

DEEFAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY

**11-16 years****40-55 min****8-30 participants****Media investigation
and group discussion**

***Technology required: None; optional for extended verification

OVERVIEW

Participants investigate fictional social-media posts containing manipulated or misleading images, videos, audio and digital personalities. Instead of relying only on visible mistakes, they examine the source, context, supporting evidence and purpose of each post. They develop a practical verification checklist for responding to suspicious digital content.

LEARNING OBJECTIVES

Participants will:

- understand what deepfakes and synthetic media are
- recognize that realistic digital content is not necessarily authentic
- distinguish between manipulated content and misleading context
- investigate the source, evidence and purpose of an online claim
- understand that visual clues alone cannot reliably prove whether content is genuine
- practise responding responsibly before liking, sharing or commenting



MATERIALS

- Three to five fictional social-media investigation packs
- Printed posts, images, transcripts or video screenshots
- Source cards
- Context cards
- Evidence cards
- Investigation worksheets
- A board, flip chart or large sheet of paper
- Markers
- Three conclusion signs: **LIKELY AUTHENTIC**, **LIKELY MANIPULATED** OR **MISLEADING** and **NOT ENOUGH EVIDENCE**
- Optional magnifying glasses for the detective theme
- Optional internet-connected devices for the challenging variation

DEEFAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY



MATERIALS

Each investigation pack should contain:

- the original post or claim
- information about the account that shared it
- one or two supporting or contradicting sources
- contextual information
- the correct explanation for the facilitator

PREPARATION

Prepare fictional cases that avoid using real participants or sensitive local controversies. Suitable examples include:

CASE 1: THE FICTIONAL MAYOR

A video screenshot appears to show a fictional mayor announcing that schools will close for one month. The post comes from an account with a name similar to—but not identical to—the official account. No official website or established news organization reports the announcement.

CASE 2: THE UNUSUAL ANIMAL

A dramatic image appears to show a newly discovered glowing animal. The image is AI-generated, but the post presents it as a real scientific discovery. No scientific institution is identified.

CASE 3: THE OLD STORM PHOTOGRAPH

A genuine photograph of flooding is shared as evidence of a storm occurring today. A credible source reveals that the photograph was taken several years earlier in another country.

CASE 4: THE VIRTUAL INFLUENCER

A social-media personality recommends a health or beauty product without clearly revealing that the character is digitally created and controlled by a company.

CASE 5: THE CHANGED AUDIO CLIP

A short transcript claims that a fictional athlete insulted their team. A longer transcript reveals that the sentence was edited and removed from its original context.

Ensure that the fictional accounts and organizations cannot be confused with real ones. Keep the answer key separate.

DEEFAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY



INTRODUCTION

Ask the participants:

- Have you ever seen an image or video online that seemed too surprising to be true?
- What can AI now create or imitate?
- If a video looks realistic, does that prove it is genuine?
- Why might someone create or share manipulated content?
- What should we do before sharing something surprising?

Explain:

AI can generate or alter images, voices and videos. A deepfake is synthetic or manipulated media that convincingly represents a person as saying or doing something they did not actually say or do. However, not all misleading content is a deepfake. A genuine image can also be shared with a false caption or in the wrong context.

Tell participants that their task is not simply to spot visual mistakes. They must investigate the complete claim.

INSTRUCTIONS

1) EXAMINE THE INITIAL POST

Divide participants into small detective teams. Give each team one investigation pack, beginning with only the original post.

Ask them to record their immediate response:

- What does the post claim?
- What emotions does it create?
- Does it encourage an immediate reaction?
- Would you initially believe, doubt or remain uncertain about it?
- Would you share it?

Ask teams not to make a final decision yet.

2) INSPECT THE CONTENT

Teams examine the image, screenshot or transcript carefully.

They may look for:

- inconsistent visual details
- unnatural movement described in a video transcript
- unusual lighting or shadows
- mismatched voice and mouth movement
- strange wording
- missing information
- a caption that makes an unusually dramatic claim

Explain that these can be warning signs, but none provides definite proof. AI-generated media can contain no obvious mistakes, and authentic media can appear strange because of compression, lighting or poor recording quality.

DEEFAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY



INSTRUCTIONS

3) INVESTIGATE THE SOURCE Give teams the source card.

Ask:

- Who originally published the content?
- Is the account official or verified through another channel?
- Does the account name imitate a trusted source?
- Does it identify the creator or organization?
- Has the account published reliable information before?
- Is the post copied from somewhere else?
- Is the original source available?

Teams record whether the source increases or reduces their confidence.

4) CHECK THE CONTEXT Give teams the context card.

This may reveal:

- the full date and location;
- the longer version of a statement;
- that the image previously appeared elsewhere;
- that the personality is digitally created;
- that the post omits important information.

Ask:

- Does the additional context change the meaning?
- Is the content itself altered, or is the caption misleading?
- What information was left out?

5) LOOK FOR INDEPENDENT CONFIRMATION Give teams one or more evidence cards from other sources.

Some may confirm the claim, while others may contradict it or report nothing about it.

Ask participants:

- Are other reliable sources reporting the same event?
- Do the sources appear independent of one another?
- Are several websites simply repeating the same original post?
- Would an event of this importance normally be reported elsewhere?
- Is there an official statement?

Explain that several identical posts are not necessarily several independent sources.

6) REACH A CONCLUSION Teams place their case under one of three conclusions:

- **LIKELY AUTHENTIC**
- **LIKELY MANIPULATED OR MISLEADING**
- **NOT ENOUGH EVIDENCE**

They must support their decision with at least three pieces of evidence.

Allow **NOT ENOUGH EVIDENCE** as a valid and often responsible conclusion.

DEEPAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY



INSTRUCTIONS

7) REVEAL THE CASE

Each team presents:

1. the original claim;
2. its initial reaction;
3. the evidence it examined;
4. its final conclusion;
5. whether it would share the post.

Reveal the correct explanation.

Discuss which clues were most useful and whether the initial emotional reaction affected the investigation.

8) CREATE A VERIFICATION CHECKLIST

Develop a shared checklist called PAUSE, CHECK, TRACE.

PAUSE

- Does the post create a strong emotional reaction?
- Is it pressuring me to share immediately?
- Do I know whether the claim is true?

CHECK

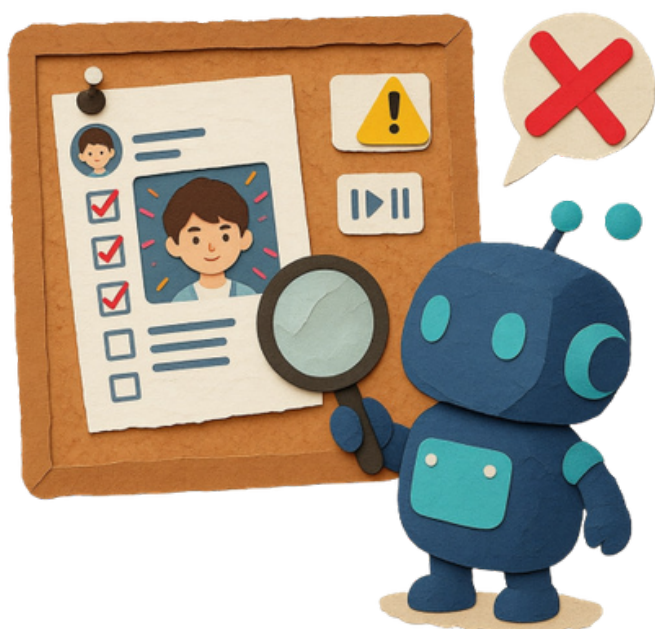
- Who published it?
- What evidence is provided?
- Are reliable sources reporting the same thing?
- Are the date, location and caption accurate?

TRACE

- Can I find the original image, video or statement?
- Has the content appeared elsewhere in a different context?
- Is there information about how it was created or edited?

Add a final instruction:

If reliable verification is not possible, do not present the claim as fact.



DEEPAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY



DISCUSSION AND REFLECTION

Ask the participants:

- Which evidence changed your opinion most?
- Was the misleading content always technically manipulated?
- Can a real photograph support a false claim?
- Why do dramatic or emotional posts spread quickly?
- Why is the original source important?
- Does a large number of likes prove that something is true?
- Can a verified or familiar account ever share incorrect information?
- Why might a digital character influence people even though it is not a real person?
- What harm could result from sharing a convincing fake?
- What should you do if you have already shared false information?
- When is "I don't know yet" the most responsible answer?

Conclude with the key idea: The most reliable verification does not depend on finding one strange pixel or obvious mistake. It combines source checking, context, independent evidence and careful reasoning.

CONNECTION TO ARTIFICIAL INTELLIGENCE

Generative AI can produce realistic:

- faces and photographs
- speech and cloned voices
- video and animation
- virtual presenters and influencers
- written posts and fabricated conversations

Deepfakes use synthetic-media techniques to make it appear that a real person said or did something that did not occur. Other AI-generated media may depict entirely fictional people, places or events.

Misleading media can take several forms:

- fully AI-generated content presented as real
- authentic content altered using AI
- real content placed in a false context
- edited content that removes important information
- digital characters presented without clear disclosure
- genuine content accompanied by an inaccurate caption

Automated detection tools may help, but they are not perfect and can produce both false warnings and missed detections. Verification therefore also requires human investigation, reliable sources and contextual evidence.

DEEFAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY



EASIER VARIATION

Use one fictional case with the whole group.

Present a dramatic image and ask participants to choose:

- share
- do not share
- check first

Reveal evidence in stages:

1. the account information
2. the date
3. the original source
4. a report from a reliable organization

After each stage, allow participants to change their choice.

Focus on three simple questions:

- Who posted it?
- Where else can we check?
- What might be missing?

MORE CHALLENGING VARIATION

Allow groups to conduct a controlled online verification exercise using safe, preselected examples.

Participants may:

- search for the original claim
- compare coverage from several reliable sources
- use reverse-image search
- locate an official statement
- check the publication date
- investigate whether an account is authentic
- look for information about how digital content was created

Ask groups to document each step rather than merely reporting their final judgment.

DEEFAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY



MORE CHALLENGING VARIATION

They can then create a short verification report containing:

- the claim
- the original source
- the evidence examined
- remaining uncertainties
- their conclusion
- a recommendation about whether the content should be shared

FACILITATOR NOTES

- Use fictional or carefully preselected cases to avoid spreading harmful misinformation.
- Clearly mark all fabricated examples as fictional after the investigation.
- Do not ask participants to create deepfakes of real people.
- Avoid using disturbing, sexual, violent or politically inflammatory examples.
- Do not teach visual imperfections as guaranteed indicators of AI generation.
- Emphasize that content may be misleading even when the image or recording itself is authentic.
- Explain that account verification symbols, popularity and professional appearance are useful context but not proof of truth.
- Encourage participants to seek the original source rather than relying on reposts.
- Treat uncertainty as a legitimate conclusion.
- If participants mention a real rumour, do not repeat it as fact.
Apply the verification checklist carefully.
- Remind participants not to harass or publicly accuse a creator when evidence remains uncertain.
- If false information has been shared, the responsible response is to correct it clearly and avoid continuing its circulation.

DEEFAKE DETECTIVES

MISINFORMATION AND MEDIA LITERACY



KEY WORDS

DEEFAKE

Synthetic or manipulated media that convincingly depicts a person saying or doing something that did not occur.

SYNTHETIC MEDIA

Digital content generated or substantially altered using technology.

MISINFORMATION

False or inaccurate information shared regardless of intent.

DISINFORMATION

False or manipulated information deliberately created or shared to deceive.

CONTEXT

Background information needed to understand content correctly.

SOURCE

The original person, organization or system responsible for publishing information.

VERIFICATION

Checking evidence to determine whether a claim is trustworthy.

INDEPENDENT CONFIRMATION

Support for a claim from sources that do not merely repeat one another.

VIRTUAL INFLUENCER

A digitally created character presented as a social-media personality.

