Lesson 4: Online Safety - Deepfakes

Spotting Fake Images, Videos & Voice Clones • 60 to 75 Min • KS3 (Ages 12 to 15)

SAFEGUARDING NOTICE

This lesson covers sensitive topics including AI-generated fake images and videos. **Before teaching:**

- Brief your safeguarding lead / designated safeguarding person (DSP)
- Do NOT demonstrate how to create deepfakes
- Use ONLY age-appropriate, non-distressing examples
- Be prepared for disclosures; students may share they've encountered harmful deepfakes
- Know your school's reporting procedures and available support
- Have CEOP (ceop.police.uk) and Childline (0800 1111) contact info displayed

Learning Objectives

- Define deepfakes and understand how they are created at a high level
- Identify at least 5 visual and audio clues that indicate AI-generated media
- Understand real-world harms: fraud, reputation damage, bullying, election interference
- Know the 5-step action plan if they encounter or are targeted by deepfakes
- Understand relevant UK law (Online Safety Act 2023)

Materials Needed

- Projector: pre-screened “Real or AI?” image set (5 to 10 safe images)
- Spot the Fake worksheet (from toolkit)
- Scenario discussion cards (printable below)
- 5-Step Action Plan handout (printable below)
- Digital polling tool (e.g., Mentimeter, hands up, or mini-whiteboards)
- Safeguarding referral information displayed in classroom

Teacher Background Knowledge

What Are Deepfakes?

Deepfakes are AI-generated or AI-manipulated media: face-swapped videos, entirely synthetic images, cloned voices, or AI-generated text. The term combines “deep learning” (the AI technique) and “fake.”

How they’re made (teacher knowledge only):

• **Face-swap videos:** AI studies hundreds of images/videos of a person’s face from multiple angles, then maps that face onto another person in a video
• **Synthetic images:** AI image generators (Midjourney, DALL-E, Stable Diffusion) create entirely new images from text descriptions using diffusion models
• **Voice clones:** AI analyses as little as 3 seconds of someone’s voice and generates new speech in that voice. Tools are freely available online
• **Video generation:** New tools (Sora, Runway) can generate entire video clips from text

Scale of the Problem (2025 Statistics)

• Deepfake videos online grew **550%** from 2019 to 2023 (Sensity AI)
• AI image generators produce **34+ million** synthetic images daily
• **96%** of deepfake videos online are non-consensual intimate imagery (Sensity)
• **500,000** deepfake voice and video scams globally in 2023
• **$25 million** fraud in Hong Kong (Jan 2024) using deepfake video calls
• **40%** of people cannot reliably distinguish AI images from real ones (MIT Media Lab)
• Average cost to create a convincing deepfake has dropped from $10,000+ to under $5

UK Legal Framework

Online Safety Act 2023:

• Creating intimate deepfakes is a **criminal offence**, up to 2 years in prison
• Sharing deepfakes intended to cause harm is prosecutable
• Platforms must remove reported deepfakes quickly or face massive fines
• Under-18s receive enhanced protections

Malicious Communications Act 1988: Sending threatening/offensive content, including deepfakes, is a criminal offence.

School responsibilities: Sharing intimate deepfakes among students is a safeguarding issue. Follow school reporting procedures. Document incidents. Contact police if criminal content is involved.

Detection Clues: The 10-Point Checklist

AI-Generated Images:

1. **Hands:** Wrong number of fingers (too many or too few), distorted joints
2. **Eyes:** Mismatched reflections, different sizes, unusual shapes
3. **Teeth:** Too uniform, too many, or merging together
4. **Text:** Garbled letters, misspelled signs, impossible text
5. **Skin:** Unnaturally smooth, waxy texture, no pores or blemishes
6. **Hair:** Plasticky appearance, merges with background, inconsistent strands
7. **Background:** Impossible architecture, melting objects, inconsistent geometry
8. **Lighting:** Shadows in wrong direction, multiple light sources that don’t match
9. **Accessories:** Mismatched earrings, glasses that merge with face, asymmetric jewellery
10. **Edges:** Blurry boundaries where face meets hair or background, visible seams

Deepfake Videos: Additional Clues

• Unnatural blinking patterns (too frequent or too infrequent)
• Lip movements slightly out of sync with audio
• Face doesn’t move naturally with the body (head rotation looks stiff)
• Inconsistent skin tone at the face boundary
• Brief “glitches” when the person turns their head quickly
• Hair around the face edge looks unnatural or “painted on”
• Emotion and tone of voice don’t match facial expressions

Voice Clone Clues

• Slightly robotic or “flat” tone with less natural emotion
• Unusual breathing patterns (too regular or absent)
• Background sounds may be missing or added artificially
• Unusual requests (urgency, money, secrecy) from a “familiar” voice
• The “person” can’t answer personal questions only they would know

Detailed Lesson Timeline

Time	Activity	Detailed Instructions
0 to 10 min	**Real or AI? Quiz**	Show 8 carefully pre-screened images (mix of real photos and AI-generated). For each image, students vote: REAL or AI? Use digital poll, mini-whiteboards, or hands up. **Recommended image types (safe, non-distressing):** • AI landscape (mountains, sunset, very convincing) • Real photo of an unusual scene (naturally looks fake) • AI portrait (look for hand/eye clues) • Real wildlife photo • AI-generated food photo (eerily perfect) • Real photo with unusual lighting • AI historical scene (anachronisms?) • Real aerial photo **After voting:** Reveal answers. Most students will get 1 to 3 wrong. **Key question:** “How did it feel to be tricked? If YOU can’t tell, who else might be fooled?”
10 to 22 min	**What Are Deepfakes?**	**Teacher script (age-appropriate, non-technical):** “Deepfakes use AI to create fake images, videos, and audio that look and sound real. The AI studies many photos or videos of a person, learning what they look like from every angle, and then generates new, fake content. There are 3 main types:” **1. Fake Images:** AI creates entirely new photos of people who don’t exist, or puts real people in situations that never happened. **2. Face-Swap Videos:** AI replaces one person’s face with another’s in a video, making it look like someone said or did something they didn’t. **3. Voice Clones:** AI copies someone’s voice from just a few seconds of audio, then generates new speech in that voice. This is used in scam phone calls. **Key facts to state clearly:** • Creating fake intimate images of anyone (including classmates) is a CRIMINAL OFFENCE • A classmate’s face can be put on anything; this is a form of bullying and abuse • Deepfake technology is increasingly easy to use and often free

Time	Activity	Detailed Instructions

22 to 30 min	**Real-World Impact**	Present 5 brief, factual, age-appropriate case studies: **1. The Pope in a Puffer Jacket (March 2023):** An AI-generated image of Pope Francis wearing a white Balenciaga puffer jacket went viral. Millions believed it was real. It was created using Midjourney and took about 1 minute. **2. Hong Kong Fraud ($25 million, Jan 2024):** Scammers used deepfake video calls to impersonate a company's CFO. An employee transferred $25 million to fraudsters because the person on the video call looked and sounded exactly like their boss. **3. Election Deepfakes (2024):** AI-generated audio of politicians saying things they never said was used to interfere with elections in multiple countries. In one case, a fake robocall from "President Biden" told voters to stay home. **4. School Bullying Cases (2023 to 2024):** Multiple reports of students creating deepfake images of classmates. In New Jersey, a group created fake images of female classmates; all students faced criminal charges. **5. Celebrity Scams:** Deepfake ads using celebrity faces to promote scam products. Martin Lewis, Elon Musk, and others have been victims.
30 to 42 min	**Spot the Fake: 10-Point Checklist**	Using the Spot the Fake worksheet, teach the 10-point detection checklist: Go through each clue with a visual example: 1. **Hands:** Count the fingers! AI often gives 6+ fingers 2. **Eyes:** Zoom in on reflections. Are they consistent? 3. **Teeth:** Too perfect? Merging together? Count them 4. **Text:** Any text in the image should be legible and correct 5. * **Skin:** Real skin has pores, blemishes, texture 6. **Hair:** Does it look like plastic? Individual strands visible? 7. **Background:** Look for impossible objects or melting buildings 8. **Lighting:** Do shadows match? Is light consistent? 9. **Accessories:** Are earrings matching? Do glasses look right? 10. **Edges:** Where does the face meet the background? Students practice with 5 to 8 prepared images.

Time	Activity	Detailed Instructions
42 to 55 min	**Scenario Discussions**	Groups of 4 receive scenario cards and discuss for 4 minutes each, then share: **Scenario 1:** A friend sends you a "shocking" video of a celebrity doing something embarrassing. They want you to share it. What do you do? -> Check if it's real first. Don't share unverified content. Report if suspicious. **Scenario 2:** You discover someone has used AI to create a fake image of a classmate in an embarrassing situation. What do you do? -> Don't share it. Tell a trusted adult IMMEDIATELY. This could be a criminal offence. Support the victim. **Scenario 3:** You receive a voice message from "Mum" asking you to urgently transfer money or share a bank password. Something feels slightly off. What do you do? -> NEVER act on urgent requests. Call Mum back on a number you know. Ask a question only she would know. Voice clones can't answer personal questions. **Scenario 4:** A news article shows a photo of a local politician at a controversial event. It's being shared widely on social media. Should you believe it? -> Apply REAL Framework. Check multiple news sources. Reverse image search. Check the original publication date and context.

55 to 70 min	**5-Step Action Plan**	Teach and distribute the action plan. Students copy into their exercise books: **STEP 1: DON'T SHARE IT** Sharing a deepfake spreads harm. Even sharing to say "look how fake this is" gives it an audience. STOP the chain. **STEP 2: SCREENSHOT AS EVIDENCE** Take a screenshot BEFORE it's deleted. Include the URL, username, date/time. **STEP 3: TELL A TRUSTED ADULT** Parent, teacher, school counsellor, safeguarding lead. Don't try to handle it alone. **STEP 4: REPORT ON THE PLATFORM** Every major platform has a report button. Report as "fake/manipulated media." **STEP 5: CHECK FACT-CHECKING SITES** Full Fact (fullfact.org), Snopes, Google Fact Check, BBC Reality Check. **Remind students:** Creating deepfakes of real people can be ILLEGAL. Under the Online Safety Act 2023, intimate deepfakes are a criminal offence.

Printable: 5-Step Deepfake Action Plan

Print one per student. A4 size recommended for classroom display.

DON’T SHARE IT

Sharing a deepfake spreads harm. Stop the chain. Don’t forward, repost, or screenshot to share.

SCREENSHOT AS EVIDENCE

Take a screenshot BEFORE it’s deleted. Include the URL, username, and date/time.

TELL A TRUSTED ADULT

Talk to a parent, teacher, or school safeguarding lead. Don’t handle it alone.

REPORT ON THE PLATFORM

Use the report button. Select “fake/manipulated media” or “bullying/harassment.”

CHECK FACT-CHECKING SITES

Full Fact, BBC Reality Check, Snopes, Google Fact Check Explorer.

Emergency Contacts:

- **CEOP:** ceop.police.uk, report online abuse
- **Childline:** 0800 1111 (free, 24/7), talk to someone
- **Internet Watch Foundation:** iwf.org.uk, report illegal content
- **Report Harmful Content:** reportharmfulcontent.com

Differentiation Strategies

Activity	Lower Ability / SEN	Core	Higher Ability
Real or AI?	5 images with clear clues. Teacher points out what to look for.	8 images, independent voting.	10 images including very convincing ones. Must explain WHAT gave it away.
Detection	Focus on 5 most obvious clues (hands, text, skin, background, edges).	Full 10-point checklist.	Research: find and test AI detection tools (Hive Moderation, AI or Not). Evaluate their accuracy.
Scenarios	2 scenarios with guided questions. Teacher-facilitated discussion.	4 scenarios, group discussion.	Create their own 5th scenario. Write detailed response including legal and ethical analysis.
Action Plan	Pre-printed plan. Students highlight key words and add examples.	Copy plan and explain each step in own words.	Write a guide for Year 7 students explaining deepfakes and the action plan. Must be accurate and age-appropriate.

Assessment Criteria

Level	Descriptor	Evidence
Emerging	Knows deepfakes exist. Can identify 1 to 2 detection clues. Knows not to share fake content.	Basic contributions during Real or AI? quiz. Copies action plan but can't explain steps.
Developing	Can define deepfakes. Identifies 3 to 4 detection clues. Understands the action plan.	Correctly identifies most images. Participates in scenario discussion.
Secure	Understands how deepfakes are created. Applies 5+ detection clues confidently. Knows UK law.	Detects most AI images. Provides thoughtful scenario responses. Can explain legal consequences.
Excelling	Deep understanding of creation, detection, and societal impact. Can teach others. Analyses ethical implications.	Detects subtle deepfakes. Leads scenario discussion. Can propose solutions to deepfake challenges.

Key Vocabulary

Deepfake: AI-generated or AI-manipulated media (images, video, audio) designed to look or sound authentic. Named from "deep learning" + "fake."

Face-Swap: Replacing one person's face with another in a video, making it appear they said or did something they didn't.

Voice Clone: An AI-generated copy of someone's voice that can speak any text. Can be created from as little as 3 seconds of audio.

Synthetic Media: Any media content (image, video, audio, text) that has been created or modified using AI technology.

GAN (Generative Adversarial Network): An AI architecture where two neural networks compete: one generates fakes, the other tries to detect them. This arms race produces increasingly convincing fakes.

Non-Consensual Intimate Imagery (NCII): Sexual or intimate images/videos created or shared without the subject's consent. Creating AI-generated NCII is a criminal offence in the UK.

Online Safety Act 2023: UK law that criminalises creating intimate deepfakes, requires platforms to remove harmful content, and gives Ofcom enforcement powers.

Post-Lesson Teacher Notes

Safeguarding concerns raised during lesson:

Student understanding of action plan:

What worked well / what to change:

Additional Resources:

- CEOP ThinkUKnow: thinkuknow.co.uk (age-appropriate online safety)
- Internet Matters: internetmatters.org/issues/deepfakes
- BBC Bitesize: What are Deepfakes? at bbc.co.uk/bitesize
- Sensity AI: State of Deepfakes Report at sensity.ai
- UK Government: Online Safety Act Factsheet at gov.uk