Uncovering Deepfakes

AI for Education

Classroom Guide + Discussion Questions

Since 2019, the presence of **Deepfakes** - hyperrealistic AI-generated video, photo, and audio forgeries - has surged by 550%, reaching a staggering 95,820 documented deepfake videos in 2023 alone.

This blurring of reality requires a new approach to vetting if online content is real or fake. This guide is designed to build student awareness of the presence and impact of deepfakes, while providing key discussion topics on the ethics of AI-generated content.

What are Deepfakes?

Deepfakes are videos, photos, or audio recordings that use artificial intelligence to make it look or sound like someone is saying or doing something that is not real.

Deepfake Techniques

- **Face Swapping:** Replacing a person's face with another in a video or photo using AI.
- **Voice Cloning:** Using AI to mimic someone's voice.
- **Lip Syncing:** Making a person's lips appear to match different audio using AI.

The Dangers of Deepfakes

While deepfakes can be used for harmless or even educational purposes, they can also be used unethically in the following ways:

- **Identity theft:** Deepfakes can impersonate individuals in videos or audio recordings, potentially allowing criminals to commit fraud under someone else's name.
- **Threats/blackmail:** Deepfakes can create convincing footage or audio that puts the subject in a bad light, which can then be used to threaten or blackmail them.
- **Cyberbullying:** Deepfakes can be used to create embarrassing or hurtful videos or audio recordings of someone.
- **Propaganda/fake news:** Deepfakes can be used to create fake videos or audio recordings that trick people and sway their opinions on important topics.
- **Election manipulation:** Deepfakes can be used to make candidates look bad or confuse voters by creating fake speeches or interviews.

How to Identify Deepfakes

AI for Education

Detecting deepfakes involves looking for inconsistencies and anomalies in videos. photos, or audio recordings that may not be immediately apparent. It's important to note that deepfake technology is quickly improving, so **it may soon be impossible to tell a fake from something real.**

Look closely at people's faces in the video. If their skin looks too smooth, their eyes don't blink normally, or their facial expressions seem off, it might be a deepfake.

If the voice doesn't sound quite right, like it's **too robotic or the tone doesn't match how the person usually speaks**, it could be a sign of a deepfake.

Pay attention to **how the lips move**. If they're not syncing up with the words perfectly, or the movement looks strange, the video might be manipulated.

Consider the video's **overall appearance**. Sometimes, the area around the person or behind them might look **blurry or glitchy**. This can be a sign that the video has been altered.

Or if the **shadows on the person's face or in the background don't match up with the lighting** in the rest of the video, it could be a clue that it's a deepfake.

*A variety of **deepfake detectors** are available online, but as these operate based on a model that calculates probabilities, which **will never be 100% reliable.***

*Large tech companies such as **Google, Meta, and OpenAI** are joining together to form the **Coalition for Content Provenance and Authenticity (C2PA)** in order to develop methods of labeling AI-generated content to clearly identify when and how it was generated or altered.*

What Can You Do?

- **Question everything:** AI-generated content has already flooded the internet, so be cautious. Always investigate if a suspicious image, video, or audio clip is real.
- **Look for the clues described above:** glitches and blurs, lip synch issues, etc.
- **Check the source:** be extra cautious with content from unknown websites or accounts.
- **Talk about it:** if you're unsure, share your concerns with trusted adults.
- **Get permission:** before using someone's face or voice when generating AI content.
- **Be transparent:** if you're creating an AI-generated copy of a person, clearly label your creation as AI-created.
- **Use your power for good:** be mindful of the potential impact of what you share or create.

Discussion Questions

Use the following scenario-based discussion questions to build background knowledge on the impact of deepfakes on society to inform students' responses to the more general questions on the following page. Links to the news sources for each scenario can be found in the Sources and Further Reading section of this document.

1. In January 2024, thousands of AI-generated robocalls impersonating President Joe Biden went out to Democratic voters discouraging them from voting in the New Hampshire primary. How might this type of deepfake use impact the integrity of elections? How can we combat this?

This question encourages a discussion on the potential manipulation of voter perceptions and behavior through false representations of political candidates.

2. One 2023 study found that 98% of the deepfake videos online consisted of nonconsensual pornography. The targets of 99% of this content were women, including pop star Taylor Swift. Why do you think this type of content is so widespread, and what are the social and ethical implications?

This question prompts a critical discussion about the pervasive use of deepfakes to create nonconsensual pornography, focusing on the disproportionate impact on women.

3. In summer of 2023, a group of New Jersey teenagers used AI to create deepfake pornographic images of an underaged female classmate, which were circulated over group text. What psychological harms could this kind of incident cause for the victim in terms of trust, safety, and mental health? Should legal action be taken against the students involved?

This question raises serious issues about the exploitation of minors using deepfake technology, as well as implications for online privacy and consent. It further prompts students to consider appropriate legal and disciplinary responses.

4. In April 2024, a high school principal in Maryland fell victim to a deepfake attack when a disgruntled athletic director faked a recording of the principal making racist remarks. What are the consequences of using deepfakes to impersonate authority figures? Consider implications for the victim, the attacker, and the system as a whole.

This question encourages exploration of the ethical, social, and legal repercussions of such digital impersonations at the system level, and how they can undermine trust within communities.

Discussion Questions cont.

AI for Education

5. What do you think has been the impact of deepfakes online? How has it changed how you interact with the internet?

This question invites students to analyze and discuss the various effects that deepfakes have had on the online environment. Considerations include influencing public perception, personal privacy, and trust in digital media.

6. Do you think deepfakes should be banned? Why or why not?

This question prompts students to consider the pros and cons of the availability of deepfake technology, weighing the implications for freedom of expression, technological innovation, and personal security.

7**. Can you think of any positive use cases for deepfake content? Under what circumstances can deepfakes be created and shared ethically?**

T*his question encourages students to identify and discuss potential positive applications of deepfake technology. Discuss scenarios such as education, entertainment, and philanthropy, and consider the benefits, risks, and circumstances that would justify deepfake use.*

8. With advancements in AI technology, deepfakes are likely to become even more sophisticated. How should society prepare for these changes? Consider the roles of individuals, educators, policymakers, and technology companies in shaping a response.

This question facilitates a discussion on the necessary preparations for future advancements in AI and deepfake technology.

Examples to Use in the Classroom

The following links are just a few of the thousands of examples of deepfakes that exist.

- Katy Perry Fake MetBall
- Stephen Fry's Voice Copied from Harry Potter Audiobooks
- Spanish AI Influencer/Model
- FKA Twigs Creates AI Avatar for Herself
- Fake Drake + Weeknd Song
- Reed Hoffman Talks to His AI Twin
- AI Impact on Elections News Piece

Sources & Further Reading

2023 State of Deepfakes: Realities, Threats, and Impact

What are deepfakes? How they work and how to spot them

One Tech Tip: How to spot AI-generated deepfake images

Political consultant behind fake Biden robocalls says he was trying to highlight a need for AI rules

Deepfake explicit images of Taylor Swift spread on social media. Her fans are fighting back

Deepfake of principal's voice is the latest case of AI being used for harm

Students at NJ school reportedly used AI to make pornographic images of classmates

How AI helped Val Kilmer get his voice back for Top Gun: Maverick

SIFT Model of Factchecking Online Content