Deepfakes Analysis Unit

An MCA Initiative

First Quarter Report 2025

April 1, 2025 - June 30, 2025

Index

What is the Deepfakes Analysis Unit (DAU)

- The Deepfakes Analysis Unit (DAU) is a collaborative task force set up by the Misinformation Combat Alliance (MCA), to ensure timely response and research on the emerging crisis of AI generated video and audio.
- The primary public touchpoint is a WhatsApp tipline that allows the public to escalate purported synthetic audio and video to the unit for further investigation and action.
- The unit is helmed by an independent secretariat within the MCA that analyses content, escalates it to external forensic and detection experts, publishes assessment reports, and coordinates responses on the tipline as well as with fact-checkers.
- The DAU also has a separate escalation channel for fact-checkers in India. Earlier this year, it opened up a similar channel for IFCN-certified fact-checkers from across the globe.
- DAU also produces content for literacy around AI generated videos. DAU aims to be a trusted resource for the public, fact-checkers, and media, that detects and responds to misleading and harmful A.I.-generated audio and video.
- The seed funding to launch the DAU was provided by Meta.

Who are DAU's Fact-Checking Partners

BOOM

FACT Crescendo

FACTLY

FACT CHECK AFWA

LF LOGICALLY FACTS

विश्वास.News

NEWSMOBILE

the quint

THIP THE HEALTHY INDIAN PROJECT

NewsMeter

newschecker

TELUGU POST

Who are DAU's Forensic & Tech Partners

Detection and Forensic Partners

DigitID
You are the Key®
to your secure world

deeptrust

GetReal

VALIDIA

RIT

Technical and Research Partner

Tattle

What is Misinformation Combat Alliance

- The Misinformation Combat Alliance, is a Section 8 not-for-profit corporation in India set up with the aim of collectively fighting misinformation and its impact.

- It is a cross-industry alliance bringing companies, organisations, institutions, industry associations and government entities together to combat misinformation and fake news and create an enlightened and informed society.

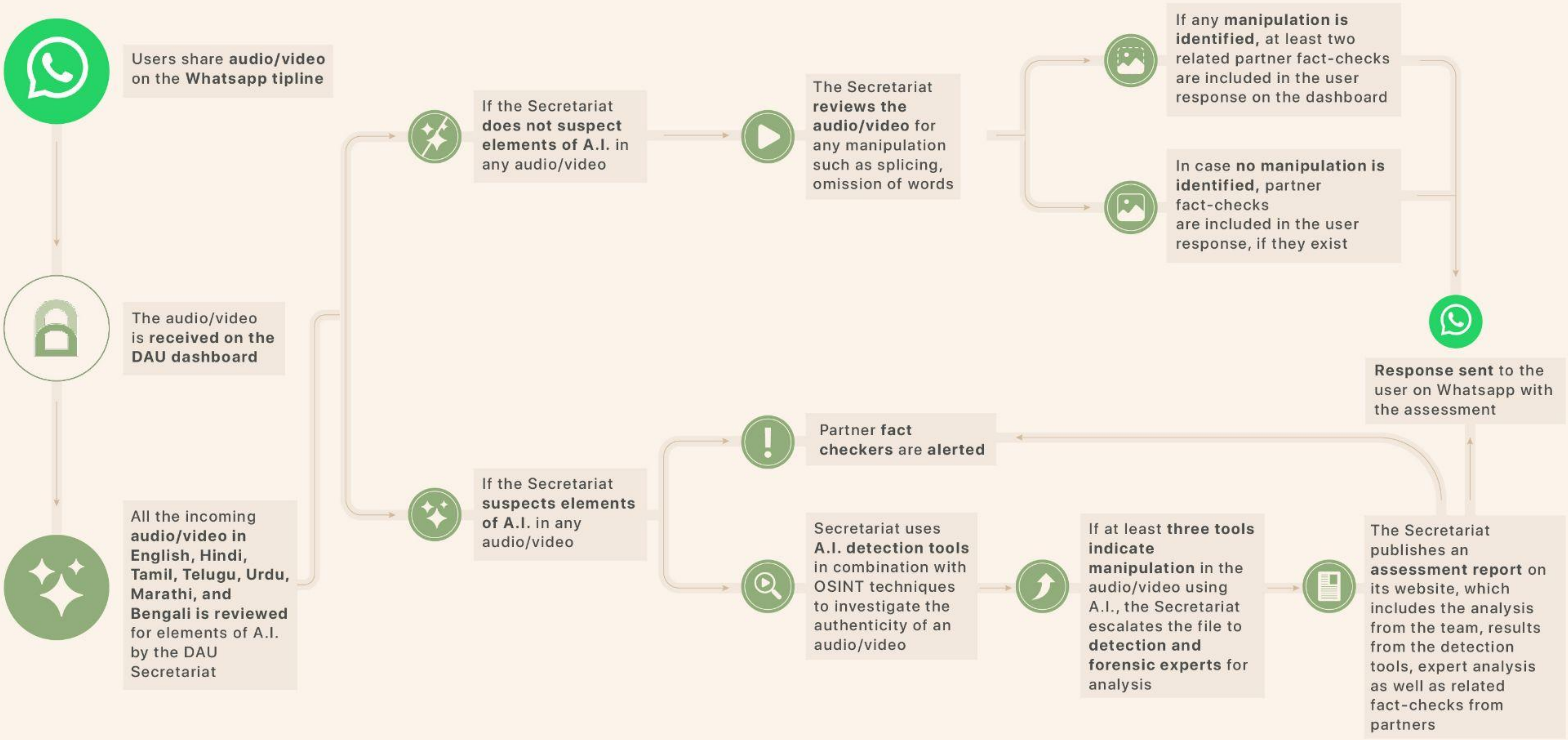
- The MCA at present has 11 members from fact-checking, media and civic tech organisations.

How the DAU Works

- Every message sent to the WhatsApp number that contains a URL, or a video or audio file is sent to a dashboard.

- The DAU secretariat looks at every media item on the dashboard. Images and text messages are filtered out and not sent to the dashboard. See the following page for a detailed description of the DAU's process.

- If the audio/video content fits within the DAU's remit (see next section), the secretariat conducts preliminary assessment using a manual review process as well as tool analysis. It checks for signs of AI manipulation as well as generation, such as imperfect lip-sync and oddities in facial features of a featured subject. It consults with experts to get further insights, if the initial investigation points to AI elements in the content.



Users share **audio/video** on the **Whatsapp tipline**

The audio/video is **received on the DAU dashboard**

All the incoming **audio/video in English, Hindi, Tamil, Telugu, Urdu, Marathi, and Bengali is reviewed** for elements of A.I. by the DAU Secretariat

If the Secretariat **does not suspect elements of A.I.** in any audio/video

The Secretariat **reviews the audio/video** for any manipulation such as splicing, omission of words

If any **manipulation is identified,** at least two related partner fact-checks are included in the user response on the dashboard

In case **no manipulation is identified,** partner fact-checks are included in the user response, if they exist

Response sent to the user on Whatsapp with the assessment

If the Secretariat **suspects elements of A.I.** in any audio/video

Partner **fact checkers** are **alerted**

Secretariat uses **A.I. detection tools** in combination with OSINT techniques to investigate the authenticity of an audio/video

If at least **three tools indicate manipulation** in the audio/video using A.I., the Secretariat escalates the file to **detection and forensic experts** for analysis

The Secretariat publishes an **assessment report** on its website, which includes the analysis from the team, results from the detection tools, expert analysis as well as related fact-checks from partners

What Does the DAU Check?

- At present the DAU only verifies audio and video content in Hindi, English, Tamil, Telugu, Urdu, Marathi and Bengali.

- Furthermore, content that is not in the public interest or is private in nature is not assessed and is considered beyond the scope of the DAU's focus.

- Pornographic or sexually explicit videos, however, go through a preliminary check by the secretariat to rule out the possibility of that video being Non-Consensual Sexual Imagery (NCSI).

- One of the three labels is assigned to an audio or video that falls outside the purview of DAU's focus:
    - Spam
    - Out of Scope
    - Unsupported Language

Motivation for this Report

- AI-generated content, including deepfakes, are an emerging challenge to a healthy online conversation.

- There are several speculations about the volume of AI-generated content online, its impact on the spread of misinformation and the ways in which the technology may be misused.

- This report is an attempt to add evidence to understand how AI-generated content is evolving and affecting online discourse.

- This is the first report of 2025, corresponding to the fifth quarter of the Deepfakes Analysis Unit's operation. To read more about how the DAU works, please visit the DAU website or read reports on previous quarters.

Scope

- This report provides aggregate statistics regarding the content that the DAU saw on its dashboard during the fifth quarter of its operations, corresponding from 1 April, 2025 to 30 June, 2025. Previous quarters' report can be found on the DAU website.

- The number of media items that the DAU sees on the dashboard is smaller than the number of messages that are sent to the WhatsApp number- text messages and images are filtered out and not analyzed by the DAU.

- This report provides aggregate statistics on:
  - the media type (audio, video, URL)
  - language of the content
  - the nature of manipulation, if manipulated (AI generated, manipulated, not manipulated, cheapfake, deepfake)
  - the broad themes in the content

- Engagement with content produced by the DAU, be it assessment reports or the media literacy videos, are not covered in this report.

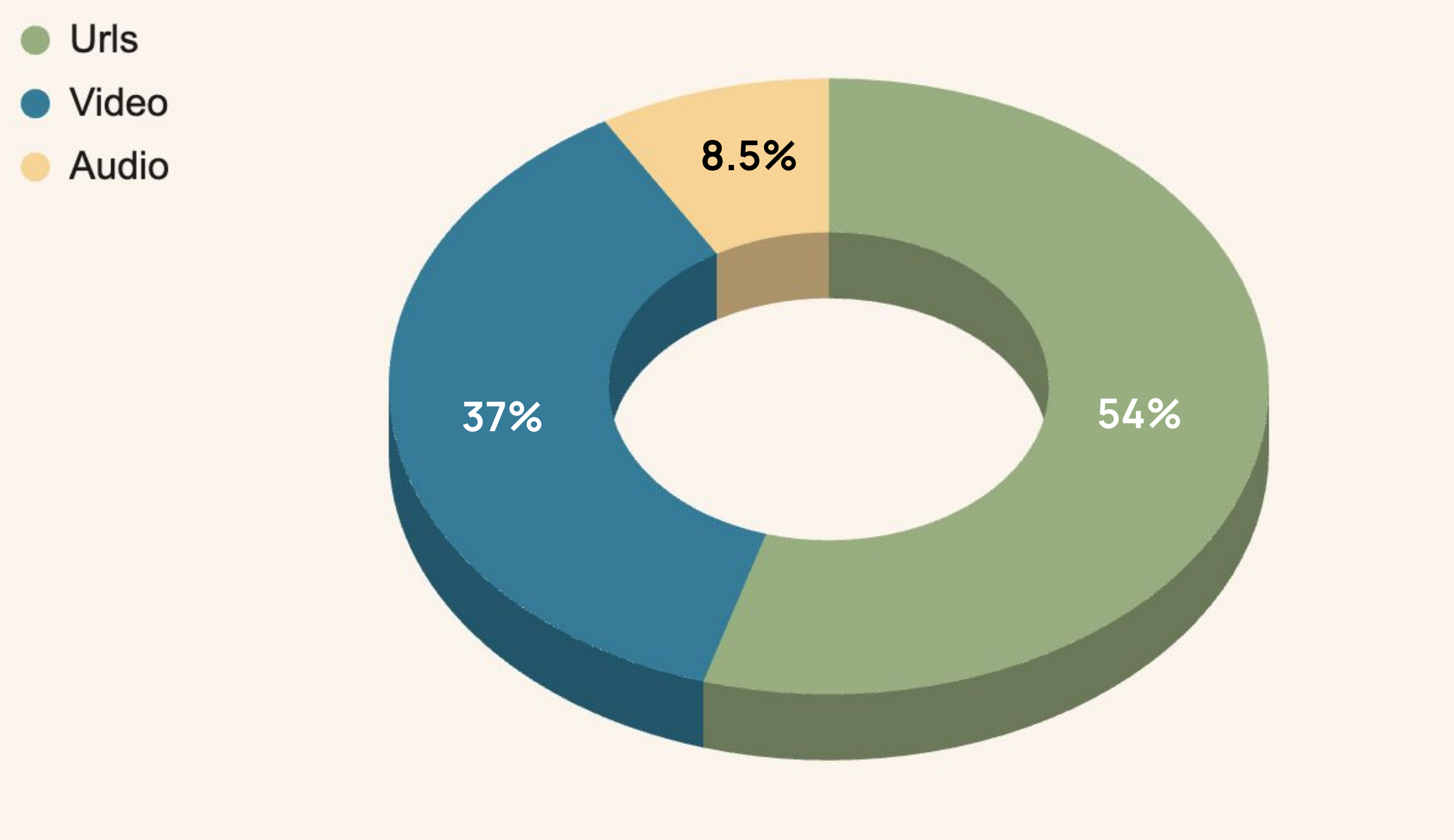
Limitations

- This report only shows data from one tipline focused specifically on deepfakes and content manipulated by A.I..
- There are several other tiplines run by fact checking groups that, while not dedicated to AI generated content, also receive it.
- The content hitting the DAU tipline should always be assumed to be a small percentage of the total AI generated content circulating online.
- The DAU database is not an exhaustive repository of all AI generated misinformation.
- Who uses the tipline, and how much the tipline is used depends on a number of factors such as ad campaigns to popularize the tipline, and the events at a time.
- For example, during a topical event where a lot of misinformation is produced, the tipline may be used more.

Aggregate Statistics

Please see Appendix B for Notes on Methodology

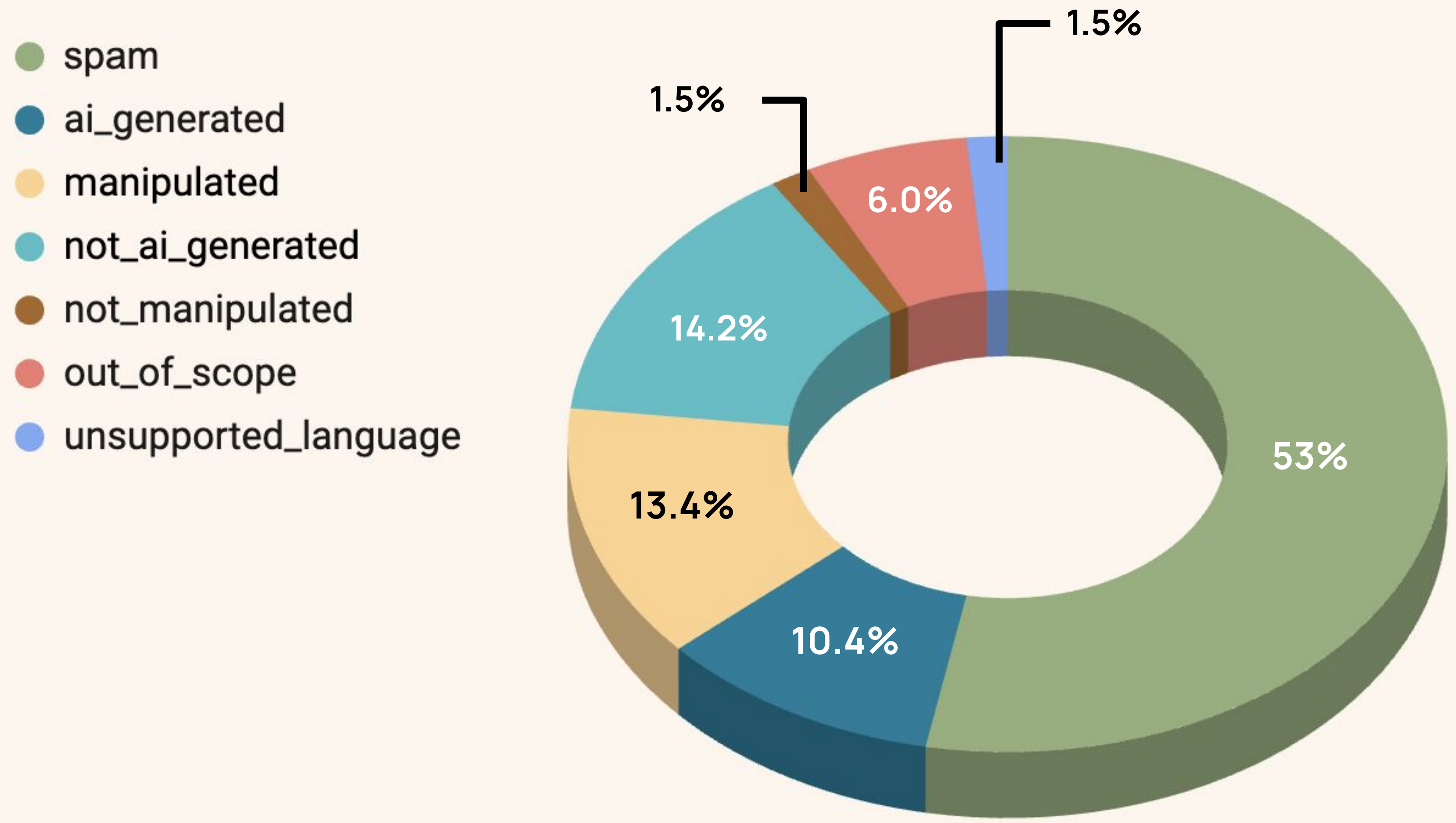
Type of Media Received



- Half (54.6%) of the content was shared as URLs, indicating that users mostly submitted links.
- Direct uploads of video and audio made up 37% and 8.5% respectively, showing a preference for link-sharing over file submissions.

Total number of media items received: 141 (including spam)

Verification Status Assigned

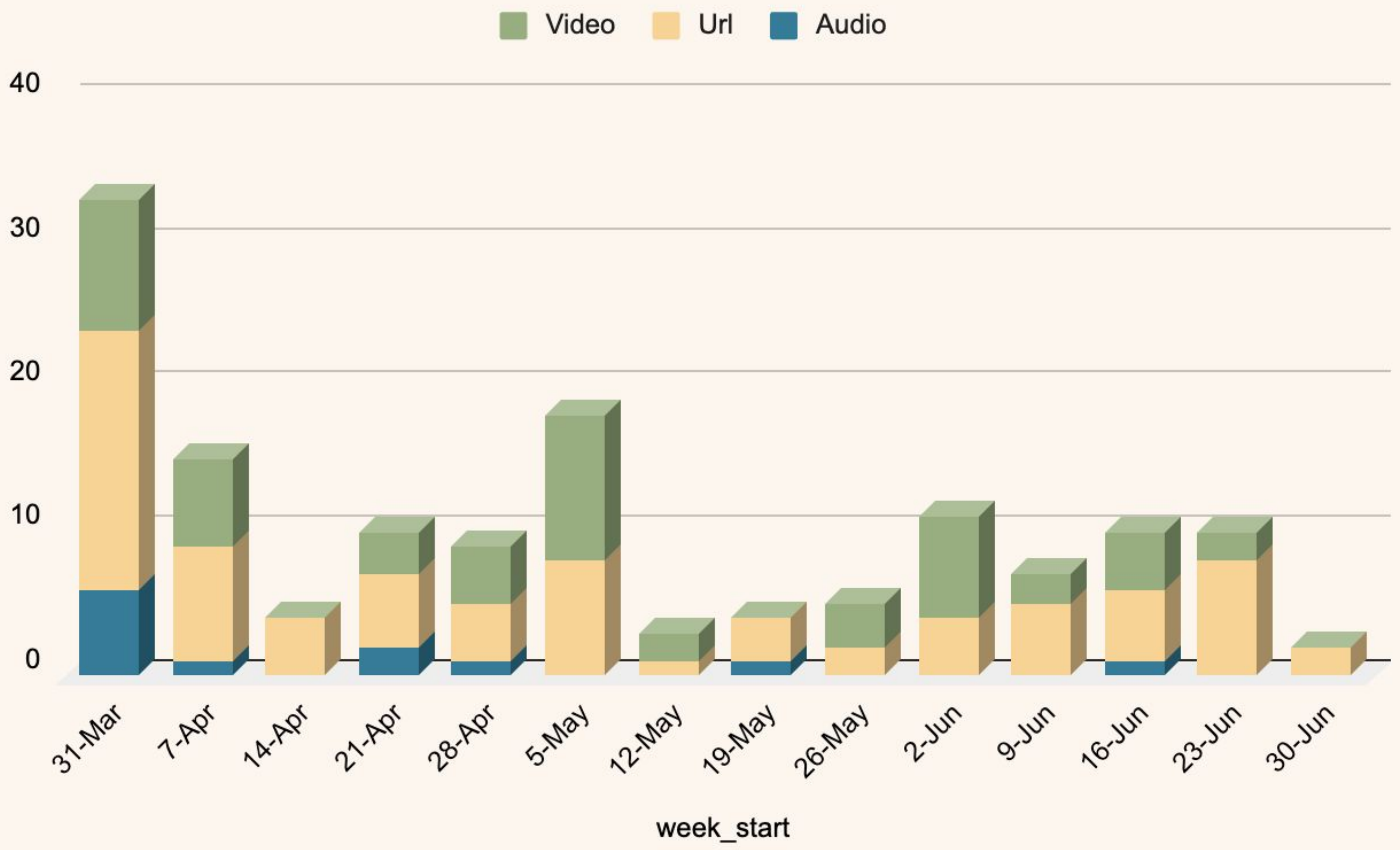


53% of the media items that hit the tipline were marked as Spam. Of the remaining:

- 19 items were labeled as Not AI-generated
- 14 items were labeled as AI-generated
- 18 items were labeled as manipulated
- No media item was labeled as Deepfake*

Media items labelled as Spam: 71
Media items assigned a label other than Spam: 63
Media items with a non-functional Url: 07 (Check Appendix B)
*One assessment report published this quarter concludes three videos as deepfakes, however, these videos were not received via tipline.

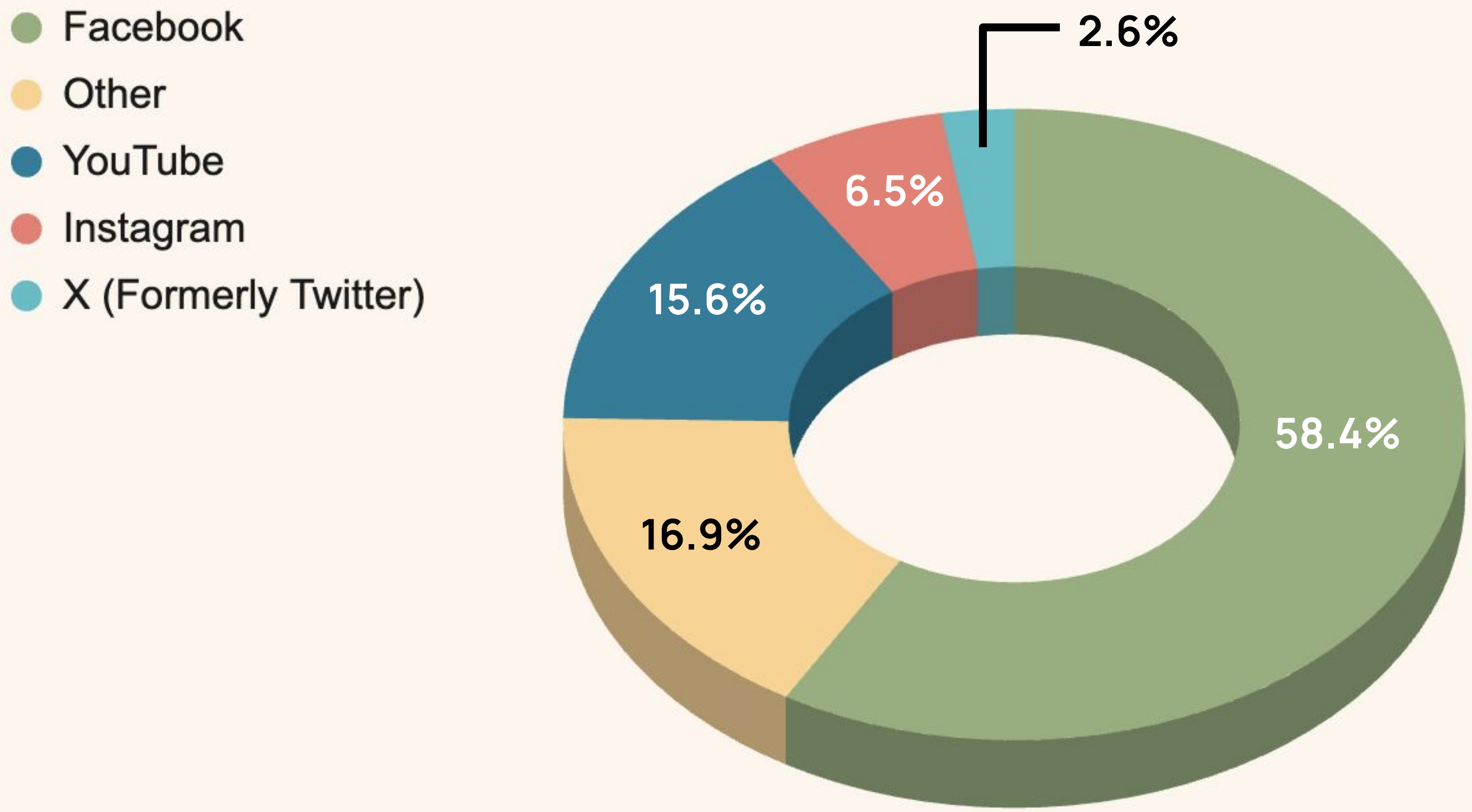
Media Type Received by Week



The highest number of URLs were recorded during the week of March 31: 18 URLs—more than double the weekly average.

April 1, 2025 to June 30, 2025

Domains Contained in URLs



- Nearly 58.4% of all URLs came from Facebook, making it the most common source by a wide margin. It could be because of ad campaigns run by the DAU on Meta.

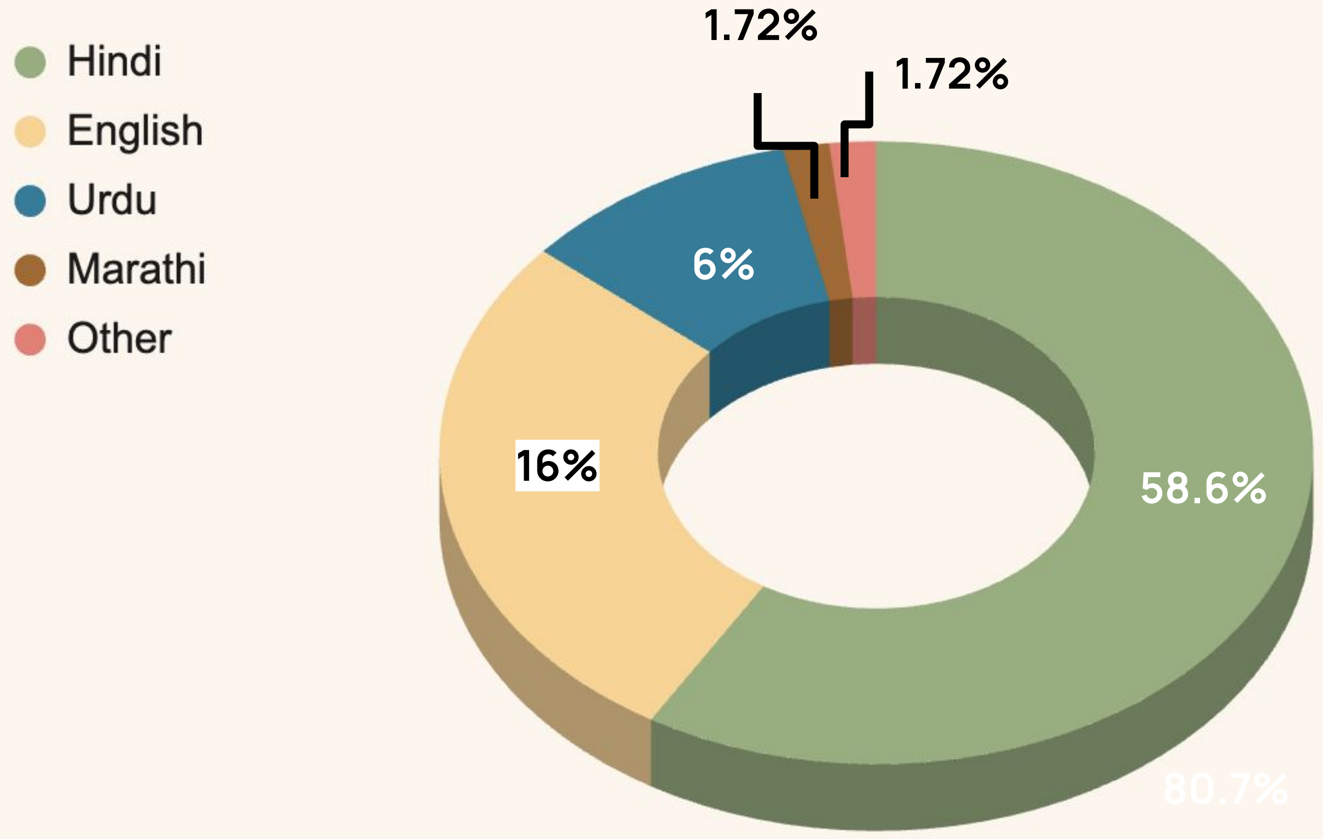
- Platforms like Instagram, YouTube, and X (formerly Twitter) contributed to 24% of all URLs.

- The “Other” category accounted for around 17% URLs, mostly spammy or low-quality domains such as 'm.par' 'hi.hello' 'goo.gl' etc.

Total Number of URLs: 77 (including spam)

April 1, 2025 to June 30, 2025

Language of Media Shared*



- The DAU added language tags (for seven supported languages) manually in media items that were not marked as spam.
- Over 58% of the media items (34) were in Hindi.
- 16 media items were in English and 06 items were in Urdu.
- DAU didn't receive any media in Tamil, Telugu and Bengali in Q1 2025 and only one in Marathi.
- Media without any associated language was labeled as "Other"

*These numbers do not reflect the language in which the user chose to interact with the tipline

Number of Media Items in DAU purview with a Language Label: 58

12 media items were not given a language tag as they had an unsupported language

April 1, 2025 to June 30, 2025

Themes in the Content Received

Here are some highlights, trends and themes we found in the media reviewed by the DAU:

- Bulk of the content verified by the DAU in this quarter comprised A.I.-manipulated videos with a few deepfakes that surfaced during and after the India-Pakistan conflict that took place in May.

- Most of the A.I.-manipulated videos pertained to content peddling supposed "automatic" or "A.I.-powered" get-rich-quick financial platforms or schemes promising high returns on investment. This has been a continuing trend, which has been highlighted in the previous quarterly reports as well.

- The manipulations in the financial scam videos involved replacing original audio tracks from authentic videos with A.I.-generated audios to make it appear that politicians, top business leaders, and celebrities were endorsing financial platforms or schemes when they were not. Most of these videos were packaged as interviews from news segments, they opened with visuals of popular television journalists.

- Among politicians, India's Finance Minister Nirmala Sitharaman was frequently linked to such scam videos. Anant Ambani from Reliance Industries and N.R. Narayana Murthy, co-founder of tech giant Infosys, were the most targeted business leaders. We also debunked one such video that apparently featured Bollywood actor Shah Rukh Khan.

Continued to next page

Themes in the Content Received

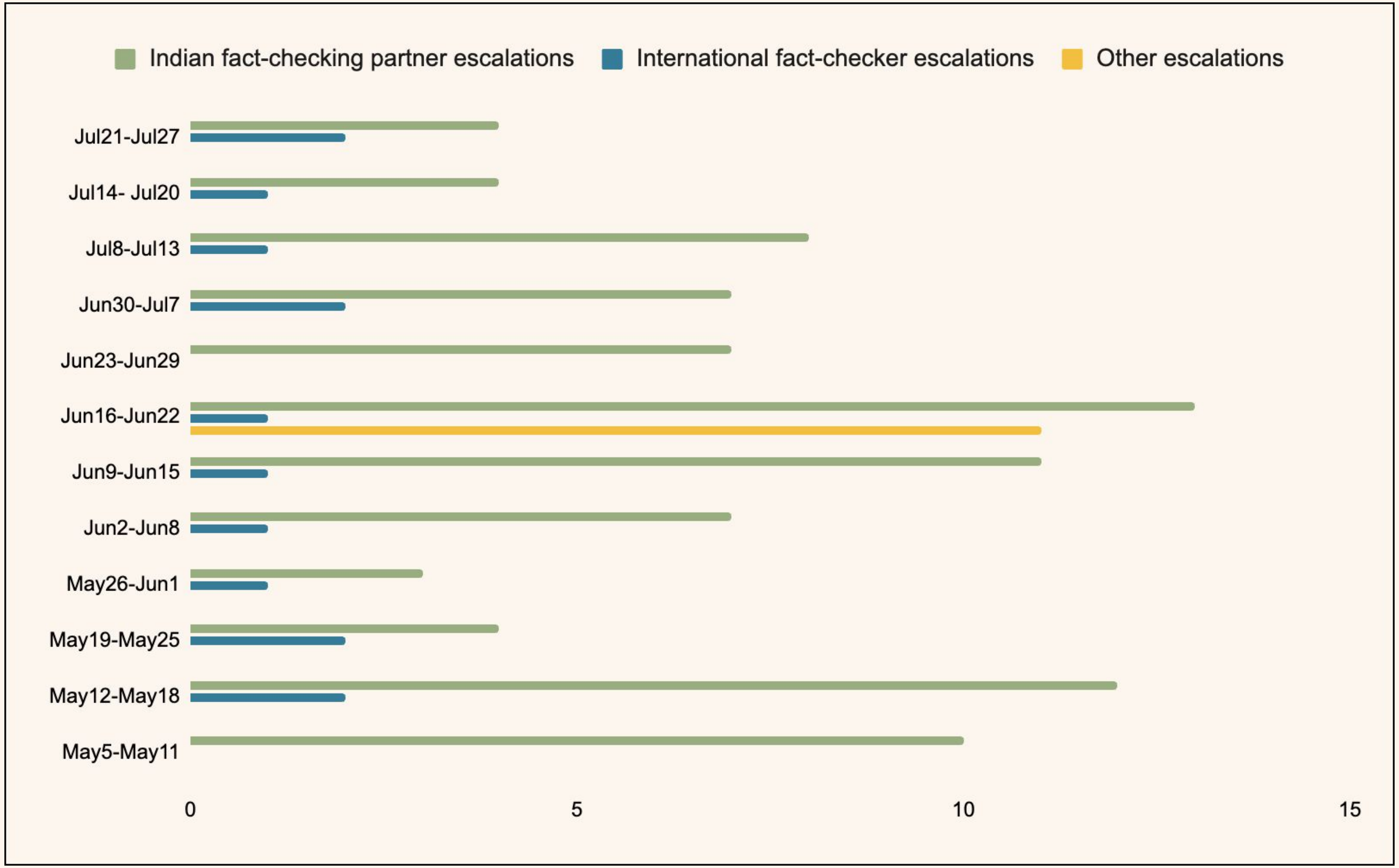
- We spoke to experts to understand why most of the financial scam videos "recommended" an "initial investment" amount of 21,000 rupees.

- The fabricated videos on the India-Pakistan conflict peddled narratives about defeat and loss. One such A.I.-manipulated video apparently showed that Pakistan’s Prime Minister Shehbaz Sharif was admitting defeat against India.

- Three separate deepfake videos, apparently featuring Prime Minister Narendra Modi, External Affairs Minister Dr. S.Jaishankar, and Home Minister Amit Shah, respectively, were in circulation around the same time. Each of these videos were similar in packaging as well as messaging, through the audio tracks they conveyed that India was conceding to Pakistan.

- Another viral A.I.-manipulated video that the DAU debunked apparently featured Indian Army officer Col. Sofiya Qureshi, one of two female officers, who led the media briefing after India’s military attacked Pakistan. This fabricated video tried to invoke her faith using synthetic audio with original video footage from the media briefing.

Media escalated to the DAU by fact-checking partners and international fact-checkers



- 90 media items, comprising of audios, videos and images, were escalated to the DAU by Indian fact checkers.

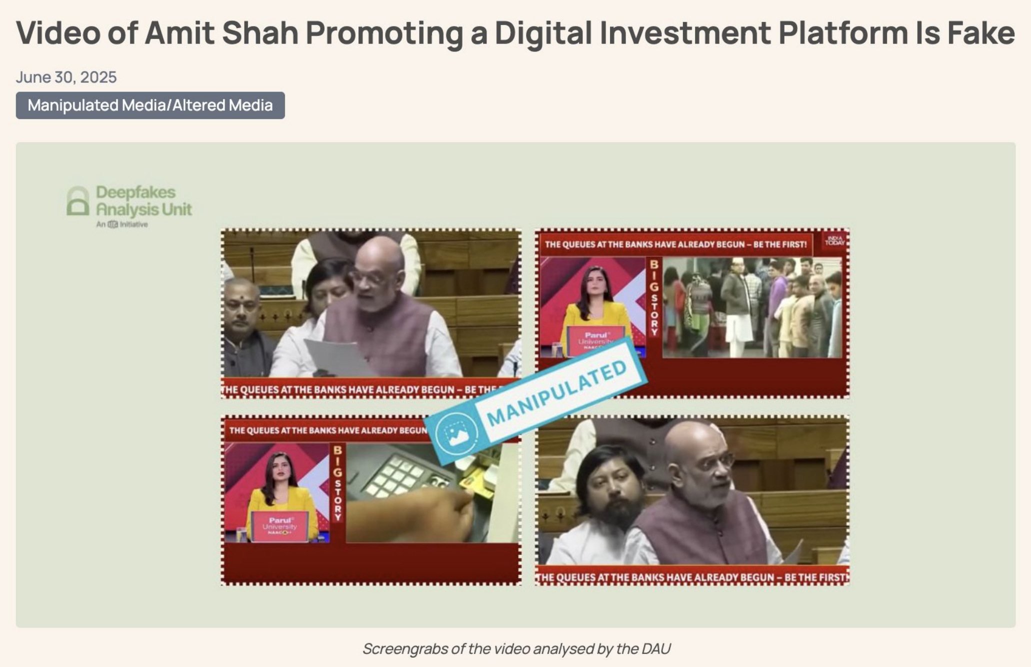
- 14 media items were escalated by International fact-checkers.

*Escalations refers to when a fact-checker and other organisations request DAU to assess a media item

Assessment Reports

The DAU published 12 assessment reports in Q1 of 2025.

- 8 reports were based on media received via the DAU Tipline
- 1 report was on media escalated by a fact-checking partner
- 2 reports were on media identified through social media monitoring
- 1 report was published after the DAU took suo motu cognizance of videos circulating online

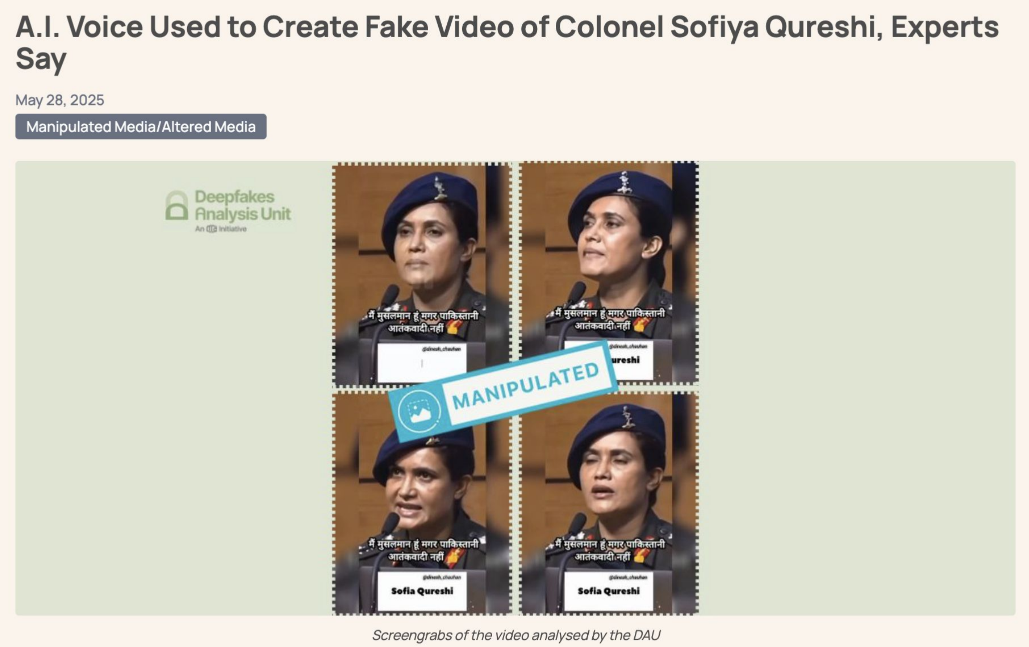


Video of Amit Shah Promoting a Digital Investment Platform Is Fake

June 30, 2025

Manipulated Media/Altered Media

Screengrabs of the video analysed by the DAU

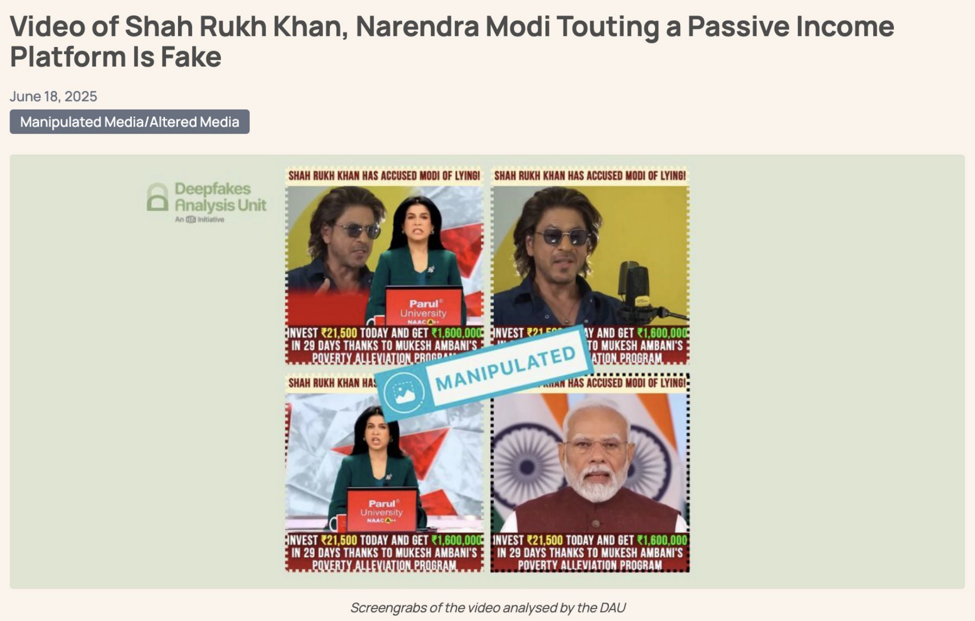


A.I. Voice Used to Create Fake Video of Colonel Sofiya Qureshi, Experts Say

May 28, 2025

Manipulated Media/Altered Media

Screengrabs of the video analysed by the DAU



Video of Shah Rukh Khan, Narendra Modi Touting a Passive Income Platform Is Fake

June 18, 2025

Manipulated Media/Altered Media

Deepfakes Analysis Unit
SHAH RUKH KHAN HAS ACCUSED MODI OF LYING!
INVEST ₹21,500 TODAY AND GET ₹1,600,000 IN 29 DAYS THANKS TO MUKESH AMBANI'S POVERTY ALLEVIATION PROGRAM
MANIPULATED

Screengrabs of the video analysed by the DAU

All assessment reports are available on [DAU website]

Feedback and Questions

Email:contactdau@mcaindia.in

Appendix A

DAU Labels and Definitions

DAU Labels and Definitions

For content that falls in the purview of the DAU, the secretariat assigns the media items one of the following labels:

- AI Generated
- Deepfake
- Manipulated
- Not Manipulated
- Not AI Generated
- Cheapfake

The secretariat produces an assessment report only for media items that may be detected to be manipulated using some element of AI. These could include media items labeled as deepfake, cheapfake, manipulated or AI Generated. You can read DAU Assessment Reports on the DAU Website.

AI Generated

A video or audio, created using AI, that depicts an event, a person, or an interaction between people that never actually occurred or that alters the reality, *which may or may not mislead people,* is labeled as AI generated. If content has been produced with the consent of the person depicted in the video, or if the content is produced for humorous or creative purposes, it is labeled as AI generated.

For example, an AI avatar of a politician from Rajasthan depicting him speaking in Tamil was labeled as AI generated.

Example of a Video Labeled as AI Generated

Deepfake

AI generated content that includes fictional creations about a living or dead person, or a non-existent person, intended to deceive or cause harm is labeled as Deepfake. Common examples of deepfakes include swapping the face of a person (dead or alive) or cloning their audio to an extent where the generated content shows them doing something they didn't do or saying something they didn't say.

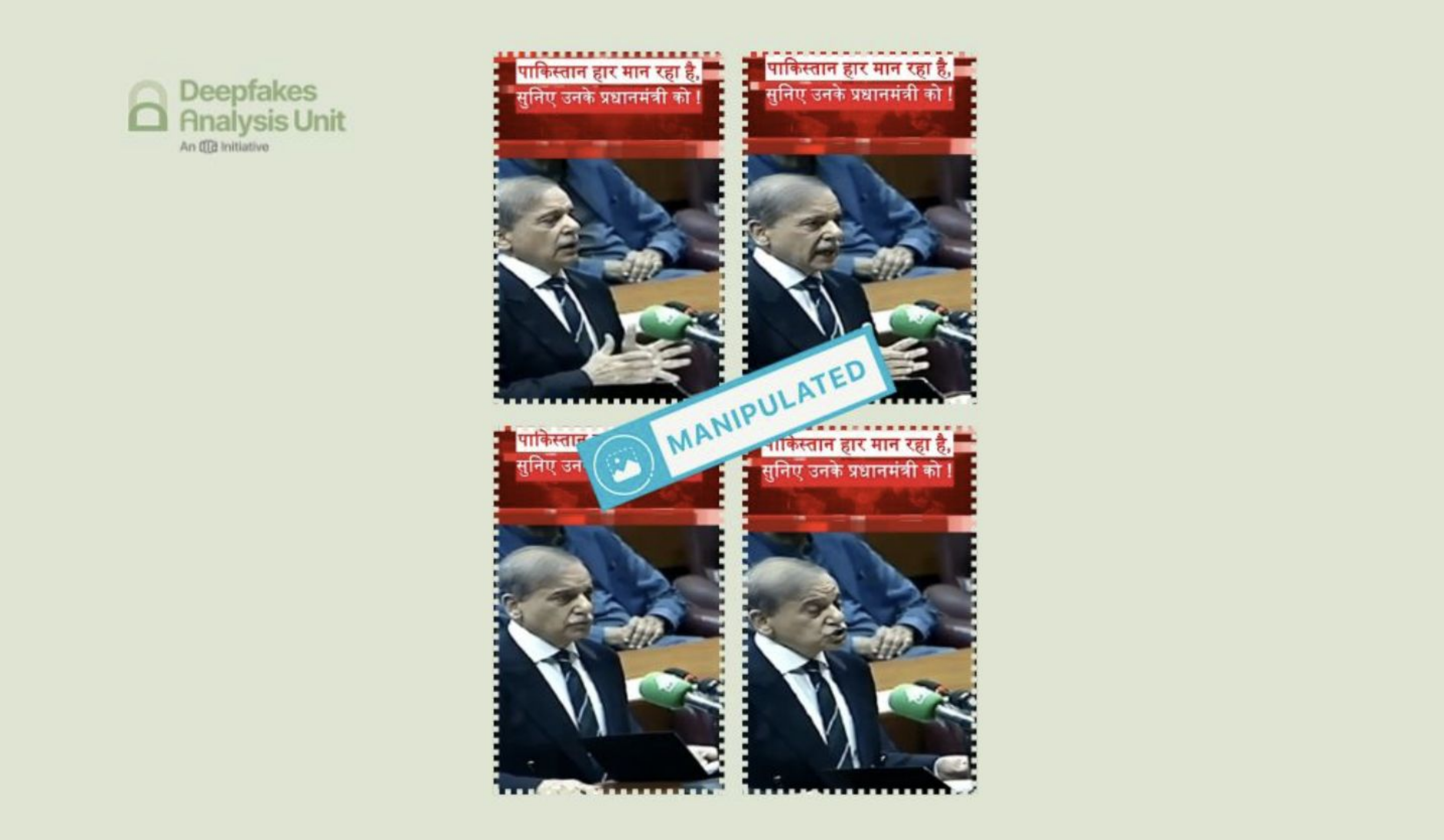
Every deepfake is AI-generated but every AI generated piece of media item is not a deepfake. AI generated content that shows nudity is labeled as a deepfake since the creator may or may not have sought consent from the subject(s) featured.

Example of Videos Labeled as Deepfake

Manipulated

Video or audio modified using simple editing software or AI technology to create a false or potentially misleading narrative, are labeled Manipulated. This includes cases where the original audio track from a video may have been replaced with AI-generated audio.

If the audio or video can be traced to an original source, the content is labeled as Manipulated, not AI Generated. In some cases, media artefacts such as the lips or mouth movement is blurred. This is seen as an attempt to synchronize the original visuals with the synthetic audio.



Example of a Video Labeled as Manipulated

Not Manipulated

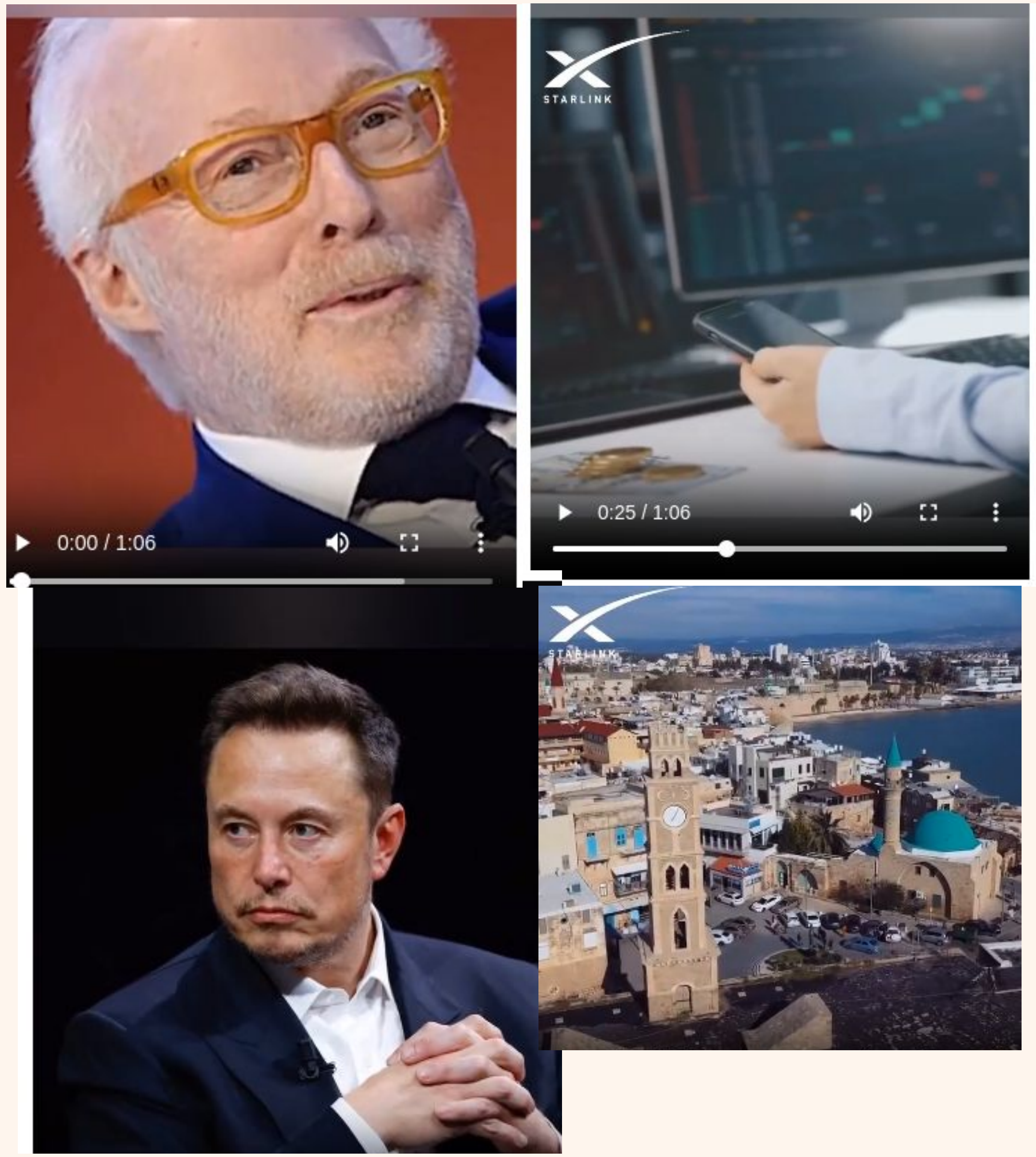
A video or audio which has not been tampered with and is identical to the source audio or video is labeled as Not Manipulated.

Not AI Generated

A video or audio, that is misleading, or has the potential to mislead but is not digitally altered is labeled as Not AI Generated. The DAU addresses these media items by surfacing fact checks, if any, from partners about the media item.

Cheapfake

A Cheapfake is a subtype of manipulated content where the quality of production is poor. This could include manipulation techniques such as slowing down, speeding up, reassembling, re-contextualising, and editing footage to create a false or a potentially false narrative; or satire or parody. The examples of content labeled as “cheapfake” by the DAU have included cases where the audio has been synthetic but the visual of a subject has been akin to a cutout or tucked in a corner of video with multiple other elements such as fleeting graphics or other moving visual elements.



Stills from a Video Labeled Cheapfake

Appendix B

Methodology Notes

Notes on Methodology

- The analysis is conducted on media items as received. Thus, the data may contain duplicates i.e. the same audio/video might have been counted more than once.
- Of all the content that is sent to the tipline, audio, video and URLs are filtered and sent to the dashboard for perusal by the DAU secretariat. While audio and video is detected through the file extension, URLs are detected by the presence of the terms:'http' and 'www'. URLs that do not contain these terms or extensions of it ('https'), or are of an unsupported media type, do not hit the DAU dashboard.
- Of the media items that hit the dashboard, the following media items are considered out of purview:
  - Spam
  - Out of Scope: Content that is not in the public interest or is private in nature
  - Unsupported Language: Content in languages other than English, Hindi, Tamil, Telugu, Urdu and Bengali.

Notes on Methodology

- The secretariat attempted to assign all media items in its purview with a verification status, and a language tag. Where the URL was broken, that is the URL did not lead to a functioning website, the secretariat could not provide a verification status or a language tag. In Q4, DAU marked 36 items as broken Urls. Similarly, where the media did not have speech or text, the secretariat could not add a language specific tag.
- Users who sent spam content did not receive a response from the secretariat.
- When a media item is tagged on the dashboard with a topical tag, it is tagged with multiple tags. These tags may overlap with each other. For example a media item tagged with 'Amitabh Bachchan' could also be tagged with 'Celebrity'.