Agenda

○ FYI: Media Manipulation
FYI: Manipulated Media
Issue

When independent fact-checkers identify misinformation, we reduce the distribution of the content and inform users of the fact checkers’ warning. People are concerned that this approach may not be sufficient to counter the persuasive effect of synthetic media (e.g., deepfakes); however, removing all “manipulated media” would lead to difficult decisions around what is actually “manipulated” and may also lead to removal of political speech and satire.
Manipulated Media

Overview

Recommendation:

• Maintain current approach of removing content that violates our Community Standards*, and reducing + informing most manipulated media; improve enforcement through better detection, and stronger inform treatments for photo and video content.

• Adopt a stronger enforcement standard for manipulated video when it is misleading.

External Outreach: 50 global external engagements

Working Groups: 3 XFN working groups focused on regional teams; small group meetings with Policy, Product, Partnerships, and Artificial Intelligence teams.

*Manipulated media that contains nudity or hate speech or otherwise violates our Community Standards, for example, would be removed.
Manipulated Media

Status Quo

• Independent third-party fact-checkers make determinations on whether content is misinformation.

• When content is rated “false,” we reduce its distribution and inform people.

• Text, image, audio, and video are eligible for fact-checking (and also subject to our Community Standards).

• Many of our partners have expertise in visual verification techniques, such as reverse image searching and analyzing image metadata. However, detection can be difficult.

• We continue to invest in technical detection, including through AI and machine learning.
We spoke to 50 experts including academic researchers, journalists, media professionals, pro-expression advocates, and experts in media manipulation.
Manipulated Media

External Outreach

- Maintain Reduce; Stronger Inform
- Stronger Reduce & Inform
- Remove some misleading manipulated media
- Remove all misleading manipulated media

Academic (US)
Journalist (APAC)
Journalist (US)
Journalist (EMEA)
Journalist (India)
Civil Society (US)
Civil Society (US)

Academic (US)
Journalist (EMEA)
Safety Org (EMEA)
Civil Society (US)
Civil Society (US)

Academic (US)
Journalist (India)
Academic (US)
Journalist (EMEA)
Academic (LATAM)
Civil Society (US)
Civil Society (US)

Academic (US)
Journalist (India)
Academic (US)
Journalist (EMEA)
Academic (US)
Civil Society (EMEA)

Academic (US)
Journalist (EMEA)
Academic (US)
Civil Society (EMEA)
Journalist (EMEA)
Manipulated Media

Define Misleading Manipulated Video as media that will cause irreversible harm (civic, reputational)

**Pros**
- In line with our current stance of removing misinformation re voting and/or physical harm
- Provides opportunity to take action ahead of an important civic moment such as an election

**Cons**
- Determining what amounts to “irreversible harm” can be highly subjective
Manipulated Media

Define Misleading Manipulated Video as media that is an **overt misrepresentation**

**Pros**
- Addresses fear of ‘deep fakes’
- May be more objective than Option 1

**Cons**
- Still subjective
Manipulated Media

Define Misleading Manipulated Video as media that is a technical “deepfake”

**Pros**
- Addresses public concern (given public use of this vocabulary)
- Facebook could leverage its Artificial Intelligence and Machine Learning teams to strengthen technical efforts at detection

**Cons**
- There is no standard definition of “deepfake”
- Tough to detect (adversarial space, will evolve)
- Deepfake does not necessarily equal harmful
Proposed Definition of Misleading Manipulated Video (Recommendation)

Video content that:
- Has been edited or synthesized
- Beyond adjustments for clarity or quality,
- In ways that are not apparent to an average person, and
- Would likely mislead an average person to believe that a subject of the video said words that they did not say

And:
- is the product of artificial intelligence or machine learning, including deep learning techniques (e.g., a technical deepfake), that merges, combines, replaces, and/or superimposes content onto a video, creating a video that appears authentic.

This policy does not extend to content that is parody or satire or is edited to omit words that were said or change the order of words that were said.
Manipulated Media

Examples

Misleading Manipulated Video (hypothetical/unconfirmed)
Manipulated Media
What are good supplementary approaches?

In parallel to in-product solutions, we will:

- Creating shared manipulated media effort across industry to aid and enhance detection speed.
- Further developing our taxonomy of manipulated media in partnership with industry and journalists.
- Identifying new partners who are experts in technical analysis.
- Investing in expanded manipulated digital and media literacy and resilience, similar to efforts done with a fact-checking partner during the European Parliamentary elections.
- Explore industry standards and best practices on manipulated media.