 2019) as backbone and contains five dense classification layers. To produce visual explanations, they investigated the use of the Anchors (Ribeiro et al., 2018) and LIME (Ribeiro et al., 2016) methods, and conducted evaluations based on a limited set of examples. Aghasanli et al. (2023), described a deepfake detection model that relies on Vision Transformers and can be used for distinguishing original and fake images generated by various diffusion models. For explaining the model's output, Aghasanli et al. used SVM and xDNN (Angelov et al., 2020) classifiers to understand the model's behaviour by analysing the closest support vectors and prototypes for each classifier, respectively. The evaluation of the produced explanations, though, was based on the qualitative analysis of a few samples. Haq et al. (2023) described a neurosymbolic deepfake detection method that is based on the idea that deepfakes exhibit inter- or intra- modality inconsistencies in the emotional expressions of the person being manipulated. Their method performs inter- and intra- modality reasoning on emotions extracted from audio and visual modalities using a psychological and arousal valence model for deepfake detection, and provides textual explanations that localise the timestamp and identify the fake part. However, it was evaluated only in terms of deepfake and emotion detection, while its explainability dimension was discussed only theoretically. Finally, Gowrisankar et al. (2024) described an evaluation framework for explanation methods, which is based on the intuition that the identified salient visual concepts by such a method after correctly classifying a real image as a non-manipulated one, could be used to flip the prediction of the detector for its fake counterpart. Initially, Gowrisankar et al., investigated the appropriateness of generic data removal/insertion approaches for modifying the spotted salient pixels or segments of the input image (e.g., zeroing, replacement with a uniform random value and blurring based on the neighbouring pixels), and found out that these approaches may produce less meaningful results when applied on deepfake detection models, as they can distort facial regions and produce completely unexpected detection results (e.g., increase of deepfake detection accuracy). Based on this finding, Gowrisankar et al., described a framework which applies a number of adversarial attacks (using Natural Evolution Strategies (NES) (Wierstra et al., 2014) in regions of a fake image that correspond to the identified salient visual concepts after explaining the (correct) classification of its real counterpart, and evaluates the performance of an explanation method based on the observed drop in the accuracy of the deepfake detector. Thus, their evaluation framework takes the unusual step of using the produced explanation after correctly classifying a real (non-manipulated) image, in order to assess the capacity of an explanation method to explain the detection of a fake (manipulated) image.

4.3.3. Comparative study setup

Deepfake detection model

We used a model that relies on the second version of the EfficientNet architecture (Tan et al., 2021) for deepfake detection. Building on the first version of EfficientNet - which leveraged Inverted Bottleneck convolutions (MBConv) and compound scaling to achieve high performance with fewer parameters compared to models with similar ImageNet accuracy (Tan et al., 2019) - the employed version introduced Fused Inverted Bottleneck convolutions (Fused-MBConv), leading to even faster training and improved efficiency (Tan et al., 2021). We chose EfficientNet due to its widespread adoption, efficiency, and effectiveness in deepfake detection tasks, either as a part of an ensemble or as a backbone of more advanced methods (Shiohara et al., 2022; Zhao et al., 2021). Notably, an ensemble of five EfficientNet-B7 models achieved the winning performance in Meta's DFDC challenge (Dolhansky et al., 2019). Moreover, EfficientNet has been shown to outperform alternative CNN architectures, such as XceptionNet (Chollet, 2017) and MesoNet (Afchar et al., 2018) (that were taken into account in (Gowrisankar et al., 2024)), on various deepfake datasets (Huang et al. 2023; Li et al., 2023; Nadimpalli et al., 2023). Finally, it achieves similar performance to other vanilla CNNs on the ForgeryNet dataset (He et al., 2021) while requiring fewer parameters.

Explanation methods

We produce visual explanations by highlighting the regions of the image (or video frame) with the biggest influence on the deepfake detection model's decision. In our study, we considered the following explanation methods:

- Grad-CAM++ (Chattopadhay et al., 2018), is a back-propagation-based method that generates visual explanations by leveraging the information flow (gradients) during the back-propagation process.
- RISE (Petsiuk et al., 2018), is a perturbation-based method that produces visual explanations by randomly masking out portions of the input image and assessing their impact on the model's output.
- SHAP (Lundberg et al., 2017), is an attribution-based method that leverages the Shapley values from game theory. It constructs an additive feature attribution model that attributes an effect to each input feature and sums the effects, i.e., SHAP values, as a local approximation of the output.
- LIME (Ribeiro et al., 2016), is a perturbation-based method that creates visual explanations by randomly masking out portions of an input image to assess their impact on the model's output.
- SOBOL (Fel et al., 2021), is an attribution-based method that employs a mathematical concept called Sobol' indices, to identify the contribution of the input variables on the variance of the model's output.

Evaluation framework and measures

Based on the reported findings in (Gowrisankar et al., 2024), about the adequacy of generic data removal/insertion approaches for perturbing the input image, we also do not apply such approaches on the image regions that have been promoted by an explanation method, in order to assess this method's performance. We evaluate the performance of an explanation method by extending the evaluation framework in (Gowrisankar et al., 2024) so that it takes into account the produced explanations for fake images. We argue that the provision of an explanation after detecting a fake image is more meaningful for the user, as it can give clues about regions of the image (the highlighted ones by the visual explanation) that were found to be manipulated. On the contrary, the provided explanation after classifying an image as "real" would demarcate specific regions of the image as non-manipulated (see Figure 14), while someone would expect that the entire image has not been manipulated at all.

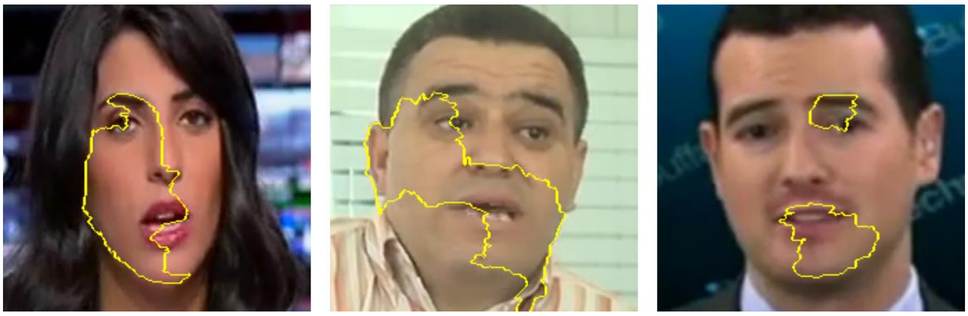


Figure 14: The produced explanations by the LIME method (the best performing one according to our evaluations), for three non-manipulated images of the FaceForensics++ dataset, that were correctly classified as "real".

Let us assume a fake image and the produced visual explanation for the deepfake detector's decision, by an explanation method (see the orange coloured part of Figure 15). We assess the performance of this method by examining the extent to which the indicated regions in the visual explanation as the most important ones, can be used to flip the deepfake detector's decision (and thus classify the image as "real"). For this, we segment the input image into super-pixel segments using the SLIC algorithm (Achanta et al., 2012). Then, we quantify the contribution of each segment to the deepfake detector's decision by overlaying the created visual explanation to the segmented image and averaging the scores of the explanation for the pixels of the segment - as a note, in the case of LIME (Ribeiro et al., 2016) we pass the SLIC-based segmentation mask of the input image and get the top-k scoring segments directly. Following, we focus on the top-k scoring segments and apply NES to progressively generate a variant of the input image that is classified as "real" by the deepfake detector. This iterative adversarial image generation and evaluation process, that is illustrated in the green coloured area of Figure 3.3.1, stops if the deepfake detector classifies the adversarial image as "real" or a max number of iterations is reached.

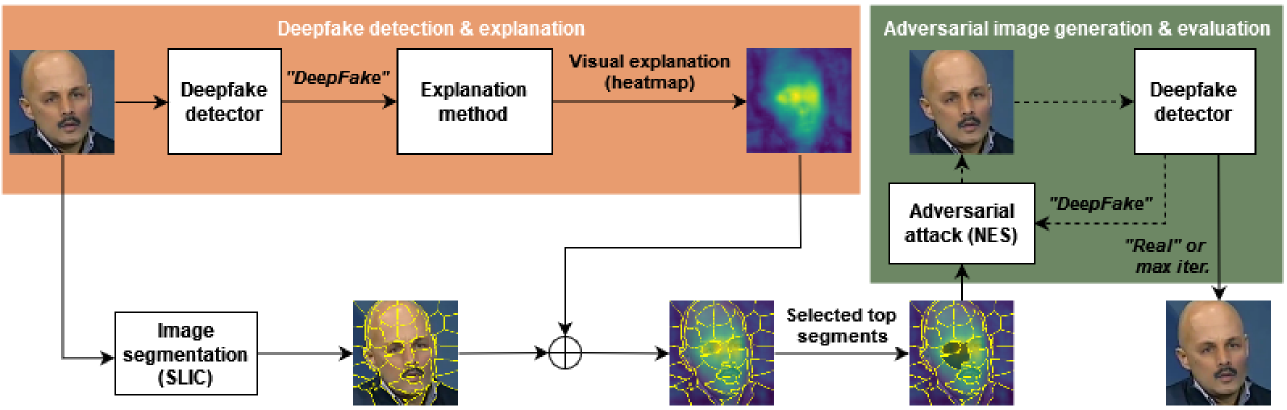


Figure 15: The proposed framework for evaluating the performance of explanation methods.

To quantify the performance of an explanation method, we calculate the accuracy of the deepfake detection model on the set of returned adversarial images after the completion of the adversarial image generation and evaluation process, when the adversarial attacks target the top-1, top-2 and top-3 scoring segments of the input images by the method. This measure ranges in [0, 1], where the upper boundary denotes a 100% detection accuracy. We anticipate a larger decrease in accuracy for explanation methods that spot the most influential regions of the input image for the deepfake detector's decision, more effectively. Complementary to the aforementioned measure, we quantify the sufficiency of explanation methods to spot the most influential image regions for the deepfake detector, by calculating also the difference in the detector's output after applying adversarial attacks to the top-1, top-2 and top-3 scoring segments (following the paradigm in Liu et al., 2022). This measure ranges in [0, 1], where low/high sufficiency scores indicate that the top-k scoring segments by the explanation method have low/high impact to the deepfake detector's decision, and thus the produced visual explanation exhibits low/high sufficiency.

4.3.4. Experiments

Dataset and implementation details

Our experiments were conducted on the FaceForensics++ dataset (Rossler et al., 2019). This dataset contains 1000 original videos and 4000 fake videos created using one of the following four classes of AI-based manipulation (1000 videos per class): “FaceSwap” (FS), “DeepFakes” (DF), “Face2Face” (F2F), and “NeuralTextures” (NT). The videos of the FS class were created via a graphics-based approach that transfers the face region from a source video to a target video. The videos of the DF class were produced using autoencoders to replace a face in a target sequence by a face in a source video or image collection. The videos of the F2F class were obtained by a facial reenactment system that transfers the expressions of a source video to a target video while maintaining the identity of the target person. The videos of the NT class were generated by modifying the facial expressions corresponding to the mouth region, using a patch-based GAN-loss as utilised in Pix2Pix (Isola et al., 2017). The dataset is divided into training, validation, and test sets, comprising 720, 140 and 140 videos, respectively.

For deepfake detection, we sampled the videos keeping 1 frame per second and used the RetinaFace face detector (Deng et al., 2020) to obtain bounding boxes for the present faces. Following suggestions in (Rosslet et al., 2019), we enlarged each bounding box by a factor of 1.3 to capture any relevant background information that might aid in discriminating between real and fake samples. The cropped faces were stored and used as input to train and test the deepfake detector. For training, we leveraged a pre-trained model on the ImageNet 1K dataset obtained from the timm library[15]. Then, the deepfake detection model was trained for 30 epochs using the AdamW optimizer (Loshchilov et al., 2019) with a learning rate of 5×10^{-5} and a weight decay of 1×10^{-1}, and the Cross-Entropy loss for multiclass classification. To mitigate overfitting and improve generalisation, we employed the following data augmentation techniques: Random Erasing, Random Resized Crop, and AugMix (Hendrycks et al., 2020). Additionally, to improve robustness to unseen data and encourage the model to learn more reliable features, we incorporated Stochastic Depth (Huang et al., 2016) with a drop path rate of 4×10^{-1}. As a result, there was a 40% chance of dropping a residual block connection during each forward pass.

To obtain the data for evaluating the different explanation methods, following (Gowrisankar et al., 2024), we used 127 videos from each different class of the test set and we sampled 10 frames per video, thus creating four sets of 1270 images. The generation of visual explanations was based on the following settings:

- For Grad-CAM++, we took the average of all convolutional 2D layers.
- For RISE, we set the number of masks equal to 4000 and kept all the other parameters with their default values.
- For SHAP, we set the number of evaluations equal to 2000 and used a blurring mask with kernel size equal to 128.

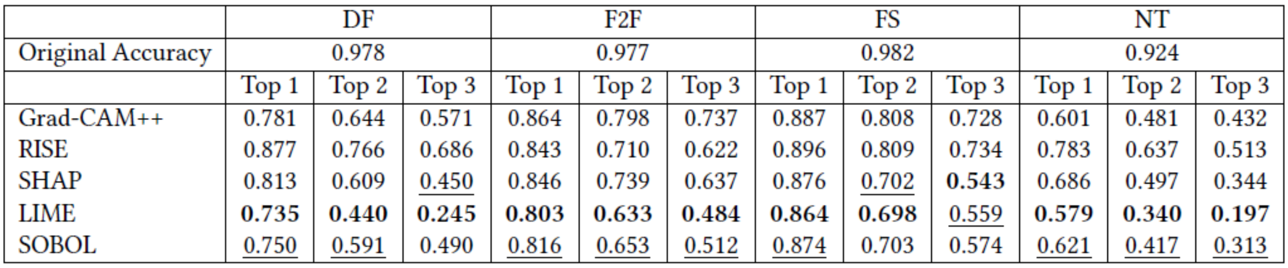
[15] https://github.com/huggingface/pytorch-image-models

- For LIME, we set the number of perturbations equal to 2000 and used the SLIC algorithm with a target number of segments equal to 50.
- For SOBOL, we set the grid size equal to 8 and the number of design equal to 32, and kept all the other parameters with their default values.

With respect to NES, we set: the number of maximum iterations equal to 50, the learning rate equal to 1/255, the maximum distortion equal to 16/255, the search variance equal to 0.001, and the number of samples equal to 40.

Quantitative results

Table 3.1.1 reports the accuracy of the employed deepfake detection model for the different types of fakes in the FaceForensics++ dataset, on the original set of images (second row) and the adversarially-generated variants of them after modifying the image regions corresponding to the top-1, top-2 and top-3 scoring segments according to the different explanation methods. As shown in this table, the used deepfake detection model exhibits very high performance on all types of fakes of this dataset (achieving approx. 98% accuracy on DF, F2F and FS and over 92% on NT), documenting its state-of-the-art performance. With respect to the considered explanation methods, LIME appears to be the most effective one, as it is associated with the largest decrease in the detection accuracy for all types of fakes and in almost all experimental settings. As expected, the observed accuracy decrease is smaller when the adversarial image is generated based on the top-1 scoring segment and significantly larger when the adversarial attack is performed on the top-2 and top-3 scoring segments. However, this decrease is even more pronounced in the case of LIME. Therefore, LIME appears to be more effective compared to the other methods at highlighting the most influential segment of the input image for the decisions of the used deepfake detector, and noticeably better at spotting the top-2 or top-3 image segments with the highest impact on the detector's decision. Concerning the remaining methods, SOBOL seems to be the most competitive in most cases, while SHAP shows good performance in the case of DF and FS samples when spotting the top-2 or top-3 regions of the image. Finally, a comparison of the reported results across the different types of fakes, reveals that the different explanation methods can more effectively explain the detection of DF and NT classes, while the explanation of fakes from the remaining two classes is a more challenging task.



	DF			F2F			FS			NT		
Original Accuracy	0.978			0.977			0.982			0.924		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
Grad-CAM++	0.781	0.644	0.571	0.864	0.798	0.737	0.887	0.808	0.728	0.601	0.481	0.432
RISE	0.877	0.766	0.686	0.843	0.710	0.622	0.896	0.809	0.734	0.783	0.637	0.513
SHAP	0.813	0.609	0.450	0.846	0.739	0.637	0.876	0.702	**0.543**	0.686	0.497	0.344
LIME	**0.735**	**0.440**	**0.245**	**0.803**	**0.633**	**0.484**	**0.864**	**0.698**	0.559	**0.579**	**0.340**	**0.197**
SOBOL	0.750	0.591	0.490	0.816	0.653	0.512	0.874	0.703	0.574	0.621	0.417	0.313

Table 7: The accuracy of the employed deepfake detection model for the different types of fakes in the FaceForensics++ dataset, on the original set of images (second row) and the adversarially-generated variants of them after modifying the image regions corresponding to

the top-1, top-2 and top-3 scoring segments based on the different explanation methods. Best scores in bold and second best scores underlined.

Table 7 presents the sufficiency scores of the considered explanation methods for the different types of fakes in the FaceForensics++ dataset, after performing adversarial attacks at the top-1, top-2 and top-3 scoring segments of the input images. These scores seem to be aligned with the results in Table 7, demonstrating once again that LIME performs consistently good for all the considered types of fakes and numbers of top-scoring segments. Moreover, its effectiveness in spotting the most influential regions of the images is more pronounced when taking into account the top-3 scoring segments according to the produced visual explanation. As before, SOBOL is the second best method and SHAP performs comparatively good on specific occasions. Finally, the most challenging cases in terms of visual explanation, still remain the ones associated with fakes of the F2F and FS classes.

	DF			F2F			FS			NT		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
Grad-CAM++	0.148	0.253	0.310	0.069	0.115	0.162	0.063	0.113	0.160	0.194	0.251	0.280
RISE	0.087	0.162	0.219	0.091	0.173	0.223	0.060	0.114	0.157	0.115	0.204	0.273
SHAP	0.137	0.300	0.402	0.092	0.158	0.222	0.073	0.181	**0.269**	0.167	0.282	0.357
LIME	**0.195**	**0.408**	**0.539**	**0.121**	**0.238**	**0.334**	**0.087**	**0.189**	0.262	**0.233**	**0.363**	**0.431**
SOBOL	0.166	0.277	0.352	0.108	0.212	0.296	0.078	0.180	0.259	0.198	0.302	0.362

Table 8: The sufficiency scores of the considered explanation methods for the different types of fakes in the FaceForensics++ dataset, after modifying the top-1, top-2 and top-3 scoring segments of the input images. Best scores in bold and second best scores underlined.

Qualitative results

The top row of Figure 16 shows four different images (sampled video frames) of the FaceForensics++ dataset and the next row contains their AI manipulated variants, where each variant is associated with a different type of manipulation. The remaining rows present the produced visual explanations by the examined methods. As illustrated in these rows, LIME successfully spots: i) the regions close to the eyes and mouth that have been modified in the case of the DF sample, ii) the regions around the nose and the cheeks that have been changed in the case of the F2F sample, iii) the regions close to the left eye and cheek that have been altered in the case of the FS sample, and iv) the regions close to the mouth and chin that have been manipulated in the case of the NT sample. With respect to the other explanation methods, Grad-CAM++ correctly focuses on regions close to the eyes in the DF and FS samples and close to the chin in the case of the NT sample. However, it fails to clearly indicate regions in the case of the F2F sample and to spot manipulations around the mouth in the case of the DF sample. RISE seems to produce explanations that highlight irrelevant (see the F2F and FS samples) or non-manipulated regions (see the NT sample) of the image, while also failing to spot the manipulated ones (see the DF sample). Finally, SHAP and SOBOL appear to perform well compared with LIME, as in most cases, they provide explanations that indicate the altered regions of the images. This finding is aligned with the performance of these methods according to the conducted quantitative evaluation.

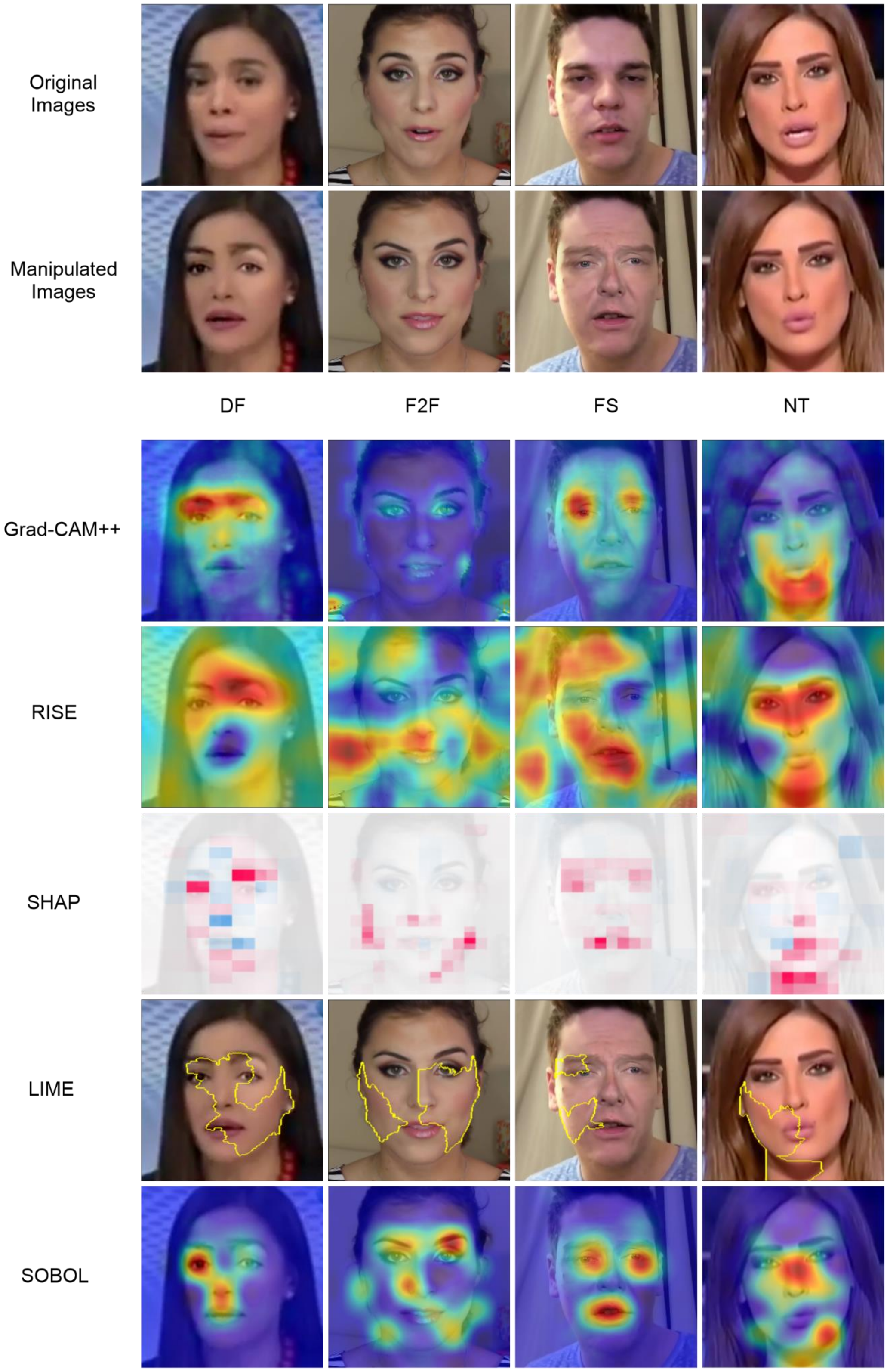


Figure 16: The obtained visual explanations from the considered explanation methods for four different images of the FaceForensics++ dataset (one per different type of manipulation). In terms of visualisation, we adopt the default supported format by each explanation method.

4.3.5. Conclusions and future steps

In this section, we presented the designed evaluation framework for explainable AI methods for deepfake detection, which measures the capacity of such methods to spot the most influential regions of the input image through an adversarial image generation and evaluation process that aims to flip the detector's decision. We applied this framework on a state-of-the-art model for deepfake detection and five explanation methods from the literature. Our experimental results demonstrated the competitive performance of the LIME explanation method across all different types of fakes, and its competency to produce meaningful explanations for the employed deepfake detection model. In the future, we will extend our study by taking into account: i) attention-based deepfake detectors, ii) attention-based explanation methods, and iii) additional evaluation measures that quantify the stability and consistency of explanations.

5. Mis- and Dis-information Mitigation Strategies

5.1. Beyond Commercial Content Moderation

Building on the frame of reference laid out in previous deliverables, all mitigating measures against the diffusion of information distortions should take stock of the conceptual distinction established between misinformation and disinformation (see deliverables 2.1. and 4.1), Intentionality is the differentiation factor between misinformation and disinformation. For this reason, the latter is much more complex to identify and counter than the former. While discrete units of content (e.g. post, video, image, audio track) can be readily qualified as instances of misinformation, **we can only make claims about disinformation operations in the presence of contextual indicators**, such as the concentrated diffusion of messages in dense network clusters. In this case, the identification of dissemination networks and the analysis of its structural components allow us to make informed judgements concerning the behaviour of certain individual and collective actors. The dissemination at scale of a narrative qualified by our tools as false, attributable to one actor or set of actors, can be taken as a proxy for intent. For example, our initial analyses of some YouTube networks suggest that, applying a statistical measure of dispersion, a concentration of comments over a given period could ascribe context to the spread of malicious content (see p. 36).

Notwithstanding the promising preliminary results of the AI tools to detect text, audio and video misinformation as well as the social network analysis methods to map the contextual dynamics of mis/disinformation, **mitigation measures cannot be equally deployed across all online platforms**

and services. Besides the different nature of the information distortions, we must also account for the variety of ontologies, functions, online (sub)cultures and social dynamics of social media platforms (SMPs). For that reason, it is important to cover and assure continuous access to data from players qualified as Very Large Online Platforms (VLOPs), such as YouTube, and to other online platforms that, regardless of size and official qualification by the applicable legislation (cf. Digital Services Act), are demonstrably relevant communication services across Europe, such as Telegram. It is only through the understanding of different online ontologies that we can start making sense of the common offline and online media ecology where SMPs, big and small, conflate and narratives tend to converge.

Until the emergence of multiple-stakeholder projects like AI4Trust, efforts to identify and counter the dissemination of online mis/disinformation were concentrated on the social media platforms themselves. Demands for accountability and efforts for regulation have typically been hampered by **asymmetric information about the circulation of online mis/information**. As much as researchers and journalists scuttered to gather scattered evidence of the nexus between online mis/disinformation and social harms, from Whatsapp fuelled lynchings in India (Samuels, 2020) to the Rohingya genocide instigated on Facebook (Mozur, 2018), the barriers to data access and the lack of comprehensive, cross-platform, multimodal analytical toolkits prevent us from having a clear picture about the societal impact of information disorders. Without systematic and systemic analyses it is difficult to demonstrate that the relationship between false information and social harms is causal rather than merely casual. Against this backdrop of information asymmetry, democracies face difficulties to make the public good prevail over the private interests of SMPs. The inscrutability of the social media black boxes, be it the synthetic or the organic dissemination of information and mis/disinformation, transform efforts of regulation into exercises of self-regulation, The emphasis put on the need to improve “content moderation” by SMPs is also an admission that we cannot overcome the current state of SMP operational opacity. According to this logic, it is incumbent on online platforms to moderate their spaces through the application of in-house mitigation measures to public data.

A non-exhaustive list of mitigation measures deployed by SMPs includes ex ante filtering by which certain categories of content are censored prior to the publication, and ex post content moderation, such as labelling, demotion, reduction (Gillespie, 2022) or even “shadow banning” (Cotter, 2023). Lastly, SMPs have been increasingly adopting suppression measures, which can take the form of single content/user removal (e.g. deplatforming) or large-scale takedowns motivated by SMP-side identification of coordinated inauthentic behaviour (CIB). The latter are typically self-reported in exercises of “corporate transparency” (Nimmo et al., 2022) that do little to elucidate the opacity of the processes and algorithms governing content circulation and moderation. Behind and beyond this unwillingness of SMPs to comply with the general guidelines of explainable and trustworthy AI, there is also lack of transparency about the human dimension of content moderation. For example, commercial content moderators, frequently sub-contracted workers in the Global South, are subject to significant psychological (Steiger et al., 2021) and labour challenges (Roberts, 2016). SMP-side mitigation thus materialises as the result of machine-driven and human-centred processes in which the machines are opaque and the humans are rendered invisible.

5.2. Hybrid Platform Mitigation

As demonstrated along this document, our approach is radically different. The **AI4Trust project proposes a hybrid platform** based on an interaction between humans and machines, **privileging human-centred AI technologies and research methods**. As a technological platform, AI4Trust tries to address the identified problems resulting from 1) information asymmetry, 2) partial accounts of the societal impact of mis/disinformation, 3) algorithmic opacity and 4) human invisibility. In this sense, the hybrid platform is a monitoring system that responds to the global call for a “comprehensive misinformation research agenda” (Watts et al., 2021) by **proposing not just to map but also to counter the information disorders**.

First, the AI4Trust platform **rebalances the current state of information asymmetry through the principle and practice of platform observability**, which “seeks to address the conditions, means, and processes of knowledge production about large-scale socio-technical systems” (Rieder & Hofmann, 2020, p. 4). The effectiveness of the platform relies on access to data streams comprising both VLOPs (e.g YouTube) and smaller yet relevant online platforms (e.g, Telegram) used by millions of people to circulate information and spread mis/disinformation. The success of the platform and its mitigation strategies is highly dependent on the access to corporate-controlled public data without which platform observability cannot be implemented. Going forward, **data access is a determinant of the viability of a public-side monitoring system** that is neither owned nor directly influenced by private online platforms. Denying or interrupting access to SMP data streams is a significant risk factor for the project.

Second, the platform aims to provide a global, multi-language, multi-modal and cross-platform perspective over the diffusion of mis/disinformation. Rather than dispersing the information collected about information disorders, the platform is an information aggregator geared towards professional end-users, who frequently act as information mediators and/or gatekeepers (e.g. journalists, fact-checkers, policy makers). Consequently, it **enlarges the spectrum of mitigation beyond content moderation by VLOPs into a range of contextual and content-based measures mediated by the hybrid platform.**

Third, the proposed mitigation directly addresses AI and algorithmic opacity as well as human invisibility by placing project researchers, fact-checkers and end-users at the centre of an explainable and trustworthy pipeline of AI tools and research methods. Peeking into and picking apart the informational black boxes of SMPs, the AI4Trust platform is trialling the systematic mapping and countering of mis/disinformation. The AI4Trust platform observes and identifies the content and contexts of mis/disinformation and explains the social makeup of mis/disinformation diffusion. This report shows that the AI4Trust platform rests on explainable and understandable AI tools and scientific methods made by humans and for humans, from audio and video manipulation to social network analysis.

5.3. Preliminary matching of platform outputs and mitigation inputs

Projecting the horizon of application, this deliverable trialled some of the AI tools and scientific methods that will be deployed by the AI4Trust platform and made available to its end-users. Concerning platform observability and representativity, the **currently accessible data streams, although not comprehensive, do represent some of the heterogeneity we encounter across online platforms**, In this regard, YouTube data are being used to account for the dynamics of mis/disnformation in VLOPs, whereas data from Telegram allow us to map diffusion dynamics in instant messaging platforms, a type of online platform that has been frequently identified as an important focus of mis/disinformation diffusion (Cavalini et al., 2023). As for the audio and video manipulation tools, the project is using state-of-the-art training and benchmarking datasets (e.g. PartialSpoof and FaceForensics++).

5.4. Future mitigation strategies

Future mitigation strategies are underpinned by the identification of false and inappropriate content and contexts, be it manipulated audio, video deepfakes or mis/disinformation signals from social networks. A range of tools in our platform will enable the detection of claim validity, disinformation signals, video and audio anomalies, etc, which will then be matched with a set of **mitigation inputs grounded in the provision of statistical and analytical information about mis/disinformation content and contexts.** Regardless of the scale of analysis, the mitigation pipeline follows the following process:

a) Detection of mis/disinformation signals (content and/or context);

b) Analysis of the detected signals (e.g. spread, virality);

c) Classification of the identified misinformation/disinformation according to their severity level, from illegal content and non-illegal individual harms to more systemic risks, such as threats against groups, fundamental rights, public health, and democratic forms of participation;

d) Hybrid recommendation tool. Based on the above elements, the platform can issue semi-automated recommendation reports tailored for specific categories of end-users, who then act upon the evidence-based recommendation according to their own professional procedures.

The matching of platform outputs with mitigation inputs is at the heart of the hybrid recommendation tool. Fact-checking provides a good use-case illustration. Section 3. confirms social network analysis can provide valuable insights about the social dynamics of information distortions, With SNA we can assess the virality of mis/disinformation (see section 3) and the structural composition of the social networks where those narratives spread. As the analysis shows, central actors as well as measures of engagement in denser areas of the analysed relational structures are considered relevant indicators due to their reach and propagation potential. The aggregated output of the identified disinformation signals and risks of inappropriateness at the individual and whole-network levels (see 3.1.) enables the classification of those signals and risks according to their severity level. The identification of a coordinated disinformation campaign targeting immunisation efforts against viral diseases would be qualified as a systemic risk because it represents a threat against public health. The statistical and analytical output can then inform the hybrid recommendation tool according to the end-user. In our use-case example, fact-checkers can be presented with a report detailing the mitigation pipeline, including indicators of virality and severity level, that assigns priority to the false story/narrative and to consequent need to debunk it. This capacity to distinguish signal from noise could prove an important tool to policymakers, journalists and fact-checkers alike thus representing an invaluable resource to mitigate the most severe mis/disinformation narratives.

The work in this report shows how the AI4Trust platform can develop algorithmic tools for the detection and mitigation of misinformation and disinformation that conform to ethical principles, consider socio-technical contextual information and advance the state-of-the-art in trustworthy and explainable AI.

6. References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 11 (2012), 2274–2282.

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: a Compact Facial Video Forgery Detection Network. In 2018 IEEE International Workshop on Information Forensics and Security, WIFS 2018, Hong Kong, China, December 11-13, 2018. IEEE, 1–7.

Aghasanli, A., Kangin, D., & Angelov, P. (2023). Interpretable-through-prototypes deepfake detection for diffusion models. In Proceedings of the 2023 IEEE/CVF Int. Conf. on Computer Vision Workshops (ICCVW). IEEE Computer Society, Los Alamitos, CA, USA, 467–474

Akbari, A., & Gabdulhakov, R. (2019). Platform surveillance and resistance in Iran and Russia: The case of Telegram. Surveillance and Society, 17(1/2).

Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). Neural Networks 130 (2020), 185–194.

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In International Conference on Learning Representations.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., ... & Auli, M. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE 10, 7 (07 2015), 1–46.

Bavelas, A. (1948). A mathematical model for group structures. Human organization, 7(3), 16-30.

Betzel, R. F. (2023). Community detection in network neuroscience. In Connectome Analysis (pp. 149-171). Academic Press.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.

Bovet, A., & Grindrod, P. (2022). Organization and evolution of the UK far-right network on Telegram. Applied Network Science, 7(1), 76.

Cavalini, A., Malini, F., Gouveia, F., & Comarela, G. (2023). Politics and disinformation: Analyzing the use of Telegram's information disorder network in Brazil for political mobilization. *First Monday*. https://firstmonday.org/ojs/index.php/fm/article/view/12901

CCDH. 2024. "Elon Musk vs. Center for Countering Digital Hate: Nonprofit wins dismissal of 'baseless and intimidatory' lawsuit brought by world's richest man." CCDH Blog, 25 March, 2024. https://counterhate.com/blog/elon-musk-vs-ccdh-nonprofit-wins-dismissal-of-baseless-and-intimidatory-lawsuit/.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 839–847.

Chettri, B., Mishra, S., Sturm, B. L., & Benetos, E. (2018). Analysing the predictions of a CNN-based replay spoofing detection system. In *IEEE Spoken Language Technology Workshop (SLT)* (pp. 92-97).

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Los Alamitos, CA, USA, 1800–1807.

Clark, J., & Holton, D. A. (1991). A first look at graph theory. World Scientific.

Cotter, K. (2023). "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, *26*(6), 1226–1243. https://doi.org/10.1080/1369118X.2021.1994624

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 5202–5211.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Canton-Ferrer, C. (2019). The Deepfake Detection Challenge (DFDC) Preview Dataset. CoRR abs/1910.08854 (2019).

Fel, T., Cadene, R., Chalvidal, M., Cord, M., Vigouroux, D., & Serre, T. (2021). Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. In Advances in Neural Information Processing Systems (NeurIPS).

Freeman, L.C., 1979. Centrality in social networks: Conceptual clarification. Social Networks, 1 (3), 215-239.

Ge, W., Patino, J., Todisco, M., & Evans, N. (2022). Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6387-6391).

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, *8*(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Gowrisankar, B., & Thing, V. L. L. (2024). An adversarial attack approach for eXplainable AI evaluation on deepfake detection models. Computers & Security 139 (2024), 103684.

Haq, I. U., Malik, K. M., & Muhammad, K. (2023). Multimodal Neurosymbolic Approach for Explainable Deepfake Detection. ACM Trans. Multimedia Comput. Commun. Appl. (sep 2023).

He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., & Liu, Z. (2021). ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 4360–4369.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2020). AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In 8th Int. Conf. on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.

High-Level Expert Group on Artificial Intelligence (AI HLEG). 2019. Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., & Ye, D. (2023). Implicit Identity Driven Deepfake Face Swapping Detection. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. IEEE, 4490–4499.

Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K. Q. (2016). Deep Networks with Stochastic Depth. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016. Springer, 646–661.

Isola, P., Zhu, J., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Los Alamitos, CA, USA, 5967–5976.

Jayakumar, K., & Skandhakumar, N. (2022). A Visually Interpretable Forensic Deepfake Detection Tool Using Anchors. In Proceedings of the 2022 7th Int. Conf. on Information Technology Research (ICITR). 1–6.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. “The Global Landscape of AI Ethics Guidelines.” Nature Machine Intelligence 1, pp. 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Kak, Amba and Sarah Myers West. 2023. “AI Now 2023 Landscape: Confronting Tech Power.” AI Now Institute, April 11, 2023. https://ainowinstitute.org/2023-landscape.

Krempel, L. (2011). Network visualization. The SAGE handbook of social network analysis, 558-577.

Li, X., Ni, R., Yang, P., Fu, Z., & Zhao, Y. (2023). Artifacts-Disentangled Adversarial Learning for Deepfake Detection. IEEE Trans. Circuits Syst. Video Technol. 33, 4 (2023), 1658–1670.

Lim, S. Y., Chae, D. K., & Lee, S. C. (2022). Detecting deepfake voice using explainable deep learning techniques. *Applied Sciences*, *12*(8), 3926.

Liu, H., Dai, Z., So, D., & Le, Q. V. (2021). Pay attention to mlps. Advances in neural information processing systems, 34, 9204-9215.

Liu, Y., Li, H., Guo, Y., Kong, C., Li, J., & Wang, S. (2022). Rethinking Attention-Model Explainability through Faithfulness Violation Test. In Proc. of the 39th Int. Conf. on Machine Learning (Proceedings of Machine Learning Research, Vol. 162). PMLR, 13807–13824.

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. In 7th Int. Conf. on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proc. of the 31st Int. Conf. on Neural Information Processing Systems (Long Beach, California, USA) (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

Malolan, B., Parekh, A., & Kazi. F. (2020). Explainable Deep-Fake Detection Using Visual Interpretability Methods. In Proceedings of the 2020 3rd Int. Conf. on Information and Computer Technologies (ICICT). 289–293.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern recognition*, *65*, 211-222.

Mozur, P. (2018, October 15). A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times*. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Müller, N. M., Sperl, P., & Böttinger, K. (2023). Complex-valued neural networks for voice anti-spoofing. *Interspeech*.

Nadimpalli, A. V., & Rattani, A. (2023). Facial Forgery-based Deepfake Detection using Fine-Grained Features. CoRR abs/2310.07028 (2023).

Nimmo, B., Agronovich, D., & Gleicher, N. (2022). Adversarial Threat Report. *Meta, April*. https://pressbooks.usnh.edu/com743/files/2022/04/Meta-Adversarial-Threat-Report-2023.pdf

Nuutila, E., & Soisalon-Soininen, E. (1994). On finding the strongly connected components in a directed graph. Information processing letters, 49(1), 9-14.

Peeters, S., & Willaert, T. (2022). Telegram and digital methods: Mapping networked conspiracy theories through platform affordances. M/C Journal, 25(1).

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. CoRR abs/1806.07421 (2018). arXiv:1806.07421

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (San Francisco, California, USA) (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: high precision model-agnostic explanations. In Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (New Orleans, Louisiana, USA) (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 187, 9 pages.

Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, *9*(4). https://policyreview.info/articles/analysis/towards-platform-observability

Roberts, S. T. (2016). *Commercial content moderation: Digital laborers' dirty work*. https://ir.lib.uwo.ca/commpub/12/?utm_source

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. In 2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV). IEEE Computer Society, Los Alamitos, CA, USA, 1–11.

Samuels, E. (2020, February 21). Analysis | How misinformation on WhatsApp led to a mob killing in India. *Washington Post*. https://www.washingtonpost.com/politics/2020/02/21/how-misinformation-whatsapp-led-deathly-mob-lynching-india/

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In 2017 IEEE Int. Conf. on Computer Vision (ICCV). 618–626.

Shiohara, K., & Yamasaki, T. (2022). Detecting Deepfakes with Self-Blended Images. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 18699–18708.

Silva, S. H., Bethany, M., Votto, A. M., Scarff, I. H., Beebe, N., & Najafirad, P. (2022). Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. Forensic Science International: Synergy 4 (2022), 100217.

Singh Yadav, A. K., Bhagtani, K., Xiang, Z., Bestagini, P., Tubaro, S., & Delp, E. J. (2023). DSVAE: Interpretable Disentangled Representation for Synthetic Speech Detection. *arXiv e-prints*, arXiv-2304.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., & Lease, M. (2021). The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445092

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning* (pp. 3319-3328).

Tak, H., Patino, J., Nautsch, A., Evans, N., & Todisco, M. (2020). An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification. *Odyssey*.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proc. of the 36th Int. Conf. on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97). PMLR, 6105–6114.

Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller Models and Faster Training. In Proc. of the 38th Int. Conf. on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139), Marina Meila and Tong Zhang (Eds.). PMLR, 10096–10106

Tang, L., & Liu, H. (2022). Community detection and mining in social media. Springer Nature.

Tsigos, K., Apostolidis, E., Baxevanakis, S., Papadopoulos, S., Mezaris, V. (2024). Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection. 3rd ACM Int. Workshop on Multimedia AI against Disinformation (MAD '24) at the ACM Int. Conf. on Multimedia Retrieval (ICMR'24), Phuket, Thailand, June 10-13, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J., & Schmidhuber, J. (2014). Natural Evolution Strategies. Journal of Machine Learning Research 15, 27 (2014), 949–980.

Xie, Y., Cheng, H., Wang, Y., & Ye, L. (2024). An Efficient Temporary Deepfake Location Approach Based Embeddings for Partially Spoofed Audio Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 966-970).

Xu, Y., Raja, K., & Pedersen, M. (2022). Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection. Proceedings of the 2022 IEEE/CVF Winter Conference on

Applications of Computer Vision Workshops (WACVW). 379–389.

Xue, J., Fan, C., Lv, Z., Tao, J., Yi, J., Zheng, C., ... & Shao, S. (2022). Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia* (pp. 19-26).

Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023). Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*.

Zehring, M., & Domahidi, E. (2023). German corona protest mobilizers on Telegram and their relations to the far right: A network and topic analysis. Social Media+ Society, 9(1), 20563051231155106.

Zhang, L., Wang, X., Cooper, E., Evans, N., & Yamagishi, J. (2022). The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 813-825.

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

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-Attentional Deepfake Detection. In IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2185–2194.

Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y.-G. (2020). WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In Proc. of the 28th ACM Int. Conf. on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 2382–2390.

7. Annex I

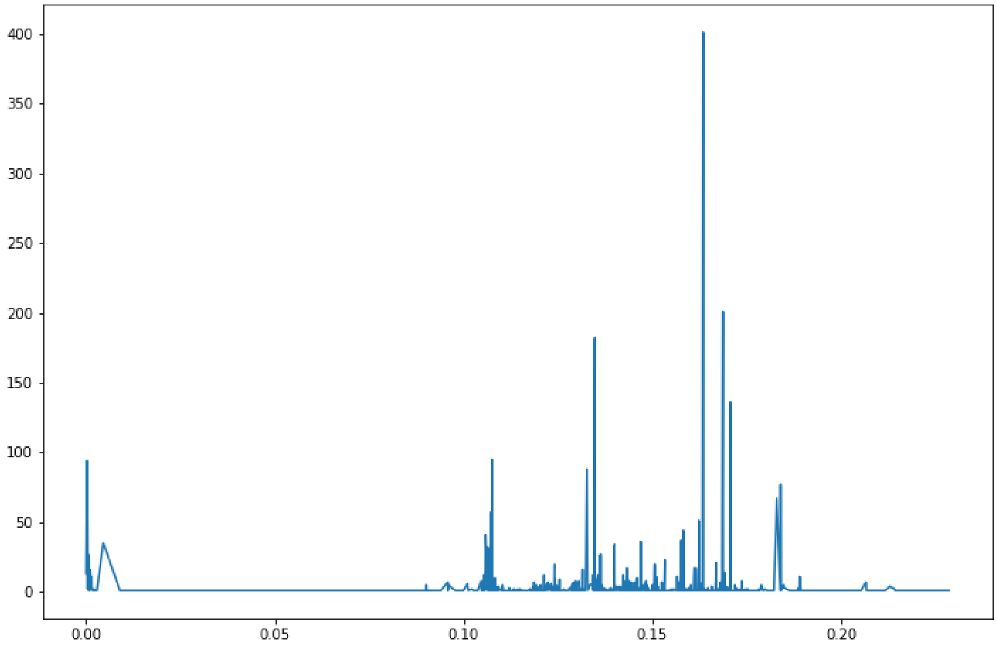


Figure 17: Distribution of closeness centrality in the Telegram network