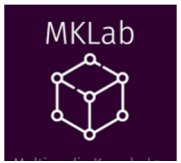
Technological, Behavioural and Policy Aspects of *Visual* Online Disinformation

Yiannis Kompatsiaris, CERTH-ITI, Director

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS



Centre for Research and Technology Hellas

CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

Founded in 2000,
Five Institutes

>1200 projects
1100 international collaborations
5 institutes
>1500 employees

In TOP-15 E.U.
research institutions in competitive research grants

Centre for Research and Technology Hellas -
Information Technologies Institute

Information Technologies Institute

The **largest** among the five institutes of CERTH

>650 employees

> 175 EU projects

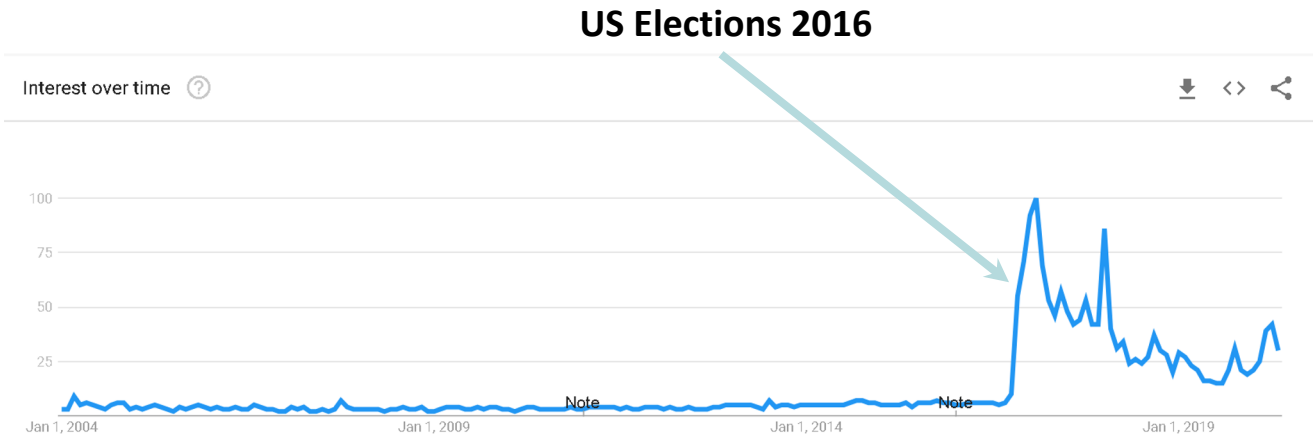
>100 national and industrial projects

> 40M € via R&D activity -last 5 years

Infrastructure
HPC, Smart Home, Robots, Autonomous Vehicle, Drones, EEG, 3D Printing, VR

The Rise of Fake News

Volume for query “fake news” over time: A key milestone has been the US Elections in 2016, which marked the beginning of large-scale coordinated disinformation campaigns.



https://trends.google.com/trends/explore?date=all&geo=US&q=fake%20news

Some definitinos

Disinformation

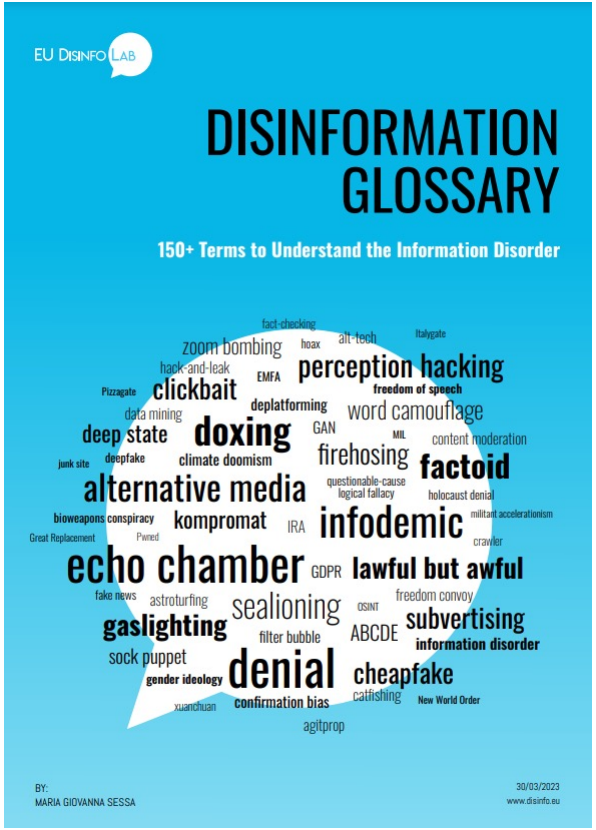
Information that is **false** and **deliberately created** to harm a person, social group, organization or country.

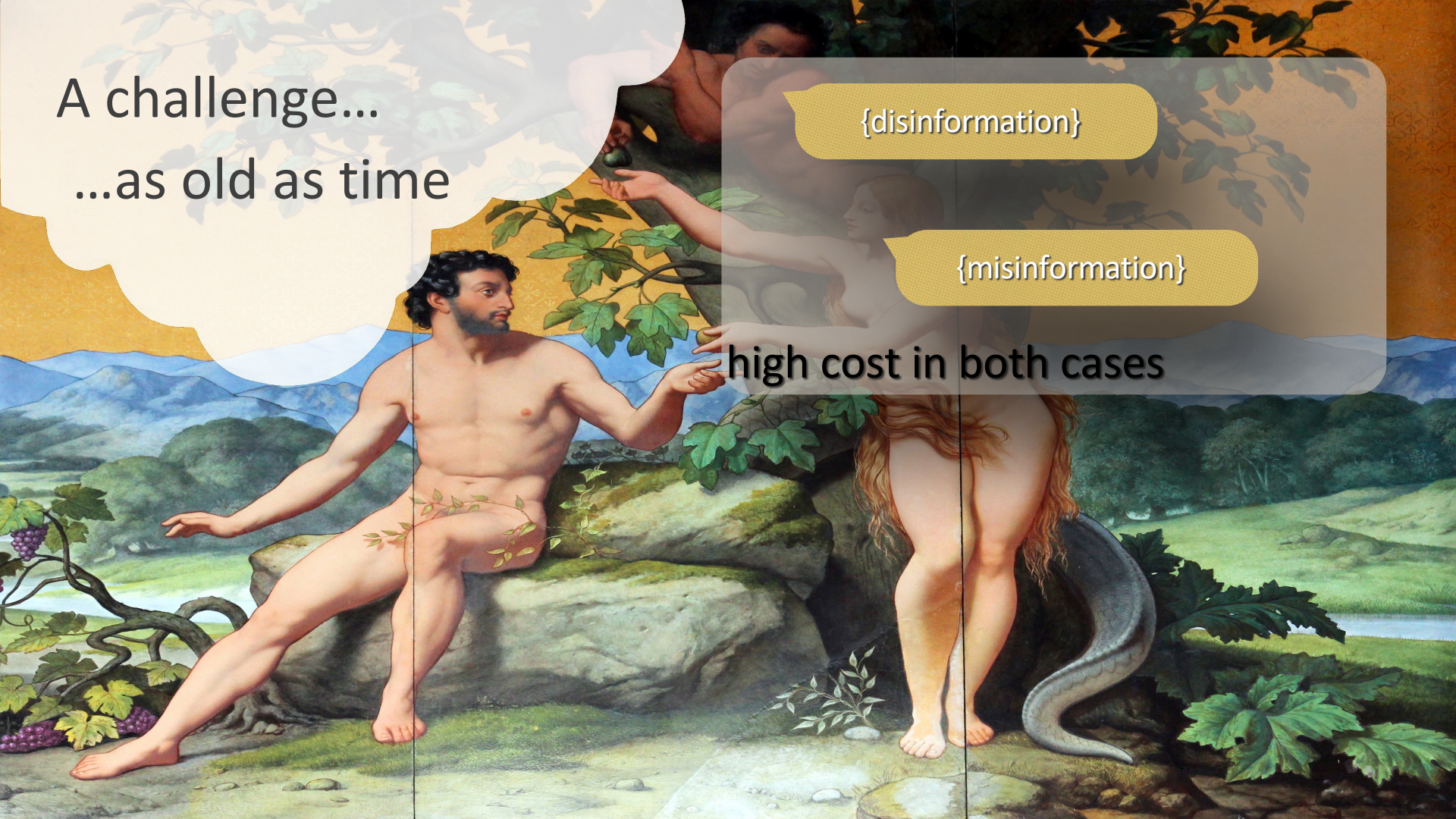
Misinformation

Information that is **false**, but **not created with the intention** of causing harm.

Malinformation

Information that is **based on reality, used to inflict harm** on a person, organization or country.




A challenge…
…as old as time
{disinformation}
{misinformation}
high cost in both cases

Much harder today

- the **rapid growth of social media and the Internet** has made it easier for false information to spread quickly and widely
- people are often more likely to share or believe false information that **confirms their pre-existing beliefs or opinions** (confirmation bias)
- the rise of **generative AI** (e.g. deepfakes, ChatGPT) is making it increasingly difficult to distinguish between real and fake
- **political polarisation and the erosion of trust** in traditional media sources has created an environment in which disinformation can thrive
- **financial models** of online platforms (click-bait)
- **disinformation campaigns** are often carried out by governments, political organisations, and other powerful groups with an agenda to manipulate public opinion and undermine democratic processes

Basic properties of disinformation

Misleading posts tend to spread faster and wider compared to accurate ones.

- **Context**: Usually before and during electoral processes and referendums, as well as during crises and events

Disinformation usually follows popular news events

- **Audience**: targeting criteria such as geographic location, language, nationality, political beliefs, social and economic profile, age, etc.
- **Impression – Emotion**: engaging content, easy, “convenient” solutions
- **Narrative**: pre-existing tensions, problems are usually exploited (e.g. in politics – elections: immigration, LGBTI, abortion, etc.)
- **Media**: articles, blogs, comments, memes, YouTube videos, hashtags, influencers

The impact can be very high
Key societal values are affected

- **Misleading the public**: the spread false or misleading information can impact individuals' decision-making and beliefs.
- Undermining **democracy**: false or misleading information can be used to manipulate elections and undermine trust in democratic institutions
- **National internal and external security**: national disasters and infowar
- Exacerbating **conflicts**: disinformation can fuel social and political tensions, leading to increased conflict and violence
- Damaging **reputations**: the distribution of various rumors and claims is frequently used to harm the reputation of individuals and organizations
- Impairing **public health**: conspiracy theories and dangerous misinformation, like non-scientifically based medical advice about causes, symptoms, and treatments

G7 Hiroshima AI Process
G7 Digital & Tech Ministers' Statement

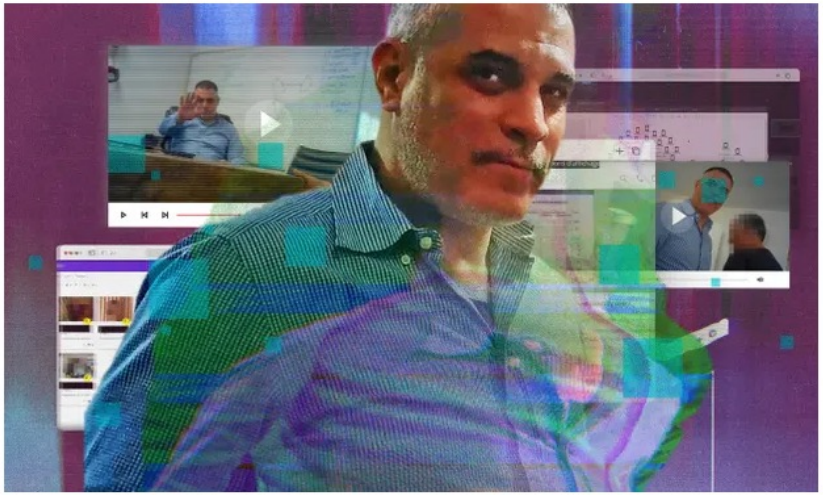
1. We, the G7 Digital and Tech Ministers and partners met virtually on September 7th, 2023 for discussions on the opportunities and challenges of advanced artificial intelligence (AI) systems, with a focus on foundation models and generative AI as

II. ***Understanding of Priority risks, challenges and opportunities from OECD's Report***

7. From a report compiled and drafted by the OECD in July/August 2023, a range of risks and opportunities were identified as priorities and as the basis for our consideration on our common understanding, position, and future action, including for collective work on generative AI. For example, the report identified transparency, disinformation, intellectual property rights, privacy and protection of personal data,

Revealed: the hacking and disinformation team meddling in elections

- **'Team Jorge' unit exposed by undercover investigation**
- **Group sells hacking services and access to vast army of fake social media profiles**
- **Evidence unit behind disinformation campaigns across world**
- **Mastermind Tal Hanan claims covert involvement in 33 presidential elections**



Tal Hanan has always denied any wrongdoing. Composite: Guardian Design/Haaretz/The Marker/Radio France

A team of Israeli contractors who claim to have manipulated more than 30 elections around the world using hacking, sabotage and automated disinformation on social media has been exposed in a new investigation.

The unit is run by Tal Hanan, a 50-year-old former Israeli special forces operative who now works privately using the pseudonym "Jorge", and appears to have been working under the radar in elections in various countries for more than two decades.

He is being unmasked by an international consortium of journalists. Hanan and his unit, which uses the codename "Team Jorge", have been exposed by undercover footage and documents leaked to the Guardian.

Hanan did not respond to detailed questions about Team Jorge's activities and methods but said: "I deny any wrongdoing."

https://www.theguardian.com/world/2023/feb/15/revealed-disinformation-team-jorge-claim-meddling-elections-tal-hanan

Visual disinformation is dangerous

- More persuasive than text
- Attracts more attention
- More tempting to share
- Can easily cross borders

Hameleers, M., Powell, T. E., Van Der Meer, T. G., & Bos, L. (2020). **A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media**. *Political Communication*, *37*(2), 281-301.

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). **Visual mis-and disinformation, social media, and democracy**. *Journalism & Mass Communication Quarterly*, *98*(3), 641-664.

Thomson, T. J., Angus, D., Dootson, P., Hurcombe, E., & Smith, A. (2020). **Visual mis/disinformation in journalism and public communications: current verification practices, challenges, and future opportunities**. *Journalism Practice*, 1-25.

ing

The Famous Shark

2005

https://www.snopes.com/photos/animals/puertorico.asp

Maury Page
@mopage19

Follow

A shark photographed on I-75 just outside of Naples, FL

This is insane. #HurricaneIrma

8:12 PM - Sep 10, 2017

1,506 5,538 12,545

SoaR Gilli ✈ Logan
@Mcgilligan

Follow

A shark was pushed inland from #HurricaneMatthew

7:27 PM - Oct 7, 2016

22 263 506

Jason Michael
@Jeggit

Follow

Believe it or not, this is a shark on the freeway in Houston, Texas. #HurricaneHarvy

9:00 AM - Aug 28, 2017

7,168 88,805 149,387

Austin. ⚠
@austinelement

Follow

Shark in road #sandy

2:53 AM - Oct 30, 2012

6 86 13

Editing examples

Out of context

- 2015 Videos
- Borders of Slovenia and Croatia
- There is some truth behind the headline
    - “Caravan of light” movement existed

https://www.france24.com/en/tv-shows/truth-or-fake/20220923-no-this-video-doesn-t-show-thousands-of-syrian-migrants-trying-to-cross-into-the-eu

→ TRUTH OR FAKE

No, this video doesn't show thousands of Syrian migrants trying to cross into the EU

Issued on: 23/09/2022 - 22:46

By: Vedika BAHL Follow

A video circulating on social media claims to show the "Caravan of Light" – a convoy of Syrian refugees in Turkey – heading for the Greek border, intending to make its way into Europe. But while reports show that plans for a caravan may indeed be underway, this viral video is not all it appears. We explain the details in this edition of Truth or Fake with Vedika Bahl.

DeepFakes

- Content, generated by deep neural networks, that seems authentic to human eye
- Most common form: generation and manipulation of human face

Source: Media Forensics and DeepFakes: an overview

Source: https://www.youtube.com/watch?v=iHv6Q9ychnA

Source: https://en.wikipedia.org/wiki/Deepfake

DeepFakes Generation

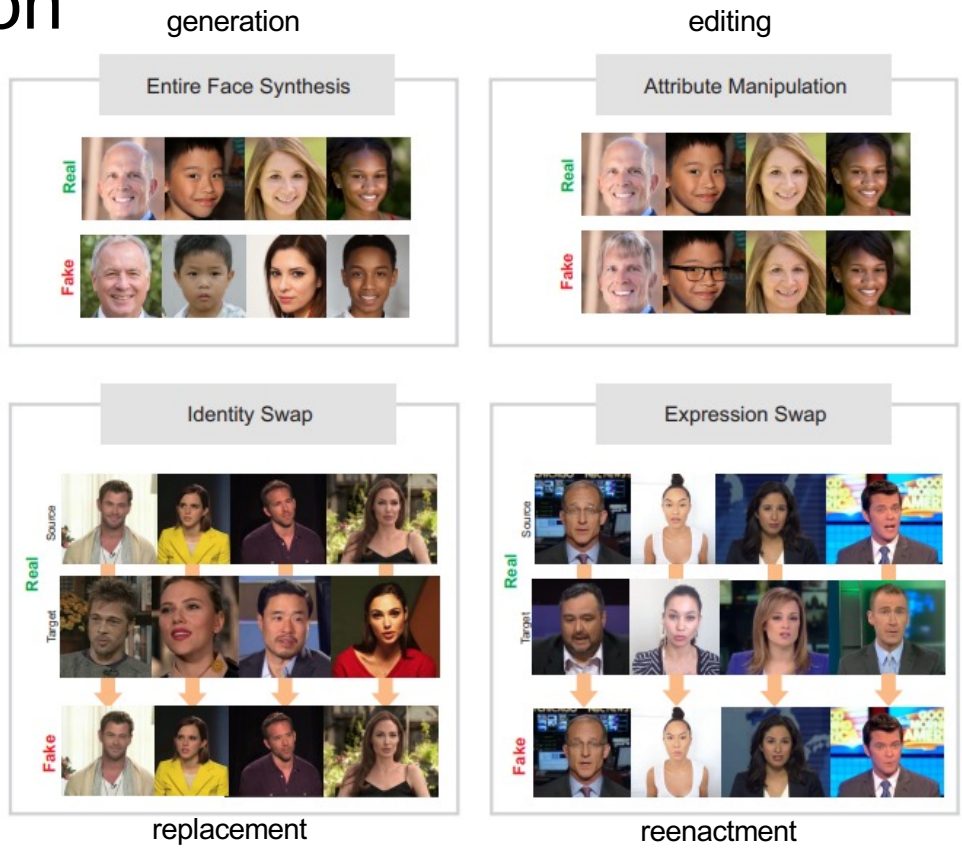
Four main types of face DeepFakes: a) *Entire face synthesis*, b) *Attribute manipulation*, c) *Identity swap*, d) *Expression swap*.

Lip syncing and *voice generation* are also common types in video and audio content.

Tolosana, R., et al. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

Verdoliva, L. (2020). Media forensics and deepfakes: an overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910-932.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), 54(1), 1-41.



Source: *DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection (Tolosana et al., 2021)*

Gaining popularity

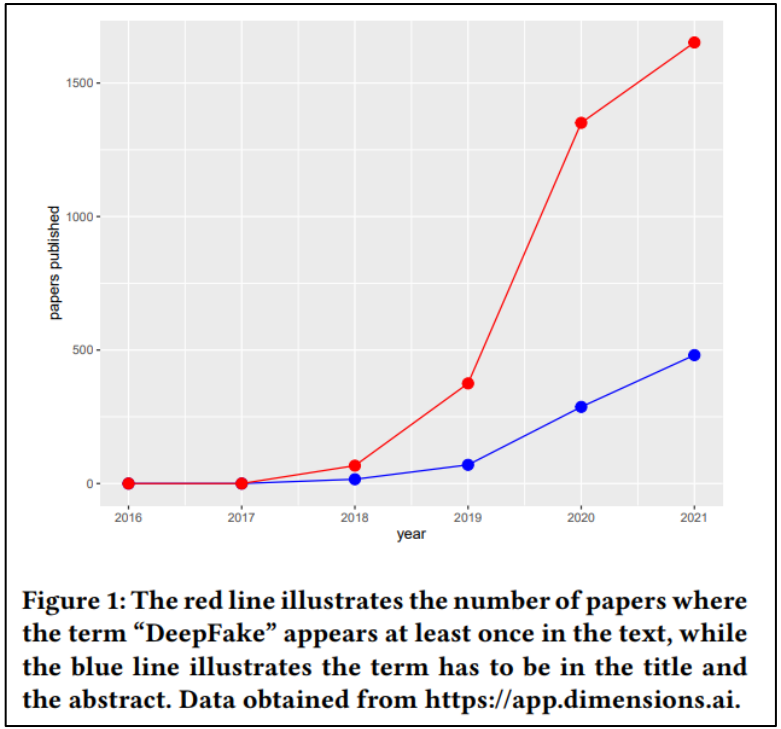


Figure 1: The red line illustrates the number of papers where the term "DeepFake" appears at least once in the text, while the blue line illustrates the term has to be in the title and the abstract. Data obtained from https://app.dimensions.ai.

Baxevanakis, S., et al. (2022). The MeVer DeepFake Detection Service: Lessons Learnt from Developing and Deploying in the Wild. Submitted to ICMR MAD 2022

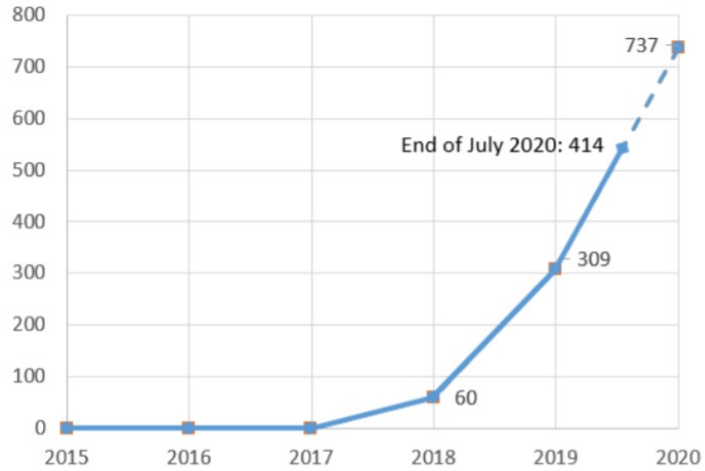
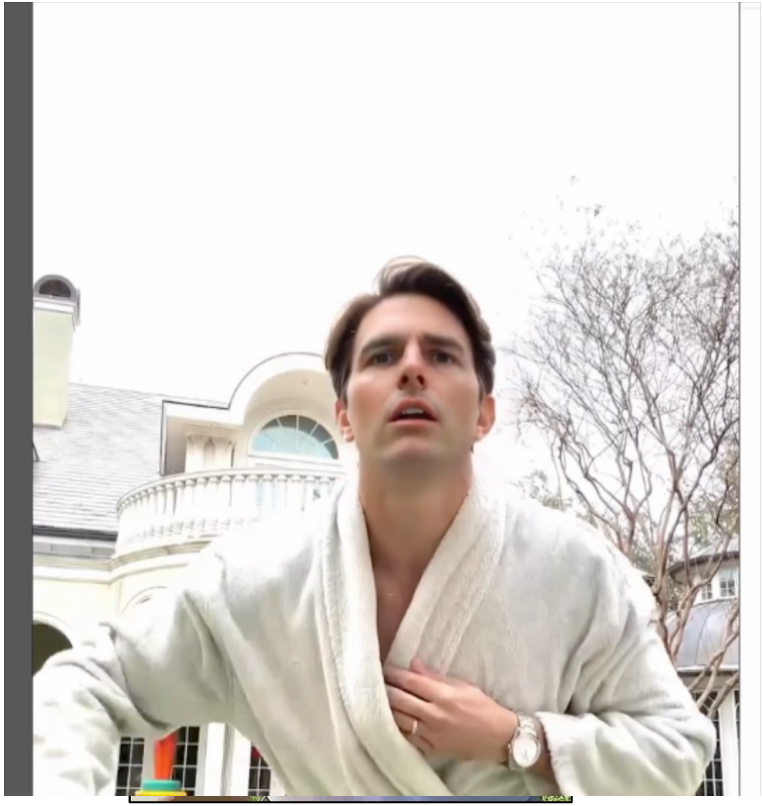


Fig. 1. Number of papers related to deepfakes in years from 2015 to 2020, obtained from https://app.dimensions.ai on 24 July 2020 with the search keyword "deepfake" applied to full text of scholarly papers. The number of such papers in 2018 and 2019 are 60 and 309, respectively. From the beginning of 2020 to near the end of July 2020, there are 414 papers about deepfakes and we linearly estimate that this number will be rising to more than 730 until the end of 2020.

Nguyen, T. T., et al. (2019). Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573, 1.*

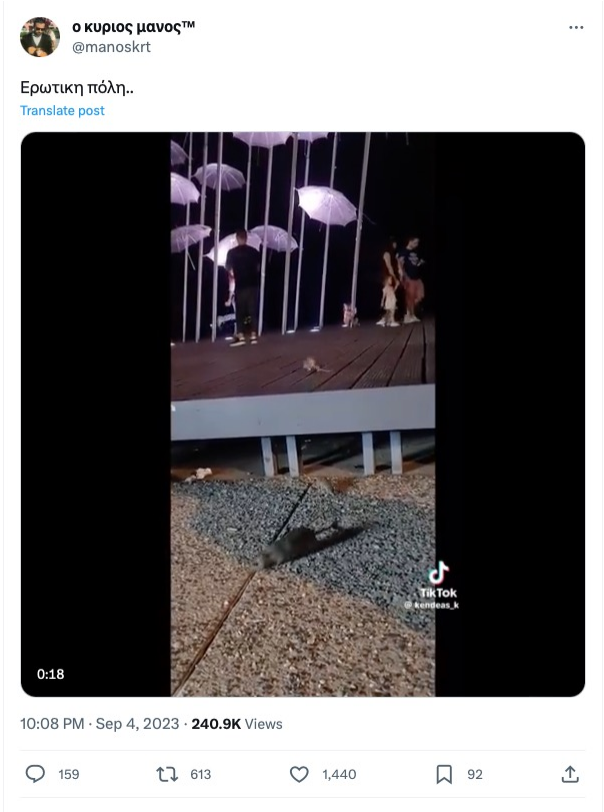
A New Level of Realism

- Created by Chris Ume, a VFX specialist
- Not detected by any of the commercial detection services
- Not discernible by human inspection
- Potential for misleading
  but to date barriers are still high
- a lot of expertise, skill and time
- an impersonator who looks like the target
  (Miles Fisher)



https://www.theverge.com/2021/3/5/22314980/tom-cruise-deepfake-tiktok-videos-ai-impersonator-chris-ume-miles-fisher

Here in Thessaloniki…



https://www.reader.gr/ellada/telika-einai-alitheia-i-ohi-pos-o-thermaikos-ehei-gemisei-pontikia/529795

Face Generation

https://miro.medium.com/v2/resize:fit:1100/format:webp/1*Y6QUypGsfleg04rEWcLpkA.png

Potential Risks and Harms

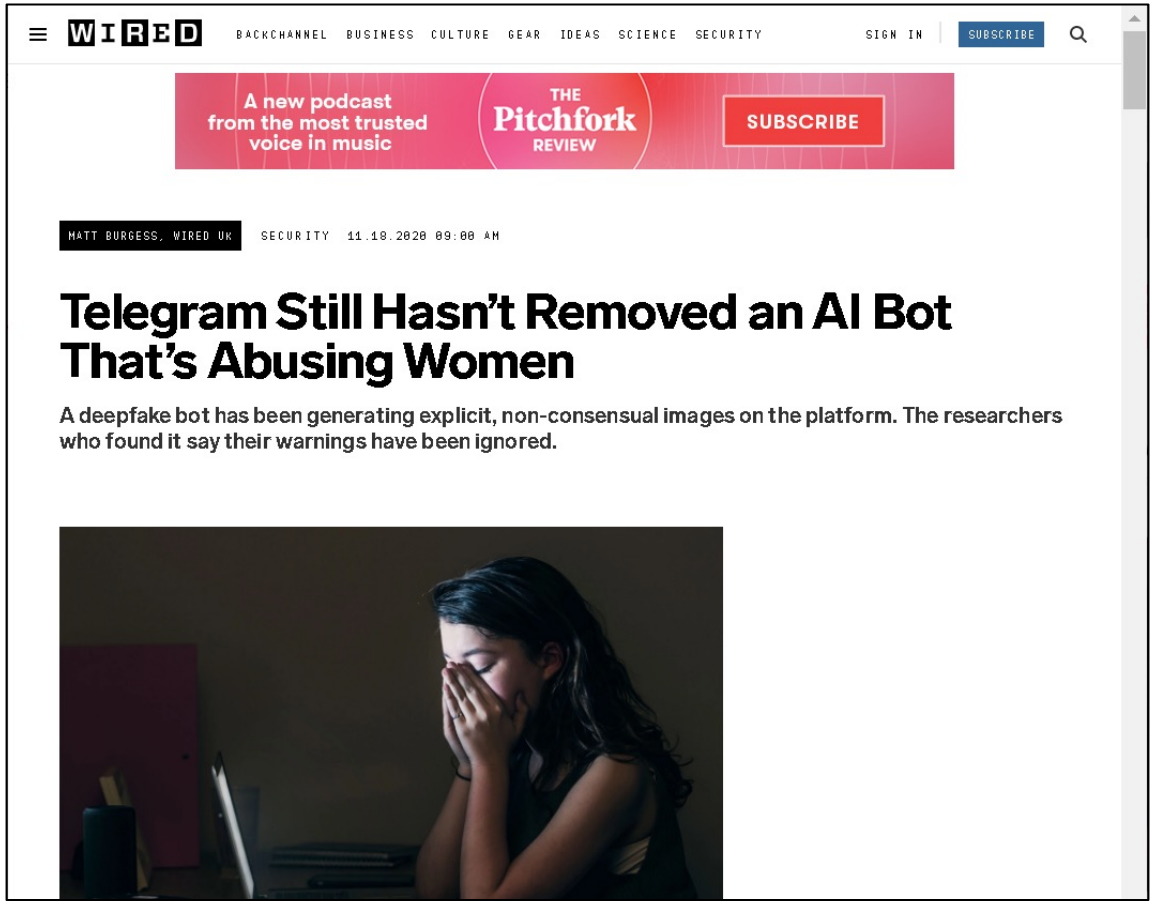
Psychological harm	Financial harm	Societal harm
• (S)extortion • Defamation • Intimidation • Bullying • Undermining trust	• Extortion • Identity theft • Fraud (e.g. insurance/payment) • Stock-price manipulation • Brand damage • Reputational damage	• News media manipulation • Damage to economic stability • Damage to the justice system • Damage to the scientific system • Erosion of trust • Damage to democracy • Manipulation of elections • Damage to international relations • Damage to national security Damage to Public Health

Tackling deepfakes in European policy, Panel for the Future of Science and Technology, Scientific Foresight Unit (STOA), July 2021

Popular Instagram Photographer Revealed as AI Fraud

“A popular “photographer” who has amassed almost 30,000 followers on Instagram has admitted that his portraits are actually generated by artificial intelligence (AI)…”

https://petapixel.com/2023/02/21/popular-instagram-photographer-revealed-as-ai-fraud/



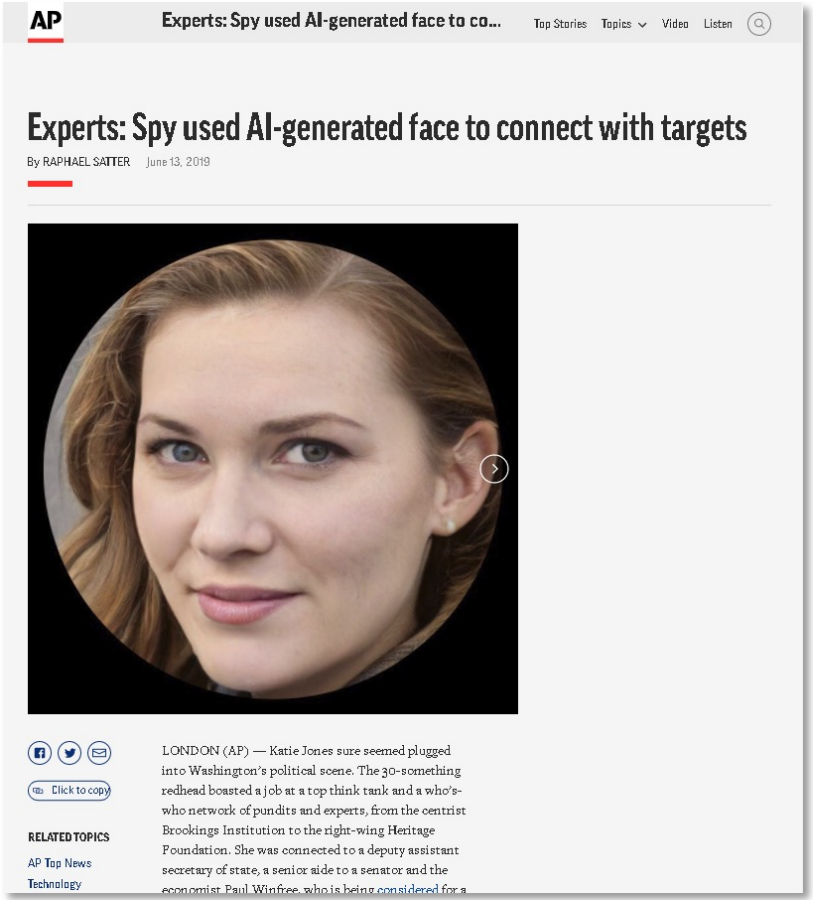
"The bot uses a version of the DeepNude AI tool, which was originally created in 2019, to remove clothes from photos of women and generate their body parts. Anyone can easily use the bot to generate images. More than 100,000 such images have been publicly shared by the bot in several Telegram chat channels associated with it. "

https://www.wired.com/story/telegram-still-hasnt-removed-an-ai-bot-thats-abusing-women/

Fake Identities

But Katie Jones doesn't exist, The Associated Press has determined. Instead, the persona was part of a vast army of phantom profiles lurking on the professional networking site LinkedIn. And several experts contacted by the AP said Jones' profile picture appeared to have been created by a computer program….

https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d



AP Experts: Spy used AI-generated face to co... Top Stories Topics Video Listen

Experts: Spy used AI-generated face to connect with targets

By RAPHAEL SATTER June 13, 2019

Click to copy

RELATED TOPICS

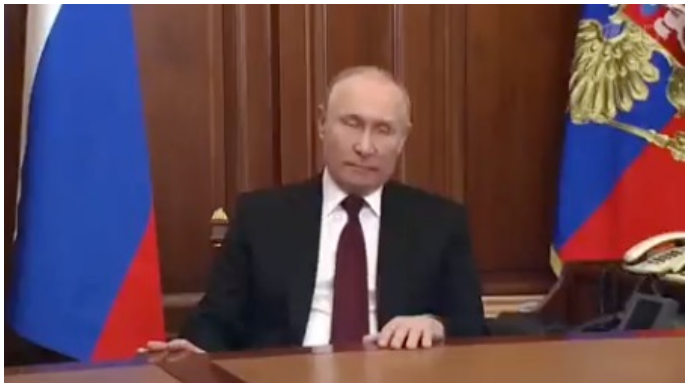
AP Top News

Technology

LONDON (AP) — Katie Jones sure seemed plugged into Washington's political scene. The 30-something redhead boasted a job at a top think tank and a who's-who network of pundits and experts, from the centrist Brookings Institution to the right-wing Heritage Foundation. She was connected to a deputy assistant secretary of state, a senior aide to a senator and the economist Paul Winfree, who is being considered for a

DeepFakes and National (cyber)Security

A 68-second deepfake video appeared in March 2022 during the third week of Russia's invasion in Ukraine, depicting Ukrainian President Volodymyr Zelenskyy calling for the surrender of arms. The video appeared on a compromised Ukrainian news web site and was then widely circulated on social media.
https://nypost.com/2022/03/17/deepfake-video-shows-volodymyr-zelensky-telling-ukrainians-to-surrender/

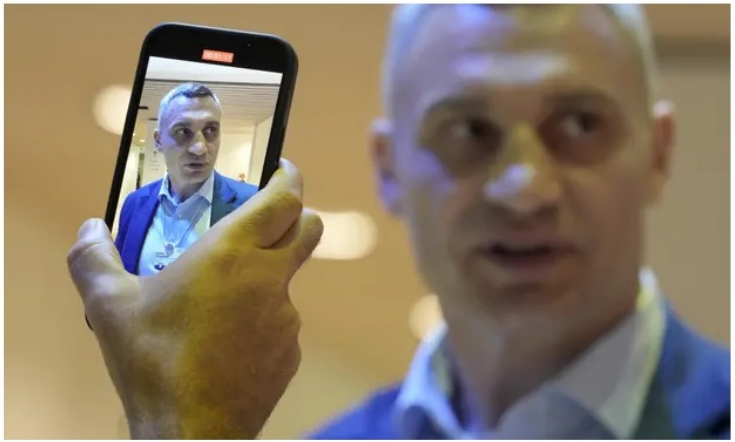


A few days later a video was circulated on social media that supposedly showed Russian President Vladimir Putin announcing that the Russian military was surrendering. A tweet sharing the video with a caption prompted Russian soldiers to lay down their weapons and go home.
https://www.snopes.com/fact-check/putin-deepfake-russian-surrender

DeepFakes and National (cyber)Security

European politicians duped into deepfake video calls with mayor of Kyiv

Person who sounds and looks like Vitali Klitschko has spoken with mayors of Berlin, Madrid and Vienna



Someone has been impersonating the mayor of Kyiv, Vitali Klitschko – the real one seen here. Photograph: Markus Schreiber/AP

The mayors of several European capitals have been duped into holding video calls with a deepfake of their counterpart in Kyiv, Vitali Klitschko.

The mayor of Berlin, Franziska Giffey, took part in a scheduled call on the Webex video conferencing platform on Friday with a person she said looked and sounded like Klitschko.

… The mayor of Berlin, Franziska Giffey, took part in a scheduled call on the Webex video conferencing platform on Friday with a person she said looked and sounded like Klitschko.

“There were no signs that the video conference call wasn’t being held with a real person,” her office said in a statement.

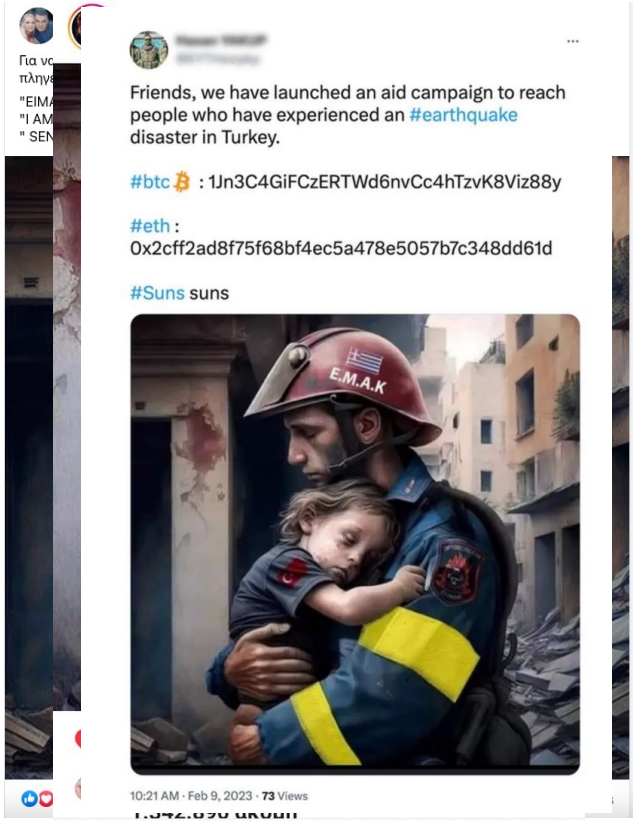
https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko

Face Swapping apps: DeepFakes going Mainstream

“The app [Reface] normalises deepfakes, and not everyone understands the concerns arising from them because not everyone has the digital know-how to differentiate what is real and what isn’t,” Apurva Singh, a privacy expert and volunteer legal counsel at Software Freedom Law Center, India….

https://www.vice.com/en/article/wxqkbn/viral-reface-app-going-to-make-deepfake-problem-worse

Multimedia manipulation and generation for all



- It was created using Midjourney and Photoshop
- Became extremely popular
- Used by scammers for fake donation campaigns (BBC: https://www.bbc.com/news/world-europe-64599553)

The Supply of Disinformation Will Soon Be Infinite

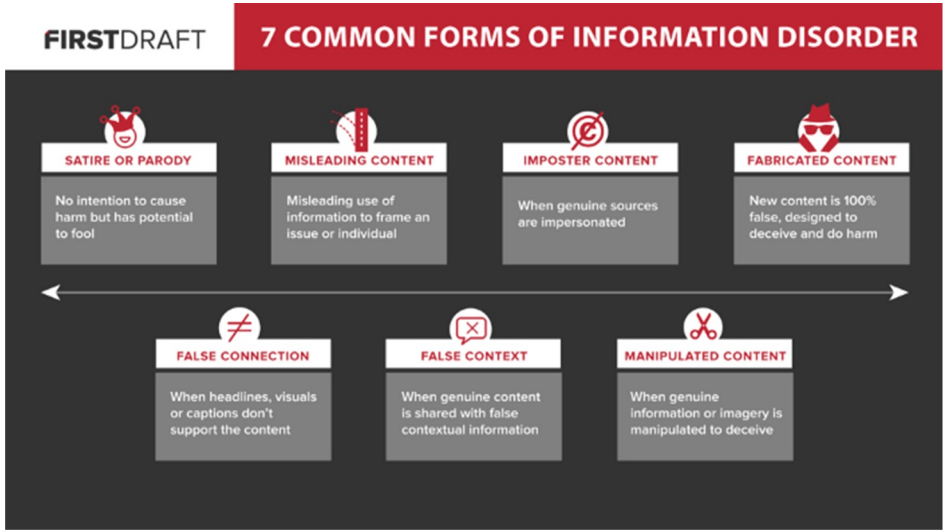
Disinformation campaigns used to require a lot of human effort, but artificial intelligence will take them to a whole new level.

By Renée DiResta

https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400/

Visual disinformation comes in many forms

- Manipulated photos/video
- Deepfakes
- Visuals out of context
- False connections
- Visual memes



https://medium.com/1st-draft/information-disorder-part-3-useful-graphics-2446c7dbb485

…many different methods and tools are needed

MeVer tools deal with multiple information aspects

Content

- Image Verification Assistant
- DeepFake Detection
- Visual Location Estimation

Context

- Near-duplicate Detection
- Context Aggregation and Analysis

Propagation

- Network Analysis and Visualization

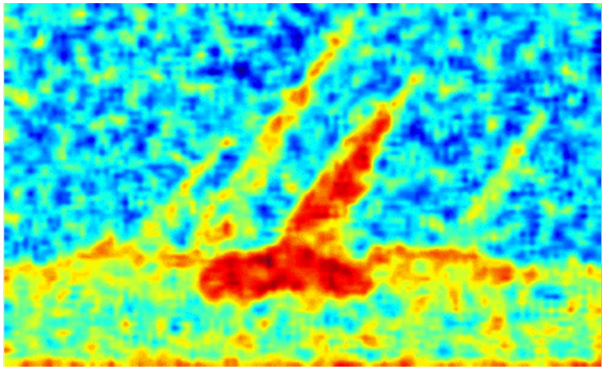
MKLab “historical” background

EU funded projects

Content Verification

Digital image manipulation

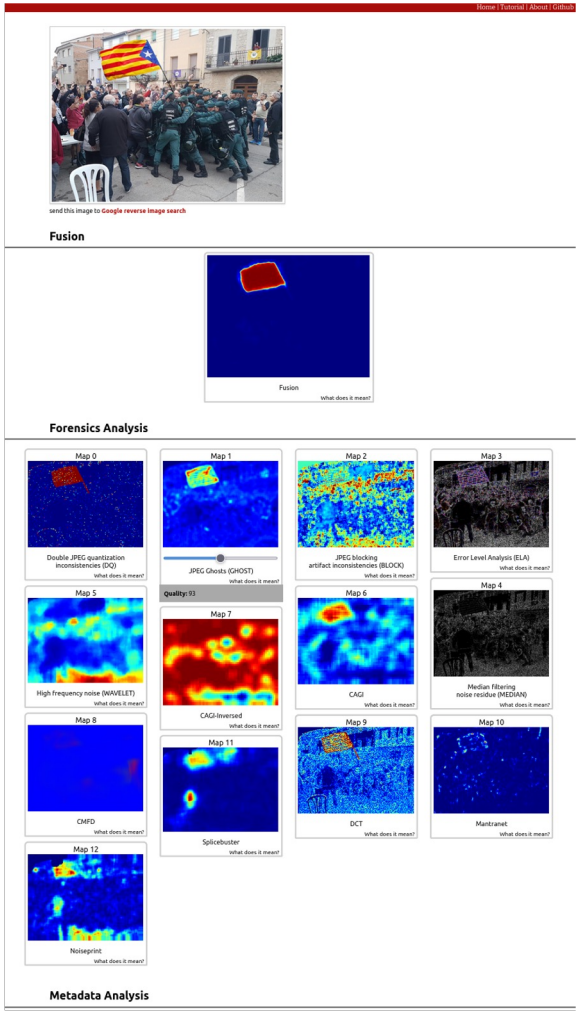
- Image tampering localisation: highlight areas in an image that have been digitally manipulated
- Types of tampering:
  - **Splicing**
  - **Inpainting**
  - **Copy-move**
  - Cropping
  - Enhancement



https://www.npr.org/templates/story/story.php?storyId=92442928

Image Verification Assistant

- Integrates 12 image forensics algorithms and a fusion algorithm
- Allows metadata inspection (EXIF, IPTC, Photoshop and more)
- Estimates the software/hardware origin of JPEG images
- Supports quick reverse image search on Google



https://mever.iti.gr/forensics/

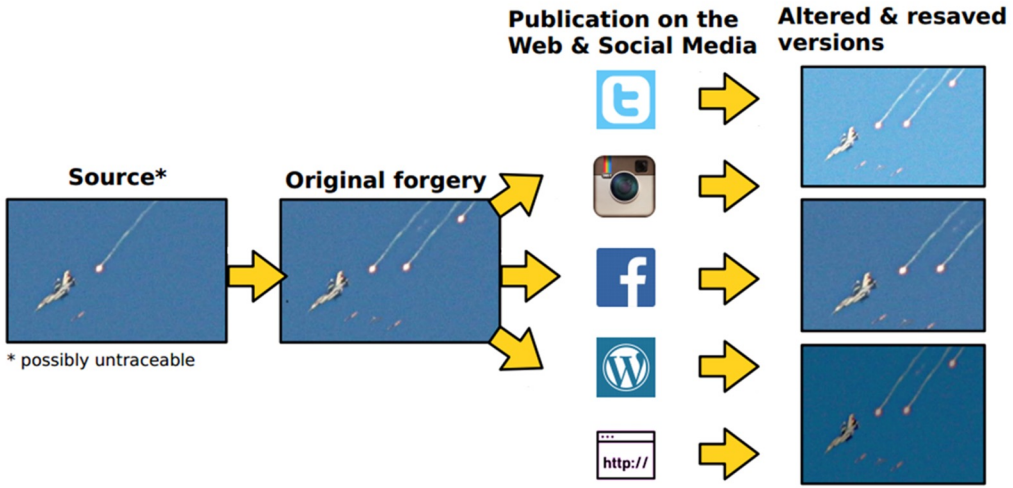
Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y., Bouwmeester, R., & Spangenberg, J. (2016, April). Web and Social Media Image Forensics for News Professionals. In *SMN@ ICWSM*.

Integrated algorithms

- Error Level Analysis (ELA) (Krawetz, 2007)
- Inconsistencies of JPEG Blocking Artifact (DCT) (Ye et al., 2007)
- JPEG ghosts (Farid, 2009)
- Double JPEG quantization inconsistencies (Lin et al., 2009)
- Median filtering noise residue
- Inconsistencies of JPEG Blocking Artifact (BLK) (Li et al., 2009)
- High-frequency noise analysis (Wavelet) (Mahdian & Saic, 2009)
- SpliceBuster (Cozzolino et al., 2015)
- **CAGI** (Iakovidou et al., 2018)
- MantraNet (Wu et al., 2019)
- Copy move forgery detection (Wu et al., 2018)
- Noiseprint (Cozzolino et al., 2018)
- **Fusion** (Charitidis et al., 2021)

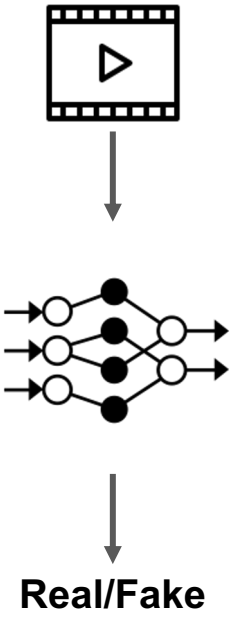
Caveat: Effectiveness of image forensics algorithms

- Certain image forensics algorithms work under very specific conditions
- Internet and social media images are particularly challenging due to multiple recompressions - often applied by the sharing platforms



Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2015). Detecting image splicing in the wild (web). In *International Conference on Multimedia & Expo Workshops (ICMEW), 2015* (pp. 1-6). IEEE

DeepFake Detection



Common architectures

- CNN
- Visual Transformers (ViT)
- Capsule Networks

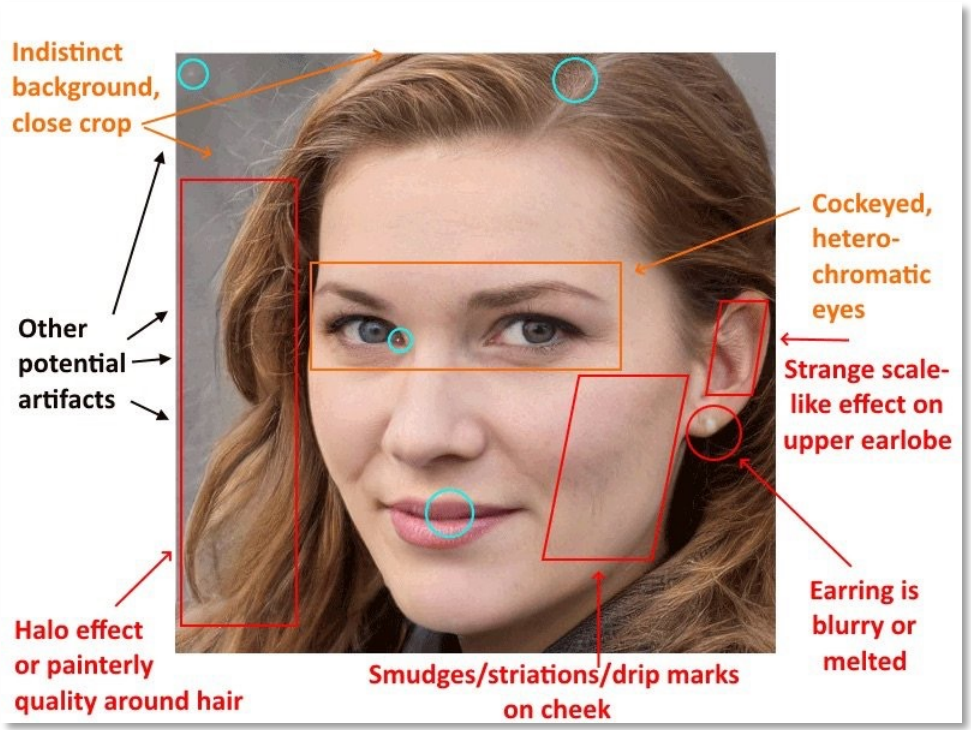
Key Ideas

- Detect abnormal changes in physiological signals e.g. **head poses**, **eye blinking**.
- Exploit left-over manipulation artifacts

Main Problem

Poor generalization to new manipulation techniques.

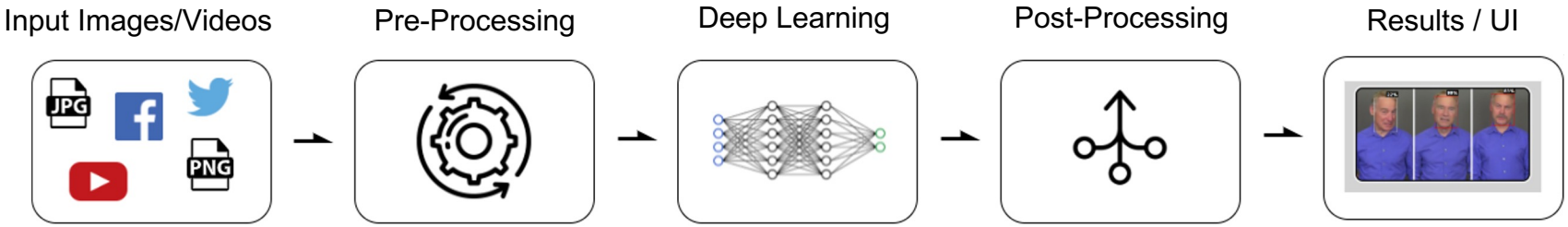
Signs of a DeepFake (in 2020)



https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d

- Different kinds of artifacts
- Blurry areas around lips, hair, earlobs
- Lack of symmetry
- Lighting inconsistencies
- Fuzzy background
- Flickering (in video)

Overview of Approach



- Shot segmentation
- 64 frames per shot
- Face detection
- Face filtering
- Face clustering per shot and filtering

- Ensemble of models (EfficientNet + Transformers)
- Trained on DFDC (120K videos) and WildDeepFake (7314 videos) datasets
- BCE / InfoNCE loss
- DF scores per face

- Average DF scores per face cluster
- Final prediction is the maximum face DF score
- Result preparation

P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos and I. Kompatsiaris. “Investigating the Impact of Pre-processing and Prediction Aggregation on the DeepFake Detection Task”. In Proceedings of the Truth and Trust Online, 2020.

DeepFake Detection Services

- DeepWare: online DeepFake scanner and Android application
- DuckDuckGoose: DeepFake detection system and chrome plugin (only for images)
- DeepFake-o-meter: accepts video link or file and results are sent to user's email

https://deepware.ai//
https://duckduckgoose.ai/
http://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/

The MeVer DeepFake Detection Service

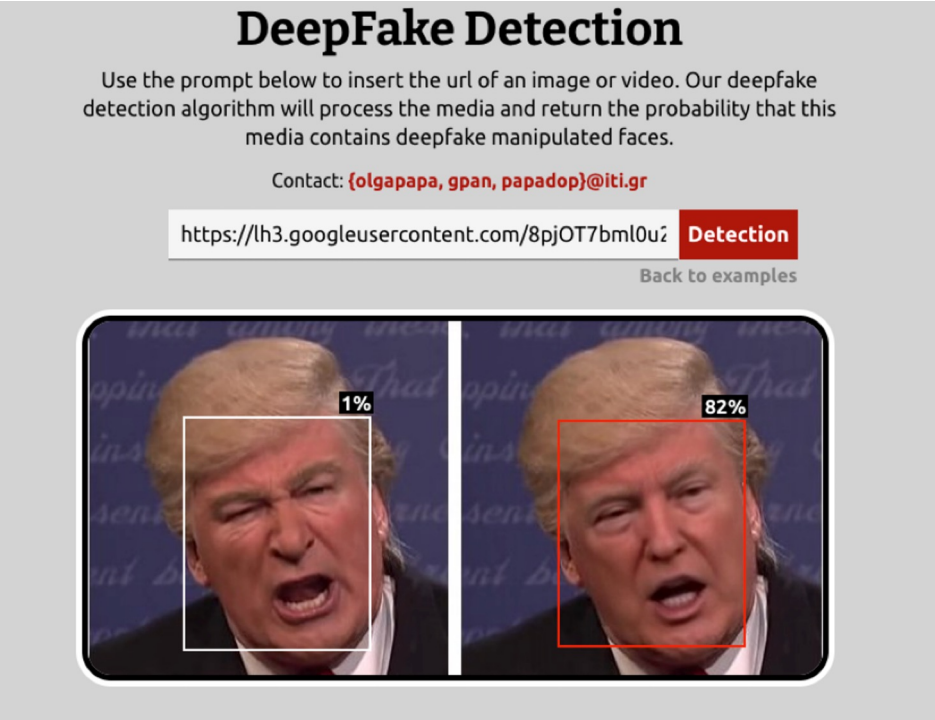
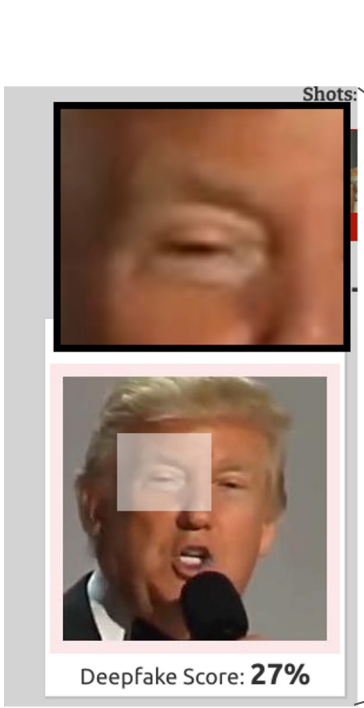


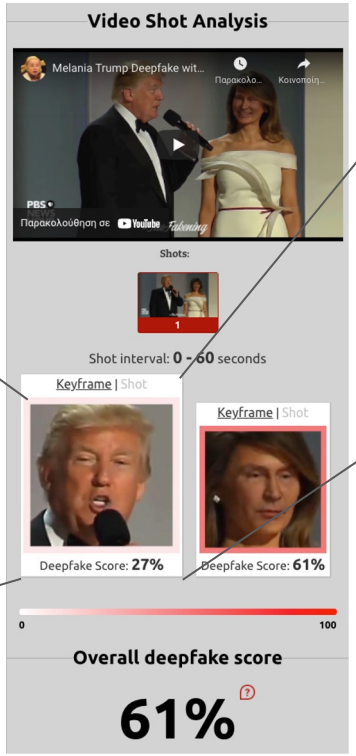
Image Analysis UI

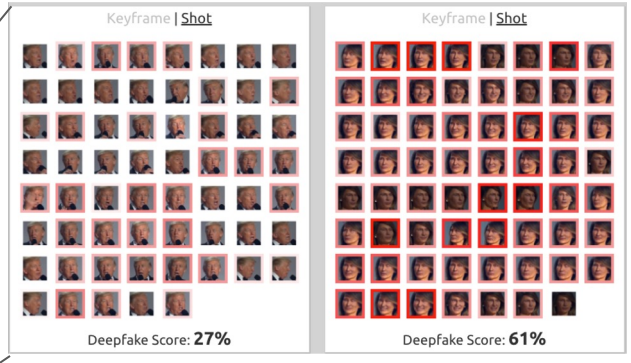
https://mever.iti.gr/deepfake

The MeVer DeepFake Detection Service



Zoom on hover





Frame collage

Video Analysis UI

Evaluation

Adversarial Robustness

Dataset	norm-1	norm-2	norm-inf
FaceForensics++	70.31%	64.04%	50.53%
CelebDF	82.75%	76.01%	50.00%
WildDeepFake	84.94%	63.04%	50.00%

Evasion attack: perform targeted alterations to an image

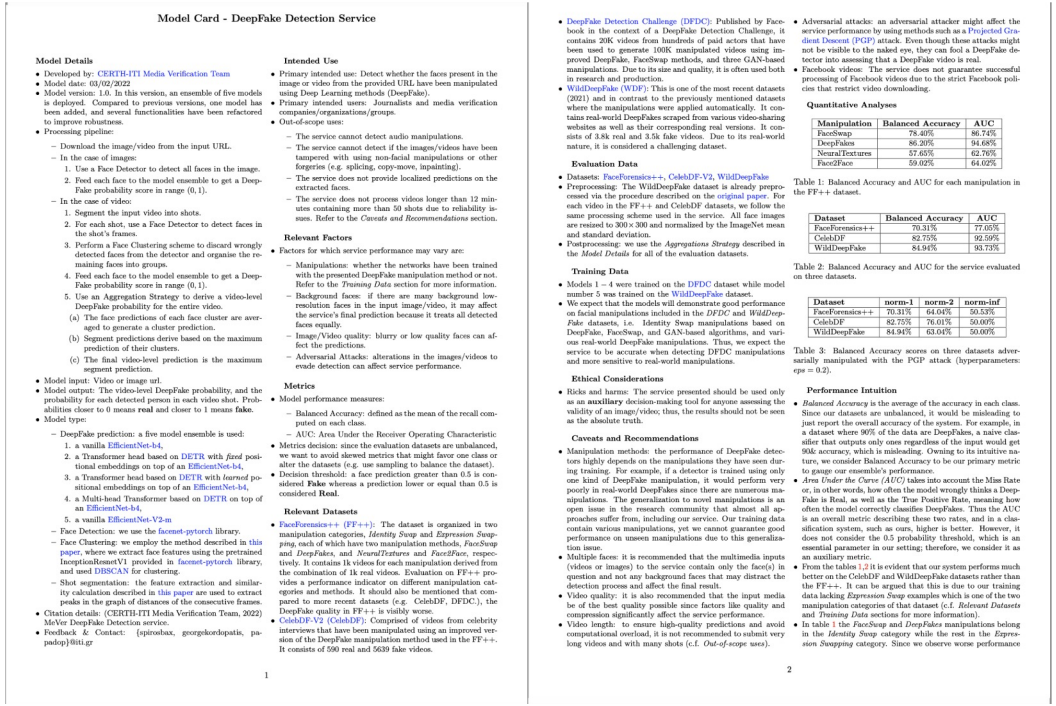
Projected Gradient Descent: white-box evasion attack

Make use of the [Adversarial Robustness Toolbox (ART) by IBM]

https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

Model Card

- inform & guide new users
- contains:
  - model architecture details
  - datasets used
  - evaluation results
  - caveats and recommendations
  - factors that affect performance



Model Card - DeepFake Detection Service

Model Details

- Developed by: CERTH-ITI Media Verification Team
- Model date: 03/02/2022
- Model version: 1.0. In this version, an ensemble of five models is deployed. Compared to previous versions, one model has been added, and several functionalities have been refactored to improve robustness.
- Processing pipeline:
  - Download the image/video from the input URL.
  - In the case of images:
    1. Use a Face Detector to detect all faces in the image.
    2. Feed each face to the model ensemble to get a DeepFake probability score in range $(0, 1)$.
  - In the case of video:
    1. Segment the input video into shots.
    2. For each shot, use a Face Detector to detect faces in the shot's frames.
    3. Perform a Face Clustering scheme to discard wrongly detected faces from the detector and organise the remaining faces into groups.
    4. Feed each face to the model ensemble to get a DeepFake probability score in range $(0, 1)$.
    5. Use an Aggregation Strategy to derive a video-level DeepFake probability for the entire video.
       (a) The face predictions of each face cluster are averaged to generate a cluster prediction.
       (b) Segment predictions derive based on the maximum prediction of their clusters.
       (c) The final video-level prediction is the maximum segment prediction.
- Model input: Video or image url.
- Model output: The video-level DeepFake probability, and the probability for each detected person in each video shot. Probabilities closer to 0 means **real** and closer to 1 means **fake**.
- Model type:
  - DeepFake prediction: a five model ensemble is used:
    1. a vanilla EfficientNet-b4,
    2. a Transformer head based on DETR with *fixed* positional embeddings on top of an EfficientNet-b4,
    3. a Transformer head based on DETR with *learned* positional embeddings on top of an EfficientNet-b4,
    4. a Multi-head Transformer based on DETR on top of an EfficientNet-b4,
    5. a vanilla EfficientNet-V2-m
  - Face Detection: we use the facenet-pytorch library.
  - Face Clustering: we employ the method described in this paper, where we extract face features using the pretrained InceptionResnetV1 provided in facenet-pytorch library, and used DBSCAN for clustering.
  - Shot segmentation: the feature extraction and similarity calculation described in this paper are used to extract peaks in the graph of distances of the consecutive frames.
- Citation details: (CERTH-ITI Media Verification Team, 2022) MeVer DeepFake Detection service.
- Feedback & Contact: {spirosbax, georgekordopatis, papadop}@iti.gr

Intended Use

- Primary intended use: Detect whether the faces present in the image or video from the provided URL have been manipulated using Deep Learning methods (DeepFake).
- Primary intended users: Journalists and media verification companies/organizations/groups.
- Out-of-scope uses:
  - The service cannot detect audio manipulations.
  - The service cannot detect if the images/videos have been tampered with using non-facial manipulations or other forgeries (e.g. splicing, copy-move, inpainting).
  - The service does not provide localized predictions on the extracted faces.
  - The service does not process videos longer than 12 minutes containing more than 50 shots due to reliability issues. Refer to the *Caveats and Recommendations* section.

Relevant Factors

- Factors for which service performance may vary are:
  - Manipulations: whether the networks have been trained with the presented DeepFake manipulation method or not. Refer to the *Training Data* section for more information.
  - Background faces: if there are many background low-resolution faces in the input image/video, it may affect the service's final prediction because it treats all detected faces equally.
  - Image/Video quality: blurry or low quality faces can affect the predictions.
  - Adversarial Attacks: alterations in the images/videos to evade detection can affect service performance.

Metrics

- Model performance measures:
  - Balanced Accuracy: defined as the mean of the recall computed on each class.
  - AUC: Area Under the Receiver Operating Characteristic
- Metrics decision: since the evaluation datasets are unbalanced, we want to avoid skewed metrics that might favor one class or alter the datasets (e.g. use sampling to balance the dataset).
- Decision threshold: a face prediction greater than 0.5 is considered **Fake** whereas a prediction lower or equal than 0.5 is considered **Real**.

Relevant Datasets

- FaceForensics++ (FF++): The dataset is organized in two manipulation categories, *Identity Swap* and *Expression Swapping*, each of which have two manipulation methods, *FaceSwap* and *DeepFakes*, and *NeuralTextures* and *Face2Face*, respectively. It contains 1k videos for each manipulation derived from the combination of 1k real videos. Evaluation on FF++ provides a performance indicator on different manipulation categories and methods. It should also be mentioned that compared to more recent datasets (e.g. CelebDF, DFDC.), the DeepFake quality in FF++ is visibly worse.
- CelebDF-V2 (CelebDF): Comprised of videos from celebrity interviews that have been manipulated using an improved version of the DeepFake manipulation method used in the FF++. It consists of 590 real and 5639 fake videos.

1

- DeepFake Detection Challenge (DFDC): Published by Facebook in the context of a DeepFake Detection Challenge, it contains 20K videos from hundreds of paid actors that have been used to generate 100K manipulated videos using improved DeepFake, FaceSwap methods, and three GAN-based manipulations. Due to its size and quality, it is often used both in research and production.
- WildDeepFake (WDF): This is one of the most recent datasets (2021) and in contrast to the previously mentioned datasets where the manipulations were applied automatically. It contains real-world DeepFakes scraped from various video-sharing websites as well as their corresponding real versions. It consists of 3.8k real and 3.5k fake videos. Due to its real-world nature, it is considered a challenging dataset.

Evaluation Data

- Datasets: FaceForensics++, CelebDF-V2, WildDeepFake
- Preprocessing: The WildDeepFake dataset is already preprocessed via the procedure described on the original paper. For each video in the FF++ and CelebDF datasets, we follow the same processing scheme used in the service. All face images are resized to 300×300 and normalized by the ImageNet mean and standard deviation.
- Postprocessing: we use the *Aggregations Strategy* described in the *Model Details* for all of the evaluation datasets.

Training Data

- Models 1 – 4 were trained on the DFDC dataset while model number 5 was trained on the WildDeepFake dataset.
- We expect that the models will demonstrate good performance on facial manipulations included in the *DFDC* and *WildDeepFake* datasets, i.e. Identity Swap manipulations based on DeepFake, FaceSwap, and GAN-based algorithms, and various real-world DeepFake manipulations. Thus, we expect the service to be accurate when detecting DFDC manipulations and more sensitive to real-world manipulations.

Ethical Considerations

- Ricks and harms: The service presented should be used only as an **auxiliary** decision-making tool for anyone assessing the validity of an image/video; thus, the results should not be seen as the absolute truth.

Caveats and Recommendations

- Manipulation methods: the performance of DeepFake detectors highly depends on the manipulations they have seen during training. For example, if a detector is trained using only one kind of DeepFake manipulation, it would perform very poorly in real-world DeepFakes since there are numerous manipulations. The generalization to novel manipulations is an open issue in the research community that almost all approaches suffer from, including our service. Our training data contain various manipulations, yet we cannot guarantee good performance on unseen manipulations due to this generalization issue.
- Multiple faces: it is recommended that the multimedia inputs (videos or images) to the service contain only the face(s) in question and not any background faces that may distract the detection process and affect the final result.
- Video quality: it is also recommended that the input media be of the best quality possible since factors like quality and compression significantly affect the service performance.
- Video length: to ensure high-quality predictions and avoid computational overload, it is not recommended to submit very long videos and with many shots (c.f. *Out-of-scope uses*).

2

- Adversarial attacks: an adversarial attacker might affect the service performance by using methods such as a Projected Gradient Descent (PGP) attack. Even though these attacks might not be visible to the naked eye, they can fool a DeepFake detector into assessing that a DeepFake video is real.
- Facebook videos: The service does not guarantee successful processing of Facebook videos due to the strict Facebook policies that restrict video downloading.

Quantitative Analyses

Manipulation	Balanced Accuracy	AUC
FaceSwap	78.40%	86.74%
DeepFakes	86.20%	94.68%
NeuralTextures	57.65%	62.76%
Face2Face	59.02%	64.02%

Table 1: Balanced Accuracy and AUC for each manipulation in the FF++ dataset.

Dataset	Balanced Accuracy	AUC
FaceForensics++	70.31%	77.05%
CelebDF	82.75%	92.59%
WildDeepFake	84.94%	93.73%

Table 2: Balanced Accuracy and AUC for the service evaluated on three datasets.

Dataset	norm-1	norm-2	norm-inf
FaceForensics++	70.31%	64.04%	50.53%
CelebDF	82.75%	76.01%	50.00%
WildDeepFake	84.94%	63.04%	50.00%

Table 3: Balanced Accuracy scores on three datasets adversarially manipulated with the PGP attack (hyperparameters: $eps = 0.2$).

Performance Intuition

- *Balanced Accuracy* is the average of the accuracy in each class. Since our datasets are unbalanced, it would be misleading to just report the overall accuracy of the system. For example, in a dataset where 90% of the data are DeepFakes, a naive classifier that outputs only ones regardless of the input would get 90& accuracy, which is misleading. Owning to its intuitive nature, we consider Balanced Accuracy to be our primary metric to gauge our ensemble's performances.
- *Area Under the Curve (AUC)* takes into account the Miss Rate or, in other words, how often the model wrongly thinks a DeepFake is Real, as well as the True Positive Rate, meaning how often the model correctly classifies DeepFakes. Thus the AUC is an overall metric describing these two rates, and in a classification system, such as ours, higher is better. However, it does not consider the 0.5 probability threshold, which is an essential parameter in our setting; therefore, we consider it as an auxiliary metric.
- From the tables 1,2 it is evident that our system performs much better on the CelebDF and WildDeepFake datasets rather than the FF++. It can be argued that this is due to our training data lacking *Expression Swap* examples which is one of the two manipulation categories of that dataset (c.f. *Relevant Datasets* and *Training Data* sections for more information).
- In table 1 the *FaceSwap* and *DeepFakes* manipulations belong in the *Identity Swap* category while the rest in the *Expression Swapping* category. Since we observe worse performance

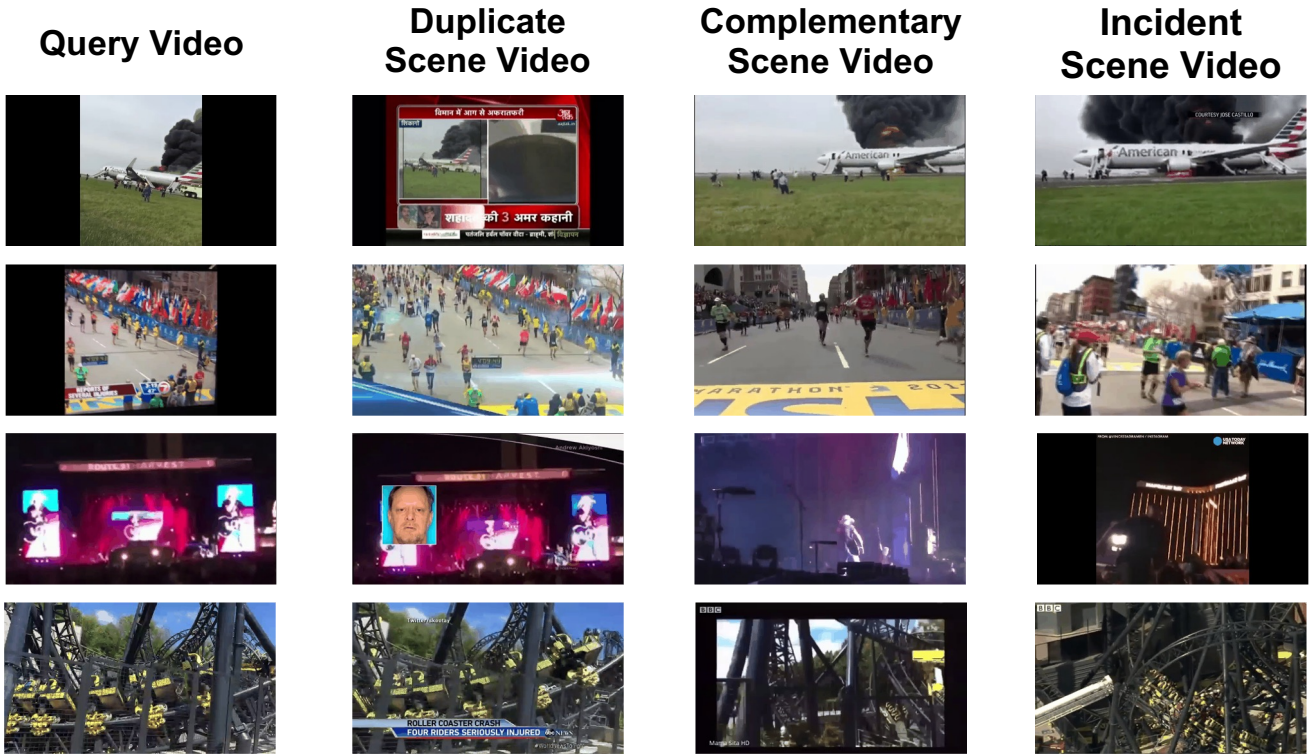
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Context Verification

Near-Duplicate Detection

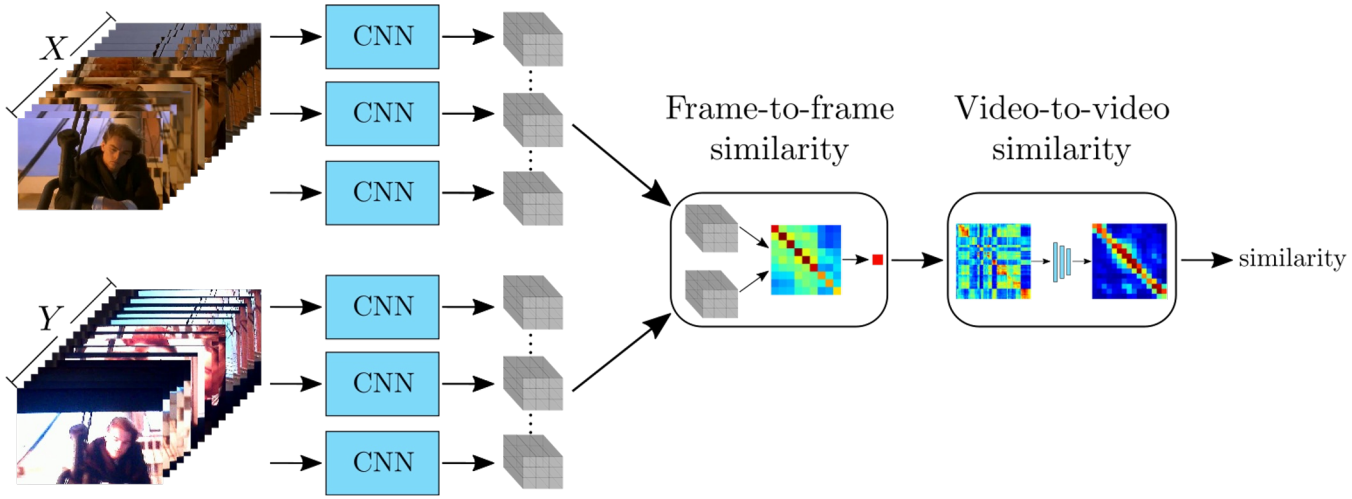
FIVR-200K dataset

- 225,960 videos
- 100 queries
- 4,687 news events
- 4 annotation labels



Kordopatis-Zilos et al. “FIVR: Fine-grained Incident Video Retrieval”. IEEE TMM, 2019

Near-Duplicate Detection

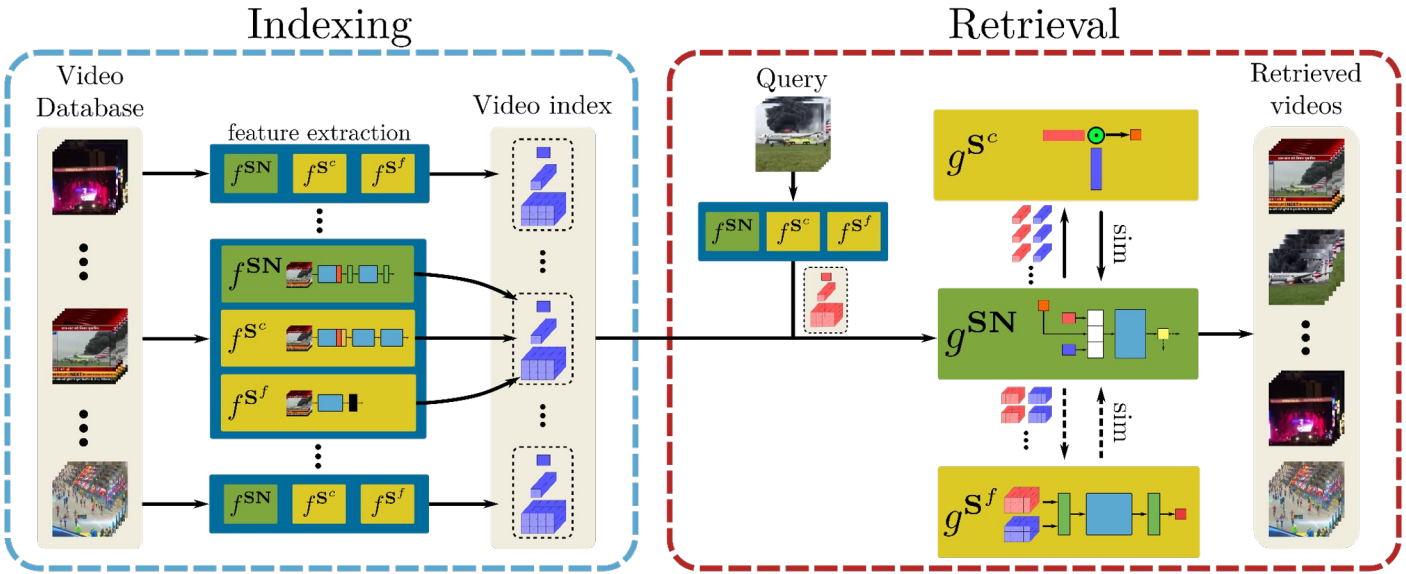


Video Similarity Learning (ViSiL)

- *Fine-grained similarity calculation*
- Learn a video similarity function that considers:
  - *Spatial structure* of video frames (intra-frame relations)
  - *Temporal structure* of videos (inter-frame relations)

Kordopatis-Zilos et al. “ViSiL: Fine-grained spatio-temporal video similarity learning”. ICCV, 2019.

Near-Duplicate Detection



DnS: Distill-and-Select for Video Indexing and Retrieval

- *Knowledge Distillation* from a teacher network to multiple students
  - Different trade-off between accuracy/efficiency
- *Selection Mechanism* between two student networks
  - Select slow but accurate or fast but less accurate student

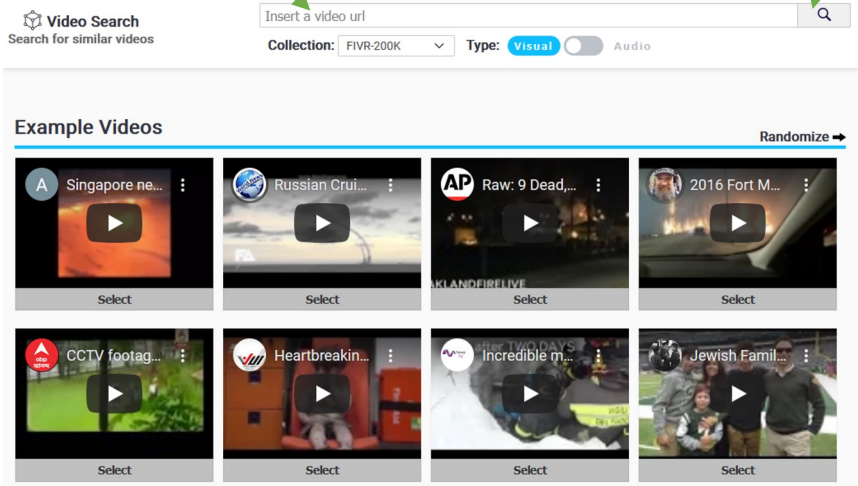
Kordopatis-Zilos et al. “DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval." IJCV, 2021.

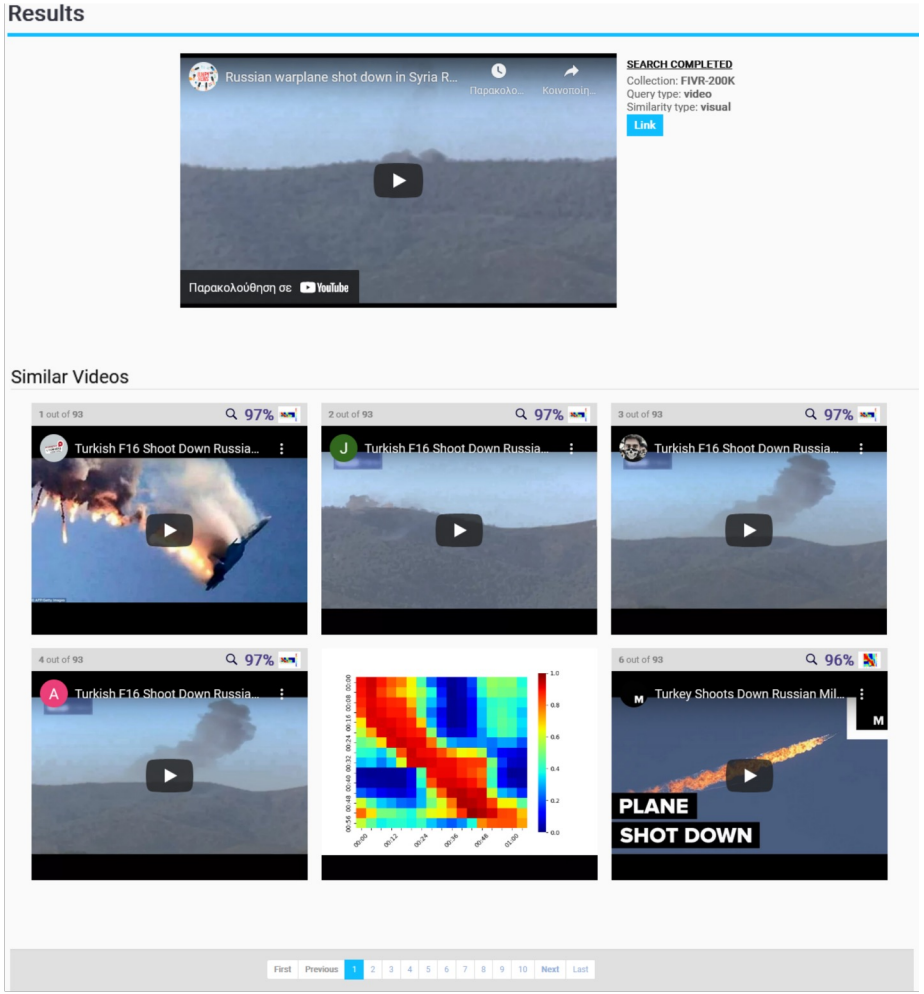
Near-Duplicate Detection

Online demo:
https://mever.iti.gr/video_search/

Provide URL

Start Search

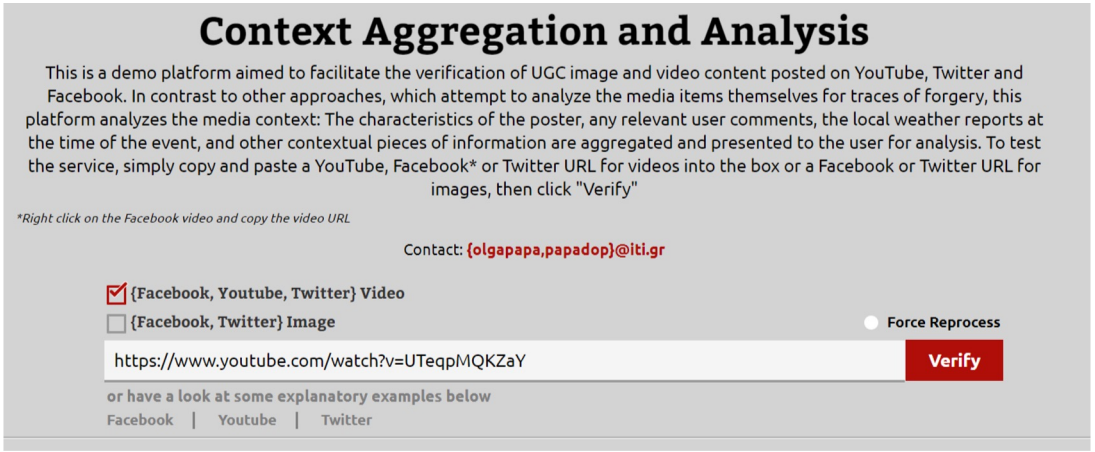


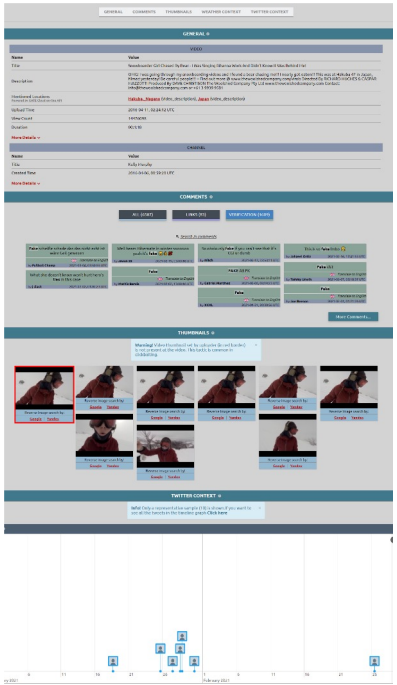


Context Aggregation and Analysis

In contrast to other approaches, which focus on the media items themselves for traces of forgery, this tool analyzes the media context

https://mever.iti.gr/caa/



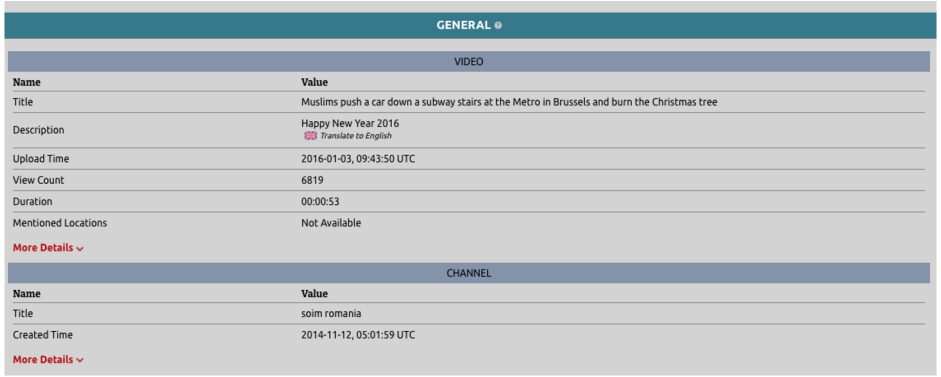


Papadopoulou, O., Giomelakis, D., Apostolidis, L., Papadopoulos, S., & Kompatsiaris, Y. (2019). Context Aggregation and Analysis: A Tool for User-Generated Video Verification. In SIGIR 2019 Workshop on Reducing Online Misinformation Exposure (ROME 2019)

Papadopoulou, O., Zampoglou, M., Papadopoulos, S., & Kompatsiaris, I. (2019). Verification of Web Videos Through Analysis of Their Online Context. In Video Verification in the Fake News Era (pp. 191-221). Springer, Cham.

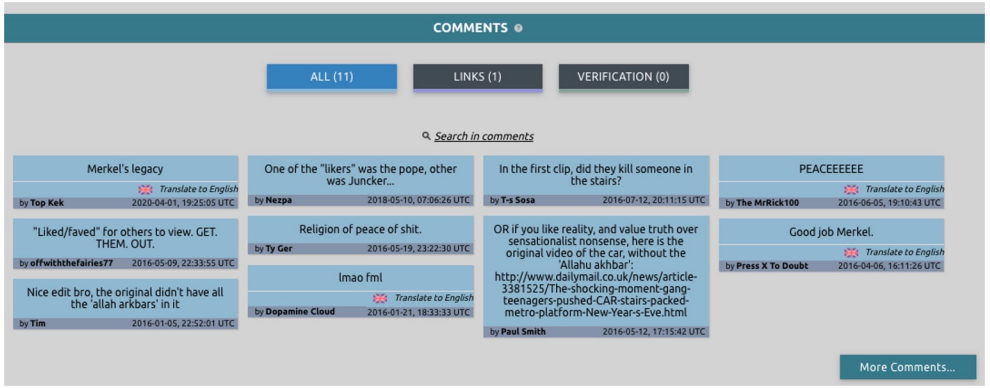
Context Aggregation and Analysis

Video and channel/user metadata



Comments:

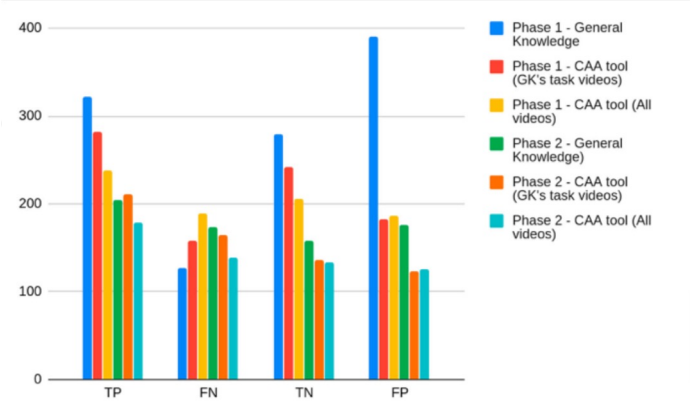
- All comments left below the video
- Links: comments containing links
- Verification: comments containing a verification-related keyword (pre-defined set of keywords)



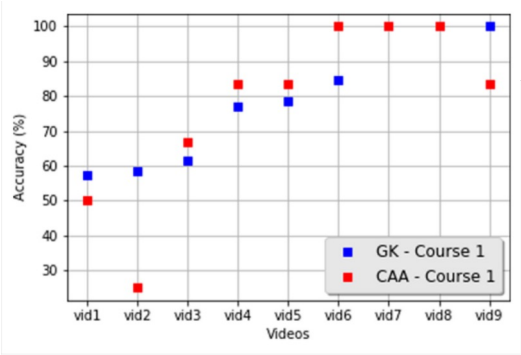
Context Aggregation and Analysis

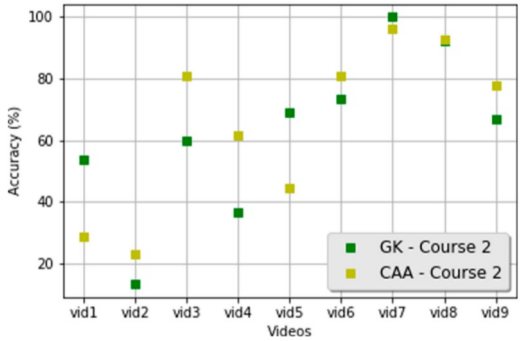
Fake Verification Corpus: Corpus of debunked and verified user-generated videos.

Study on the verification of news video content derived from social media platforms, using the CAA tool and a semi-automated verification practice.



Average time needed for the verification process





Average accuracy per video in GK (without CAA tool) and CAA tool tasks for the same set of fake videos.

Papadopoulou, O., Zampoglou, M., Papadopoulos, S., & Kompatsiaris, I. (2018). A corpus of debunked and verified user-generated videos. Online information review, 43(1), 72-88.

Giomelakis, D., Papadopoulou, O., Papadopoulos, S., & Veglis, A. (2021). Verification of News Video Content: Findings from a Study of Journalism Students. Journalism Practice, 1-30.

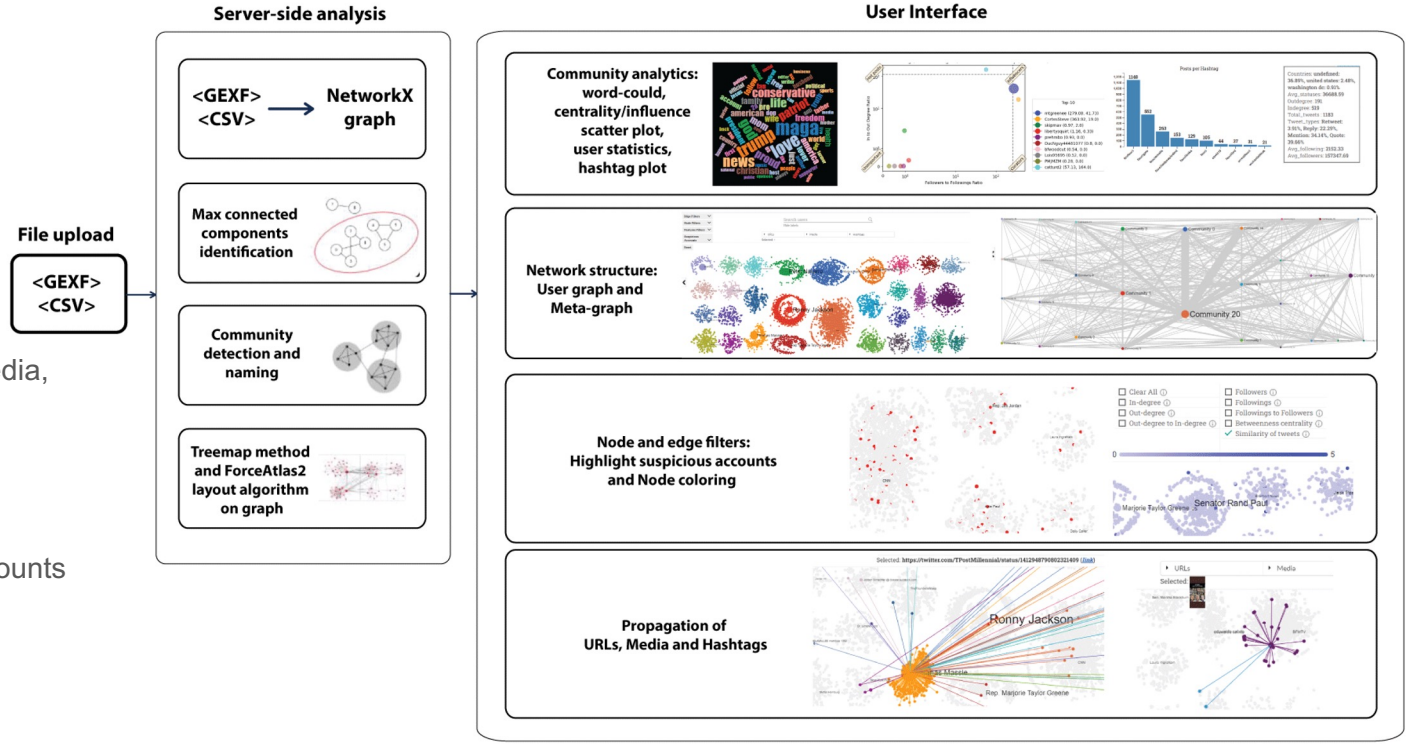
Information Propagation

Network Analysis and Visualization

MeVer NetworkX analysis and visualization tool helps users delve into **social media conversations**, helps users gain insights about how information propagates, and provides intuition about communities formed via **interactions**.

Features:

- Individual Account and Post Inspections
- Community Detection
- Community Analytics
- Metagraph
- Propagation of URLs, Media, and Hashtags
- Node and Edge Filters
- Node Coloring
- Highlight Suspicious Accounts



Network Analysis – Suspicious Accounts

- **Following rate** is the ratio of the number of followings to the number of days since an account was first created.
- **Status rate** is the ratio of the number of posts to the number of days since an account was created.
- **Average mentions per post** shows the average number of mentions in an account's tweets. A common strategy for spreading disinformation is mentioning many accounts in tweets.
- **Average mentions per word** shows the average number of mentions in a tweet's text. The tactic of posting tweets with many mentions and a single hashtag is often regarded as spam-like or suspicious. This feature is normalized to the total number of posts.
- **Average hashtags per word** calculates the average number of hashtags in a tweet's text.
- **Average URLs per word** calculates the average number of URLs in a tweet's text.

Network Analysis and Visualization - Analysis

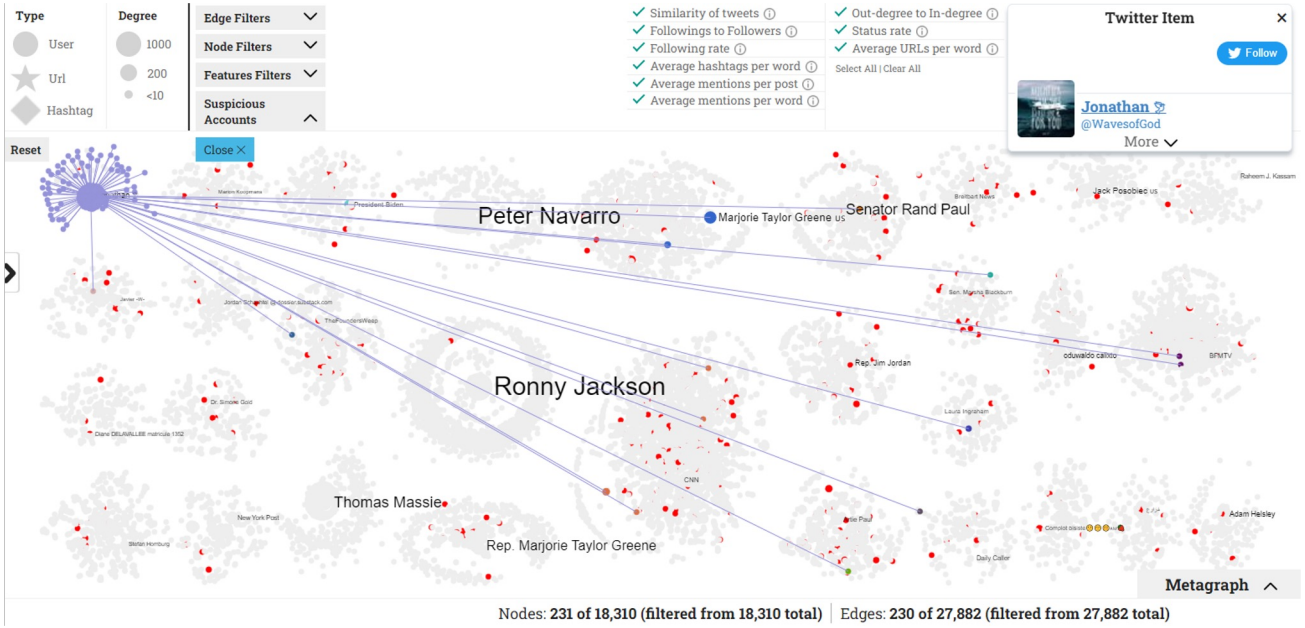
Four use cases to simulate a scenario in which, through the tool, an end user tries to identify and inspect suspicious accounts within a given dataset graph.

<u>Fauci use case</u>

425 accounts out of 18,310 highlighted as suspicious.

@WavesofGod (Community 14)

- Mentioned 228 other accounts in its tweets
- Mentioned popular accounts: President's Biden (POTUS) and Marjorie Taylor Greene ("mtgreenee").
- It is a strong supporter of Christianity and against vaccination.



This account has been suspended from Twitter

https://mever.iti.gr/networkx/

Papadopoulou, O., Makedas, T., Apostolidis, L., Poldi, F., Papadopoulos, S., & Kompatsiaris, I. (2022). MeVer NetworkX: Network Analysis and Visualization for Tracing Disinformation. Future Internet, 14(5), 147.

Key Challenges in AI against Disinformation

- Arms race nature of problem
  - Constantly improved AI models for synthetic media
  - New disinformation tactics
- Gap between research and end users
  - Journalists/fact-checkers need intuitive tools
  - Tools need to be robust and trustworthy
- Risks of discrimination
  - AI models might be biased against certain demographics
  - Use of AI models might be done in a biased manner
- Sustainability
  - AI costs a lot in terms of research, maintenance and deployment
  - Fighting disinformation is not a profitable business
- Required contribution from various disciplines
  - Content Analytics - NLP, Machine Learning, Network Analysis, Big Data Architectures,
  - Psychology – Social Sciences (patterns of presentation, sharing),Visualization

- **Continual learning**
- **Out-of-domain generalization**

- **Explainability**
- **Human agency**
- **Robustness**

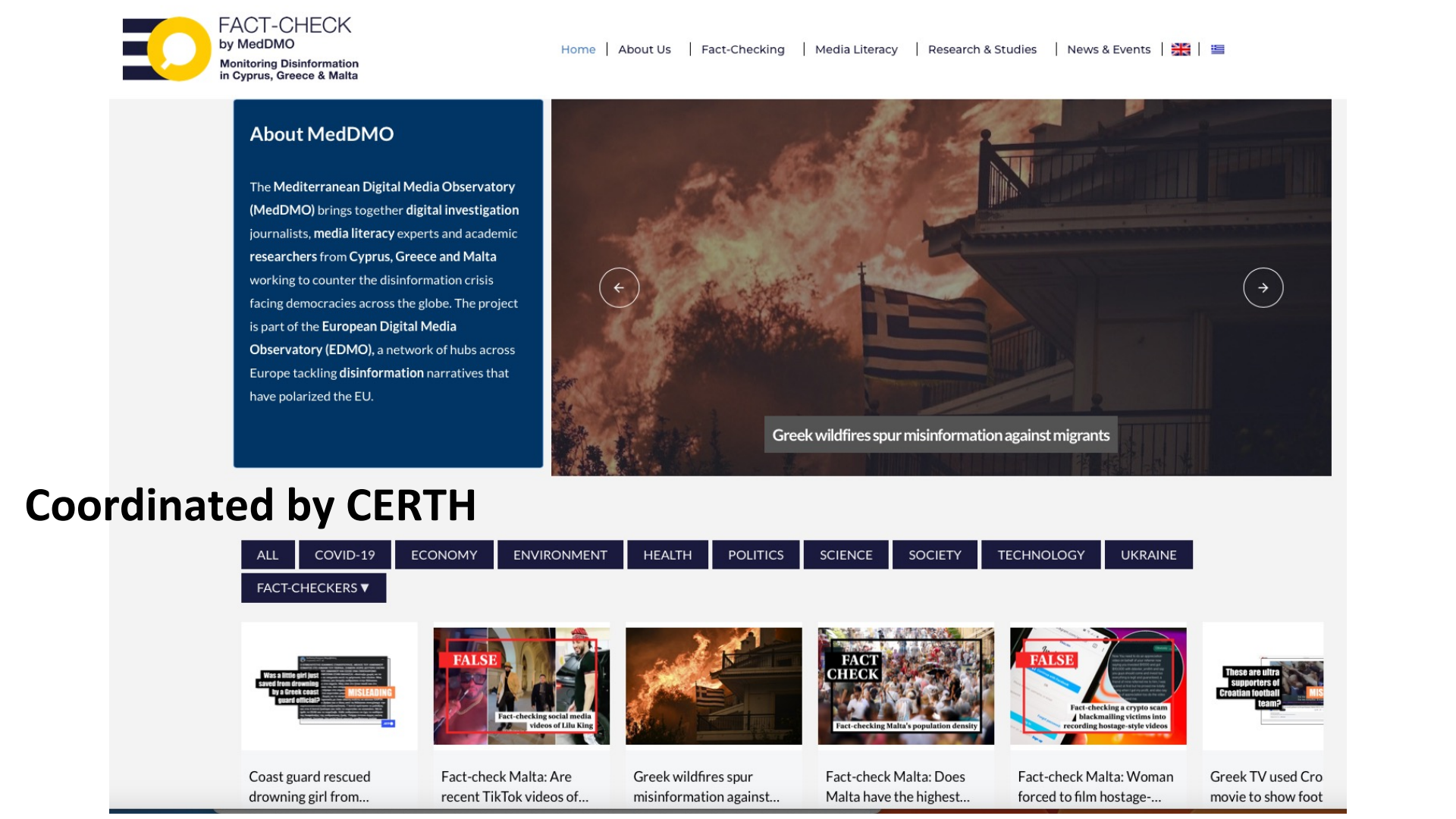
- **AI fairness**

- **Green AI**
- **Model compression**
- **Knowledge distillation**

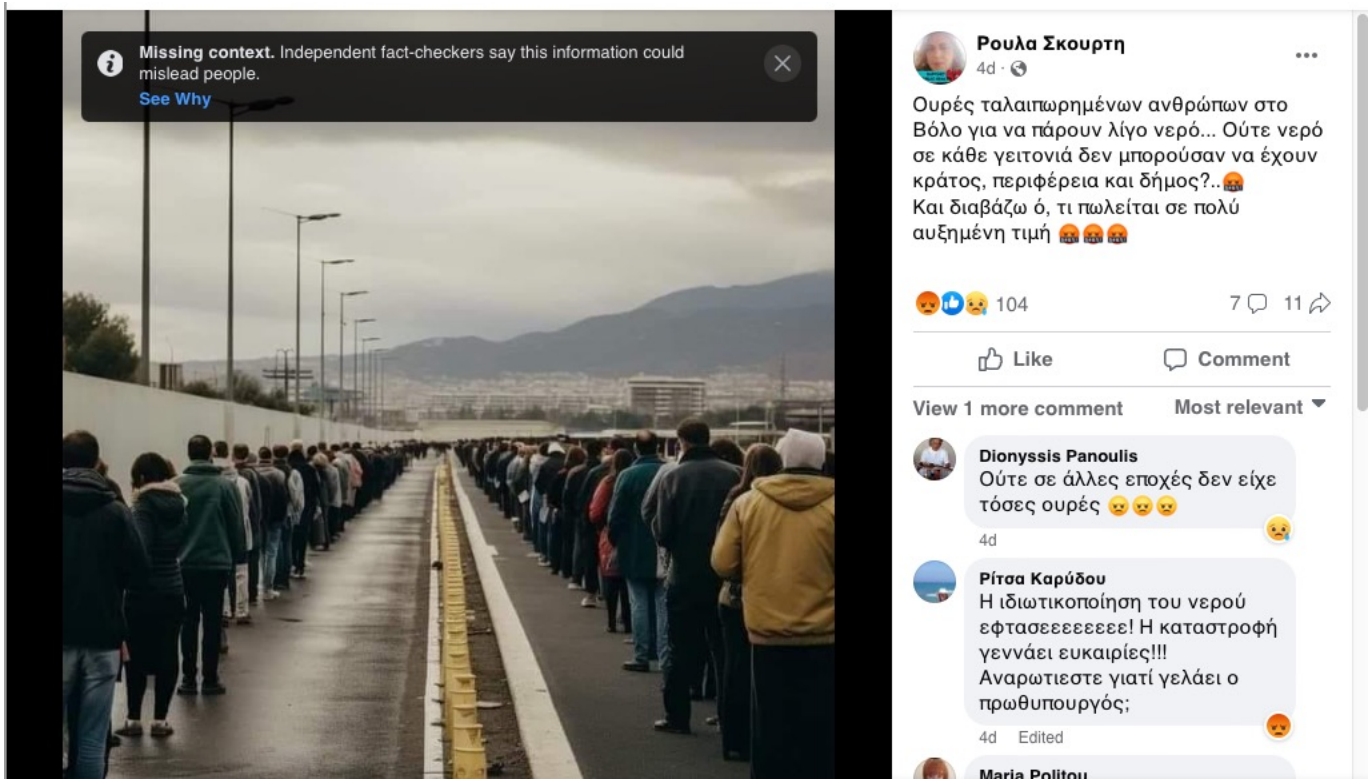
Fact checking - EFCSN just started

AFP

- 2017: One fact checker in Paris
- 2022: 130 fact checkers, 80 countries, 24 languages


FACT-CHECK
by MedDMO
Monitoring Disinformation
in Cyprus, Greece & Malta
Home | About Us | Fact-Checking | Media Literacy | Research & Studies | News & Events
About MedDMO
The Mediterranean Digital Media Observatory (MedDMO) brings together digital investigation journalists, media literacy experts and academic researchers from Cyprus, Greece and Malta working to counter the disinformation crisis facing democracies across the globe. The project is part of the European Digital Media Observatory (EDMO), a network of hubs across Europe tackling disinformation narratives that have polarized the EU.
Greek wildfires spur misinformation against migrants
Coordinated by CERTH
ALL
COVID-19
ECONOMY
ENVIRONMENT
HEALTH
POLITICS
SCIENCE
SOCIETY
TECHNOLOGY
UKRAINE
FACT-CHECKERS ▼
Was a little girl just saved from drowning by a Greek coast guard officer?
MISLEADING
FALSE
Fact-checking social media videos of Lilu King
FACT CHECK
Fact-checking Malta's population density
FALSE
Fact-checking a crypto scam blackmailing victims into recording hostage-style videos
These are ultra supporters of Croatian football team?
Coast guard rescued drowning girl from...
Fact-check Malta: Are recent TikTok videos of...
Greek wildfires spur misinformation against...
Fact-check Malta: Does Malta have the highest...
Fact-check Malta: Woman forced to film hostage-...
Greek TV used Cro movie to show foot

Another Midjourney creation



Why Tools and Fact Checking is not enough?

Human Behaviour related challenges

Cyberpsychology, Behavior, and Social Networking, Vol. 22, No. 6 | Rapid Communications

Free Access

What Debunking of Misinformation Does and Doesn't

Jeong-woo Jang, Eun-Ju Lee, and Soo Yun Shin

Published Online: 7 Jun 2019 | https://doi.org/10.1089/cyber.2018.0608

… **participants' agreement with the news position was not attenuated by the explicit post hoc correction**. Considering that information is judged as truth when it meets intuitive evaluation criteria (e.g., familiarity, compatibility with existing knowledge)… **debunking its falsehood may not be sufficient to undo it**.

Cyberpsychology, Behavior, and Social Networking > Vol. 25, No. 10 > Original Articles Free Access

How Health-Related Misinformation Spreads Across the Internet: Evidence for the "Typhoon Eye" Effect

Lei Zheng, Jincheng Cai, Fang Wang, Chenhan Ruan, Mingxing Xu, and Miao Miao

Published Online: 11 Oct 2022 | https://doi.org/10.1089/cyber.2022.0047

… **Our results highlight the importance of psychological approaches to understanding the propagation patterns of health-related misinformation.** The present findings provide a new perspective for development of prevention and control strategies to reduce the spread of health-related misinformation during pandemics …

… The more **moral outrageous language you use, the more inflammatory language, contemptuous language, the more indignation you use, the more it will get shared.** So we are being rewarded for being division entrepreneurs. The better you are at innovating a new way to be divisive, we will pay you in more likes, followers and retweets.…

Role of Social media platforms

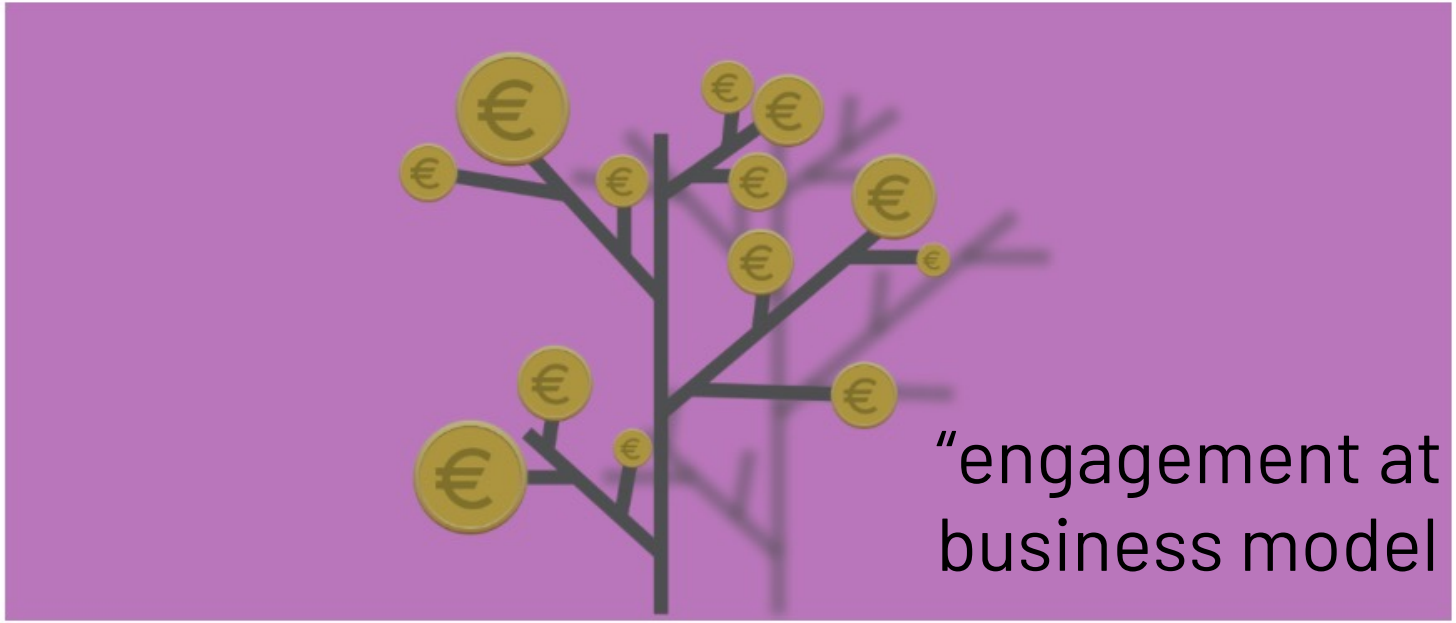
JANUARY 18, 2023

Study reveals the key reason why fake news spreads on social media

by University of Southern California

Credit: Pixabay/CC0 Public Domain

..."As it turns out, much like any video game, **social media has a rewards system** that encourages users to stay on their accounts and keep posting and sharing. **Users who post and share frequently, especially sensational, eye-catching information, are likely to attract attention"**…



“engagement at all costs” business model

News | Training | July 31, 2023

EDMO Online Training: The Economics of Disinformation

“Making people hasn’t been that quick since Eden”

Tencent Cloud announces Deepfakes-as-a-Service for $145

Three minutes of video, 100 sentences of speech, and 24 hours gets you a bot to front your livestreams and answer questions

Laura Dobberstein — Fri 28 Apr 2023 // 03:58 UTC

Tencent Cloud has announced it's offering a digital human production platform – essentially Deepfakes-as-a-Service (DFaaS).

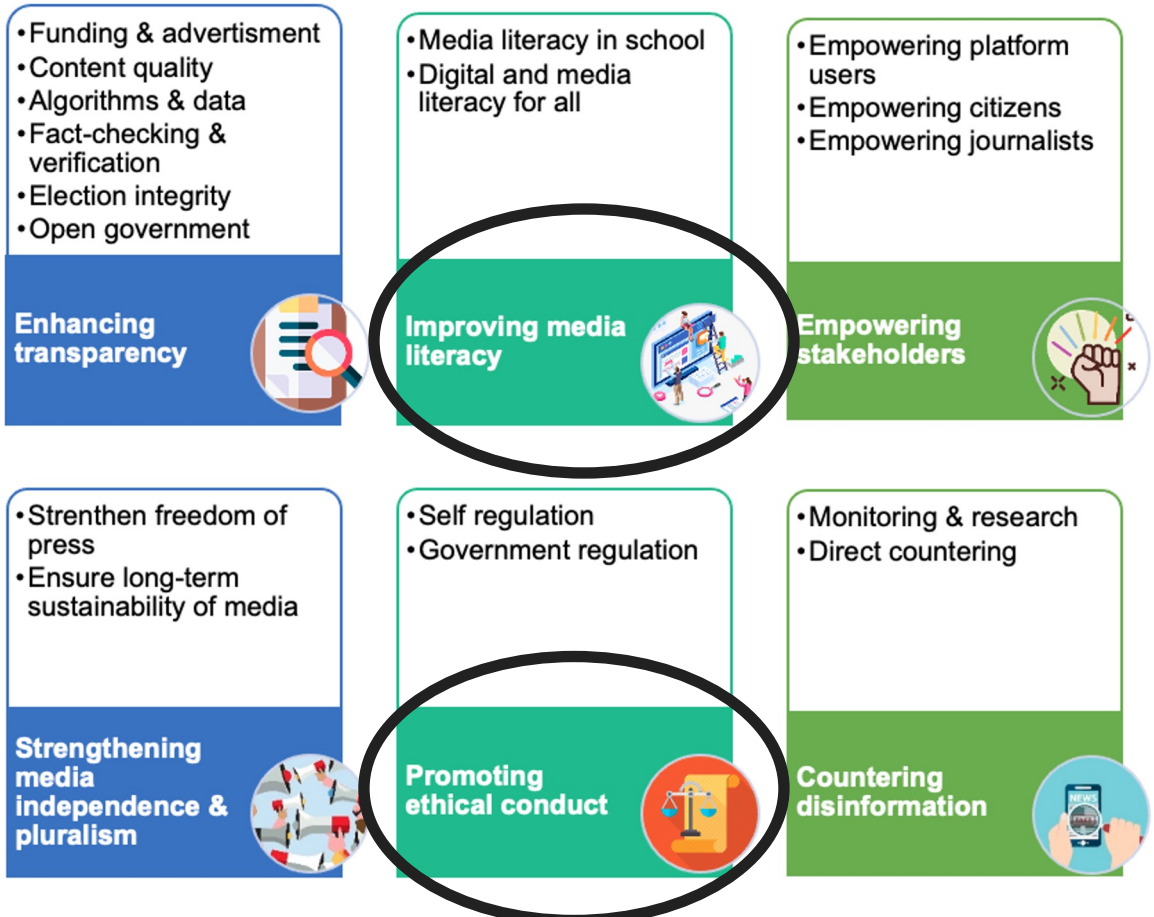
According to Chinese media and confirmed to *The Reg* by Tencent, the service needs just three minutes of live-action video and 100 spoken sentences – and a $145 fee – to create a high-definition digital human.

Gestating the creation requires just 24 hours. Making people hasn't been that quick since Eden.

The digital characters are available in half bodies or full bodies, and the service is available in both Chinese and English.

https://www.theregister.com/2023/04/28/tencent_digital_humans/

Counter-disinformation policies classification



PRESS RELEASE | Publication 11 October 2022

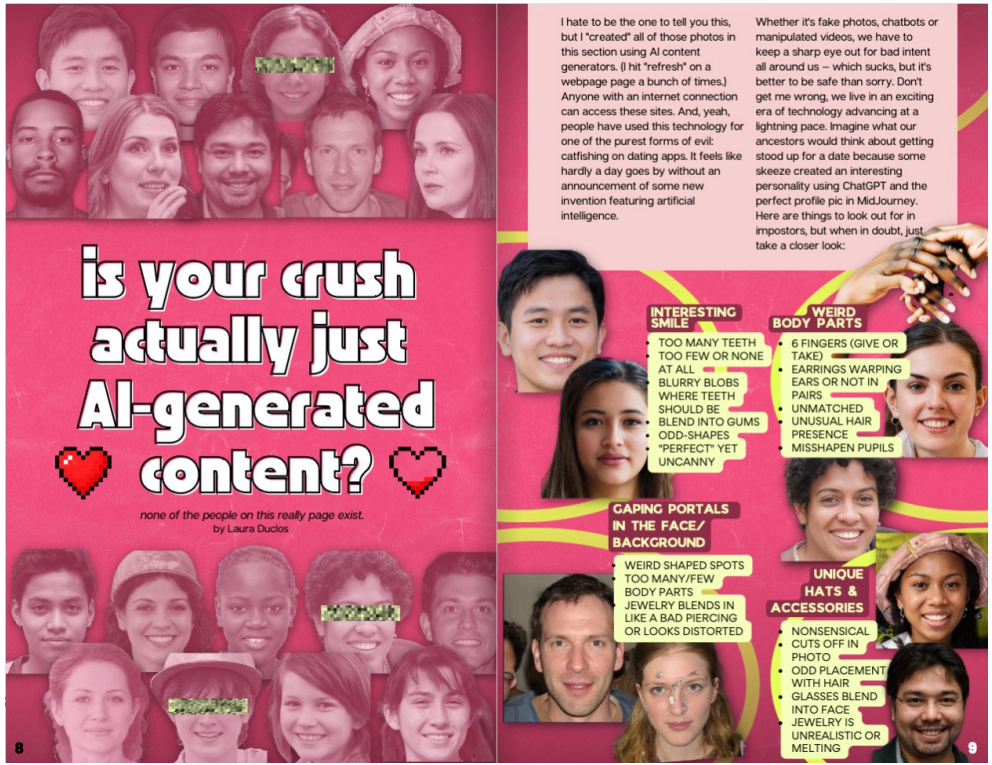
Commission steps up action to tackle disinformation and promote digital literacy among young people

The Commission has published Guidelines for teachers and educators in primary and secondary schools, on how to address disinformation and promote digital literacy in their classrooms.

The guidelines provide practical support for teachers and educators and include definitions of technical concepts, class-exercises and how to encourage healthy online habits. This toolkit covers three main topics: building digital literacy, tackling disinformation, and assessing and evaluating digital literacy.

Full press release

Media literacy to reach younger audiences



is your crush actually just AI-generated content?

none of the people on this really page exist.
by Laura Duclos

I hate to be the one to tell you this, but I "created" all of those photos in this section using AI content generators. (I hit "refresh" on a webpage page a bunch of times.) Anyone with an internet connection can access these sites. And, yeah, people have used this technology for one of the purest forms of evil: catfishing on dating apps. It feels like hardly a day goes by without an announcement of some new invention featuring artificial intelligence.

Whether it's fake photos, chatbots or manipulated videos, we have to keep a sharp eye out for bad intent all around us – which sucks, but it's better to be safe than sorry. Don't get me wrong, we live in an exciting era of technology advancing at a lightning pace. Imagine what our ancestors would think about getting stood up for a date because some skeeze created an interesting personality using ChatGPT and the perfect profile pic in MidJourney. Here are things to look out for in impostors, but when in doubt, just take a closer look:

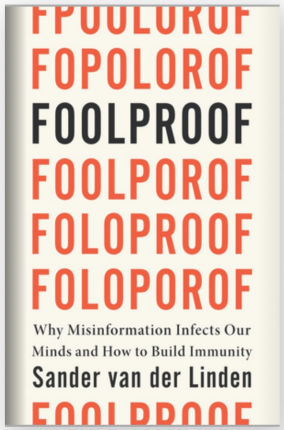
INTERESTING SMILE
- TOO MANY TEETH
- TOO FEW OR NONE AT ALL
- BLURRY BLOBS WHERE TEETH SHOULD BE
- BLEND INTO GUMS
- ODD-SHAPES
- "PERFECT" YET UNCANNY

WEIRD BODY PARTS
- 6 FINGERS (GIVE OR TAKE)
- EARRINGS WARPING EARS OR NOT IN PAIRS
- UNMATCHED
- UNUSUAL HAIR PRESENCE
- MISSHAPEN PUPILS

GAPING PORTALS IN THE FACE/ BACKGROUND
- WEIRD SHAPED SPOTS
- TOO MANY/FEW BODY PARTS
- JEWELRY BLENDS IN LIKE A BAD PIERCING OR LOOKS DISTORTED

UNIQUE HATS & ACCESSORIES
- NONSENSICAL
- CUTS OFF IN PHOTO
- ODD PLACEMENT WITH HAIR
- GLASSES BLEND INTO FACE
- JEWELRY IS UNREALISTIC OR MELTING

Build Immunity against Disinformation

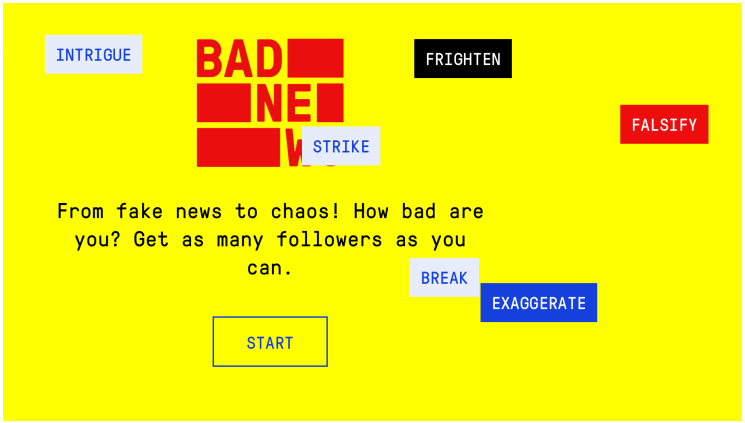


Foolproof

Why Misinformation Infects Our Minds and How to Build Immunity

*by **Sander van der Linden** (Author, University of Cambridge)*

A *Next Big Idea Club* Must-Read • A *Financial Times* Best Book of the Week • One of *Nature's* best science picks



The country inoculating against disinformation

A cyber-attack that spread disinformation in Estonia in 2007 led to violent protests on the streets of the capital Tallinn (Credit: Raigo Pajular/AFP/Getty Images)

By Amy Yee 31st January 2022

Subjected to repeated disinformation campaigns, the tiny Baltic country of Estonia sees media literacy education as part of its digital-first culture and national security.

For two days riots raged in Estonia's capital Tallinn. Protestors clashed with police and looters rampaged after the violence was sparked by **controversy** about a decision to move a military statue erected during Soviet rule. The flames of outrage among Estonia's Russian-speaking minority were fanned by false news spreading online and in Russian news reports.

The disinformation campaign then escalated into what is considered the first **cyber-attack against an entire country**. The attack, which was linked to Russia, shut down websites of Estonia's government, banks and media outlets.

In the aftermath of the attack in 2007, Estonia decided to take action. The country has now become a cyber-security leader, aimed at protecting its online infrastructure from future attacks.

https://www.bbc.com/future/article/20220128-the-country-inoculating-against-disinformation

Counter-disinformation policies

Big Tech must deal with disinformation or face fines, says EU

By Tom Gerken & Liv McMahon
Technology Team

2 days ago

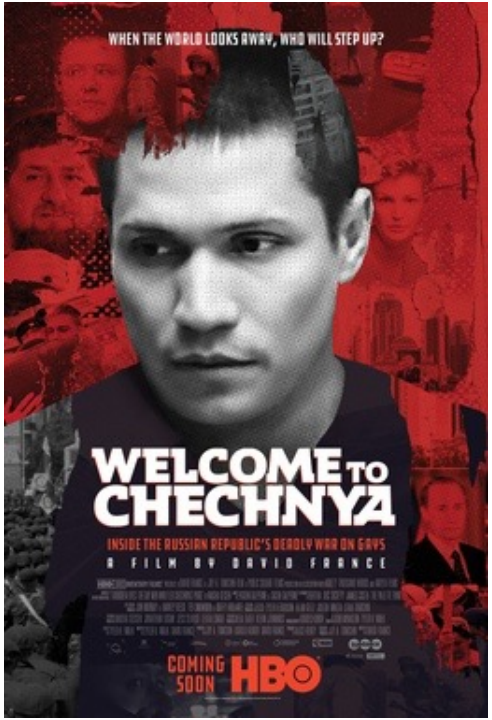
https://www.bbc.com/news/technology-61817647

The EU is working in close cooperation with online platforms to encourage them to promote authoritative sources, demote content that is fact-checked as false or misleading, and take down illegal content or content that could cause physical harm.

Positive Applications of DeepFakes

Protecting the Identity of Interviewees

- Welcome to Chechnya is a 2020 documentary film by David France
- The film centres on the anti-gay purges in Chechnya of the late 2010s, filming LGBT Chechen refugees using hidden cameras as they made their way out of Russia
- This was the first film to use DeepFake technologies to protect the identities of speakers



https://en.wikipedia.org/wiki/Welcome_to_Chechnya

Animating Faces from the Past

- DeepNostalgia by my Heritage makes it possible to create animations from still photos, e.g. of historical figures or beloved ones

Dalí Lives

https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum

Reimagining Historical Events

https://moondisaster.org/film

Sensitizing about Global Issues

- DeepEmpathy (https://deepempathy.mit.edu/): *increase empathy by making our homes appear similar to the homes of victims of disasters/wars*

- Using DF to Visualize Impacts of Climate Change

Luccioni, A., Schmidt, V., Vardanyan, V., & Bengio, Y. (2021). Using Artificial Intelligence to Visualize the Impacts of Climate Change. IEEE Computer Graphics and Applications, 41(1), 8-14.

FIGURE 2. Before and after example of the transformation carried out by our GAN approach, based on an image of a suburban house.

Related projects

With contributions from

Symeon Papadopoulos papadop@iti.gr

Senior Researcher, CERTH-ITI

MeVer Group Leader

Nikos Sarris nsarris@iti.gr

Senior Researcher

Thank you for your attention!

Get in touch!

Yiannis Kompatsiaris ikom@iti.gr

https://mklab.iti.gr