Are humans or machines better at detecting deepfakes?

Humans' ability to identify fake content is critical to neutralizing misinformation, but as AI technology improves, who can we trust to detect deepfakes: Humans or machines?

Deepfake threatens every aspect of society. Humans' ability to identify fake content is critical to neutralizing misinformation, but as AI technology improves, who can we trust to detect deepfakes: Humans or machines?

The dangers of Deepfake

As AI technology advances, the dangers of deepfakes become an increasing threat to all of us. Here's a quick summary of some of the most pressing issues posed by deepfakes:

1. **Misinformation** : Videos and voice recordings using deepfake can spread false information, such as fake news.
2. **Impersonation** : By impersonating individuals, deepfakes can damage people's reputations or deceive anyone who knows them.
3. **National Security** : A deepfake doomsday scenario is fabricated video or audio of a global leader instigating conflict.
4. **Civil Unrest** : Deceptive footage and audio can also be used by parties to stir up anger and civil unrest among specific groups.
5. **Cybersecurity** : Cybercriminals have used AI voice cloning tools to target an individual with persuasive messages from people they know.
6. **Privacy and consent** : The malicious use of deepfake takes the image of individuals without their consent.
7. **Trust and confidence** : If you cannot distinguish between truth and lies, accurate information becomes unreliable.

Deepfakes will become more and more convincing, so we need powerful tools and processes to detect them. AI is providing such a tool in the form of a deepfake detection model. However, like the algorithms designed to identify AI-generated text, deepfake detection tools are not perfect.

At this time, human decision-making is the only tool we can rely on. So, are we better than algorithms at identifying deepfakes?

Can algorithms detect deepfakes better than humans?

Deepfake is a serious enough threat that tech giants and research groups are dedicating huge resources to research and development. In 2019, companies like Meta, Microsoft, and Amazon offered $1,000,000 in prizes in the "Deepfake Detection Challenge".

The best performing model had 82.56% accuracy against a dataset of publicly available videos. However, when testing the same models with a "black box dataset" of 10,000 unwatched videos, the best performing model had only 65.18% accuracy.

There are also many studies analyzing the performance of AI deepfake detection tools on humans. Of course, results vary from study to study, but overall, humans have equal or superior success rates compared to deepfake detection tools.

A 2021 study published in PNAS found that "casual human observers" achieved slightly higher accuracy rates than leading deepfake detectors. However, the study also found that human participants and AI models were susceptible to different types of errors.

Interestingly, research conducted by the University of Sydney has shown that the human brain is, unconsciously, more effective at detecting deepfake behaviors than our conscious efforts.

Detecting visual clues in Deepfake

The science of deepfake detection is complex and analysis requirements will vary, depending on the nature of the footage. For example, North Korean leader Kim Jong-un's infamous deepfake video from 2020 is essentially a chat video with only movement at the beginning. In this case, the most effective deepfake detection method may be to analyze images (mouth movements) and phonemes (sounds) to find inconsistencies.

Human experts, ordinary viewers, and algorithms can all perform this type of analysis, even if the results may vary. MIT identifies 8 questions to help identify deepfake videos:

1. **Pay attention to the face** . Advanced DeepFake operations are almost always facial transformation operations.
2. **Pay attention to your cheeks and forehead** . Does skin seem too smooth or too wrinkled? Is the age of skin the same as the age of hair and eyes? DeepFake can be unreasonable in some respects.
3. **Pay attention to the eyes and eyebrows** . Does the shadow appear in the places you expect it to? DeepFake may not fully represent the natural physics of a scene.
4. **Pay attention to the glasses** . Is the glass glare? Does the glare angle change as the person moves? Again, DeepFake may not fully represent the natural physics of light.
5. **Pay attention to facial hair** . Does the facial hair look real? DeepFake can add or remove mustaches, sideburns or beards. However, DeepFake may not be able to perform completely natural facial hair transformations.
6. **Pay attention to moles on the face** . Do moles look real?
7. **Pay attention to your blinking** . Does the person blink too little or too much?
8. **Pay attention to your lip movements** . Some deepfakes rely on lip-syncing. So is the lip movement natural?

The latest AI deepfake detection tools can reanalyze similar elements with varying degrees of success.

Data scientists are also constantly developing new methods, such as detecting natural blood flow on the faces of people speaking on a screen. New approaches and improvements over existing ones could help AI deepfake detection tools outperform humans in the future.

Detecting audio clues in Deepfake

Detecting deepfake audio is another challenge altogether. Without the visual cues of video and the opportunity to identify audiovisual inconsistencies, deepfake detection relies heavily on audio analysis (other methods such as metadata verification are also possible). useful in some cases).

A study published by University College London in 2023 found that humans can detect deepfake speech 73% of the time (English and Mandarin). As with deepfake videos, humans often intuitively detect unnatural speech patterns in AI-generated speech, even if they can't clearly tell what's wrong.

Common signs include:

1. Whisper
2. Lack of natural expression
3. Background noise or interference
4. Inconsistent voice or speech
5. Voice lacks 'roundness'
6. The voice delivery is too "dramatic)
7. Lack of natural cues (no stumbling, throat clearing, etc.)

Again, algorithms can also analyze speech for similar signs of deepfakes, but new methods are making the tools more effective. USENIX research identified patterns in AI speech reproduction that fail to simulate natural speech. The report summarizes that the AI ??speech generator produces sounds tailored to narrow vocal ranges without the natural motion of the human voice.

Previous research from the Horst Görtz Institute analyzed real and deepfake sounds in English and Japanese, revealing subtle differences in the higher frequencies of real speech and deepfake sounds.

Both loudness and high-frequency inconsistencies can be perceived by listeners and AI detection models. In the case of high-frequency differences, AI models could theoretically become more and more accurate - although the same could happen for AI deepfakes.

Humans and algorithms are both fooled by Deepfake, but in different ways

Studies show that humans and the latest AI detection tools are capable of identifying similar deepfakes. Success rates can vary from 50% to 90+%, depending on testing parameters.

More broadly, humans and machines are fooled by deepfakes to a similar extent. Importantly, however, humans are vulnerable in a variety of ways, and this may be our greatest asset in addressing the dangers of deepfake technology. Combining human strengths and deepfake detection tools will minimize weaknesses and improve success rates.

For example, MIT research shows that humans are better at identifying deepfakes of world leaders and celebrities than AI models. The study also revealed that AI models struggled with shots with multiple people, although it suggested this could be because the algorithms were trained on shots with one person speaking.

Conversely, the same study found that AI models performed better than humans with low-quality footage (blurred, dark, etc.) that could be intentionally used to deceive viewers. Likewise, recent AI detection methods such as monitoring blood flow in specific areas of the face incorporate analytical capabilities that humans cannot perform.

As more methods are developed, not only is AI's ability to detect signs that humans cannot improve on improving, but so is deepfake's ability to deceive. The big question is whether deepfake detection technology will surpass deepfake itself.