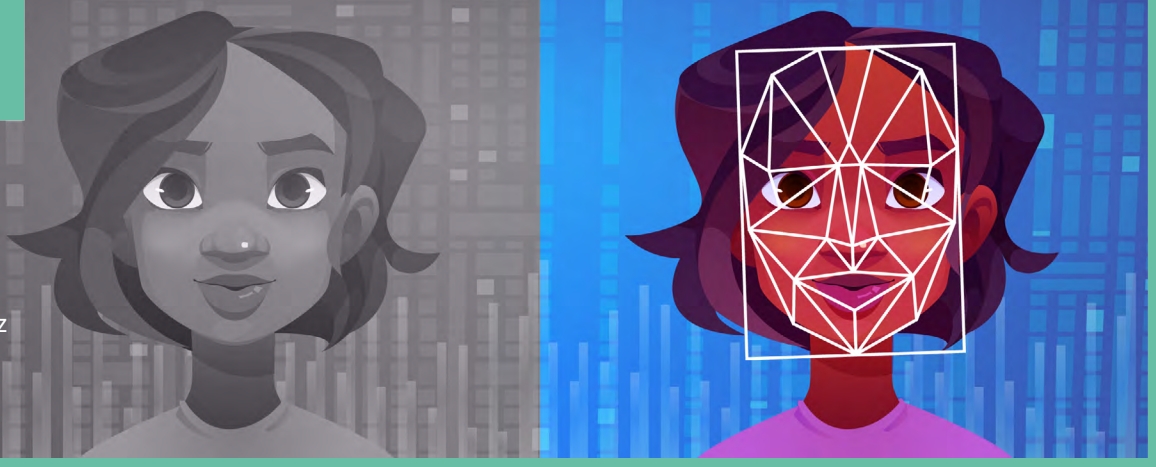
TECHNOLOGY SCIENCE JOURNAL FOR TEENS

JANUARY 2024

How can humans and AI work together to detect deepfakes?



Researchers:
Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard

Associate Editors:
Christy Ascione and Fiona Firth

Abstract

Fake news is not new on the internet, and people often change images and videos for a joke. However, deepfakes aren't only meant to make you laugh. Instead, they can spread misinformation or discredit a person or a group. As more deepfakes find their way onto the internet, we need to find the best way to detect these harmful videos. We tested whether the leading AI model or humans were better at detecting deepfakes online. We found that humans and the AI model were each good at identifying certain types of deepfakes. Maybe we could merge the abilities of both AI and humans to create the best deepfake detection model!

Introduction

Imagine you are watching a video online of your favorite celebrity and they say something very offensive. You might be shocked, or maybe even angry. But you have a feeling that something is wrong with this video – and it turns out it's a **deepfake**.

For a long time, video evidence has been the best way of indicating whether someone did or said something. Unfortunately, the rise of deepfakes means that is no longer the case. Deepfakes are videos that show people saying or doing things that didn't really happen. These videos are created by an **artificial intelligence (AI)** system based on **deep learning** – leading to the name deepfake. There have been a handful of deepfakes that have gone viral in the past few years. These include videos of well-known figures like Barack Obama, Donald Trump, and Mark Zuckerburg saying things they didn't say. But can we tell the difference between real and fake videos that we see online?

Engineers have trained **machine learning** models to try to identify deepfakes on the internet. However, the best known model can only detect deepfakes with around 65% accuracy. We wanted to know whether humans or the latest machine learning model were better at detecting a deepfake. We also wanted to see if human emotion plays a role in our ability to tell what's real and what's fake.

We then used our results to suggest how best to detect deepfakes online!

Methods

We designed a website called *Detect Fakes*, where anyone could view deepfake videos. We used this site to test how accurately ordinary people could detect a deepfake. Most of the videos were of unknown people making unimportant statements. That was to make the experiment more equal between humans and the machine learning model.

We conducted two experiments:

→ **Experiment 1** – One real and one fake video side by side

All participants in this experiment found the site while browsing. They watched a deepfake video alongside its corresponding real video, and we then asked them to choose which one was fake. There were 56 pairs of videos. We examined the accuracy of 882 participants who watched at least 10 pairs of videos (Figure 1). The machine learning model assessed all 56 videos.

→ **Experiment 2** – One real or fake video

We had two kinds of participants in this experiment: people we recruited and people who found the site while browsing. We focused on participants who had viewed at least 10 videos. In total, there were 301 recruited participants and 1,879 non-recruited participants.

Our recruited participants started the assignment with a writing exercise. Half of them just had to write about their day. The other half had to write about things that made them angry. This was to test how emotion impacts decision-making. They then watched the videos.

All participants were shown one video at a time. They had to share how confident they were, from 50–100%, that the video was real or a deepfake. The videos we used included four videos of Kim Jong-un and Vladimir Putin, two of which were fake. We then showed participants what the model predicted and allowed them to change their answers. This way we could see whether human decision-making was impacted by machine predictions. The machine learning model assessed 50 videos from a **dataset** of 4,000 videos.

In both experiments, we included **interventions**. These included **obstructed** faces and inverted videos – videos shown upside down. We did this to see if the way our brain naturally recognizes faces changes our ability to identify a deepfake.

Figure 1:
Here are two images used in Experiment 1. One of these images is from the original video and one is from the deepfake. We asked the participants to identify which of these two videos was the deepfake. Can you guess which image is the deepfake? Image: PNAS, CC BY-NC-ND

Results

→ **Experiment 1**

The leading machine learning model correctly identified 65% of the videos in the dataset (Figure 2).

82% of participants outperformed the machine learning model.

Participants were better at detecting fakes in high-quality videos. When the video was inverted or of low quality, they were 5% less accurate.

→ **Experiment 2**

Recruited participants identified deepfakes 66% of the time.

Non-recruited participants identified deepfakes 72% of the time.

Please see Figure 2 on page 3

The leading machine learning model correctly identified 80% of videos when shown 50 videos from the dataset.

When shown the entire dataset of 4,000 videos, the model achieved 65% accuracy.

Participants who wrote about things that made them angry were quicker to judge a real video as a deepfake.

The machine learning model was better at detecting low-quality deepfakes such as grainy, blurry, and inverted videos.

Participants were more likely to change their answer if it was different to the model's answer. This usually made them more accurate, except in the case of the political leader videos. The machine learning model was incorrect about both real videos of the political leaders. Meanwhile, 64% of participants correctly identified the videos of political leaders.

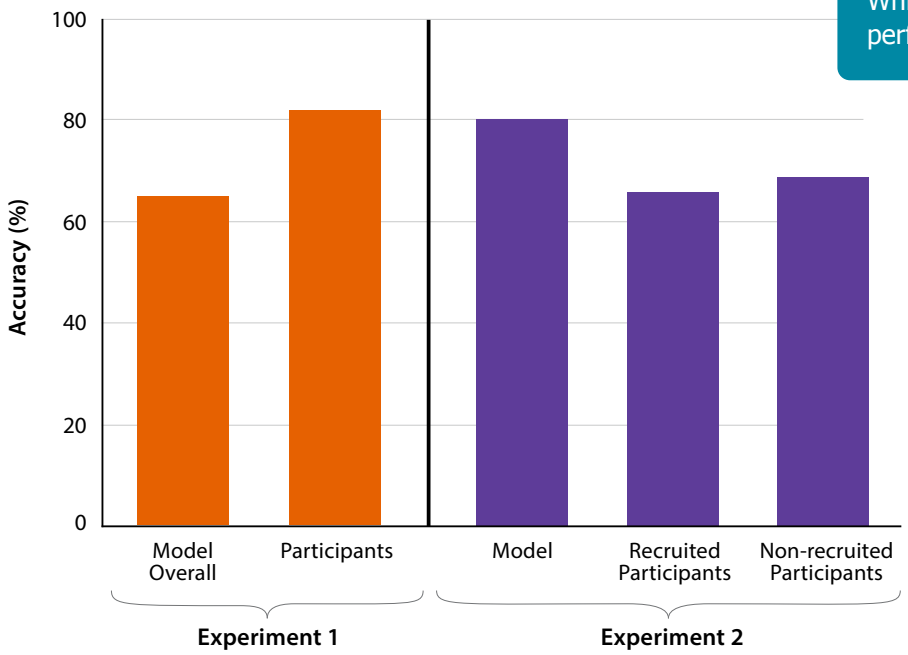


Figure 2.
Human performance compared to model performance

Discussion

Our results suggest that people are as good at identifying a deepfake as the leading machine learning model. Participants were better than the model when it came to the four videos of Kim Jong-un and Vladimir Putin. This could be because we can think critically about the content beyond the visual clues. The machine learning model was pretty certain that the authentic videos were deepfakes. This leads to questions about the model's ability to analyze the context of a video.

Emotion does seem to impact human decision-making by decreasing our accuracy. Therefore, deepfakes that provoke emotion may be harder for us to detect.

In Experiment 2, participants could change their initial answer after seeing the model's response. This helped participants improve their accuracy in most cases. However, it also led them to change their correct answers. This suggests that human-machine collaboration might not lead to more accurate results.

Conclusion

The good news is people appear to be good at detecting deepfakes. The bad news is that deepfakes will likely get more difficult to detect over time. However, we can take action. Humans are better at thinking critically about the content and context of the video. Always think before sharing viral videos, especially if they make us angry or upset! We can tell our friends and family to be careful, too.

Glossary of Key Terms

Artificial intelligence (AI) - machines and systems that can perform certain tasks. If these tasks were done by humans, they would need the thinking and learning skills we call 'intelligence'.

Dataset - a collection of data.

Deepfakes - images and videos that depict events that have not occurred.

Deep learning - pattern recognition that is based on neural networks.

Intervention - the act of purposely changing a condition within the experimental design to test something more specific.

Machine-learning model - a pattern-matching algorithm that learns from data.

Obstructed - blocked. In our research, this was something in the way of the faces, making them hard to see.

Check your understanding

1. How does emotion impact people's ability to accurately detect a deepfake?
2. What type of deepfakes were humans better able to detect than machine learning programs?
3. What type of deepfakes were machine learning programs better able to detect than humans?
4. How do you think scientists could use both AI and human detection abilities to create an accurate deepfake detection model?
5. Can you think of some situations where someone would create a deepfake video of a known person? In groups, discuss why someone might do this – try and think of both positive and negative situations.

REFERENCES

Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard (2021) *Deepfake detection by human crowds, machines, and machine-informed crowds.* Proceedings of the National Academy of Sciences.
https://www.pnas.org/doi/10.1073/pnas.2110013119

University of Virginia: What the heck is a deepfake?
https://security.virginia.edu/deepfakes

TIME for Kids: Fakeout
https://www.timeforkids.com/g56/fakeout-2/

Acknowledgment:
This article's adaptation was supported by the Akamai Foundation.