jea

Scholastic Journalism Week

Feb. 19-23, 2024

HERE TO STAY

Journalism Next

Rationale: With the advancement and intersection of technology, artificial intelligence and social media, student journalists need to have the skills and tools needed to be discerning and questioning consumers of information and media.

Essential Questions: How will journalism, news, and information continue to evolve as “times” and technology are changing? What are our responsibilities before we share information?

Learning Objectives: Students will pull from their prior knowledge and collaborate effectively with peers. Students will learn the tools to discern between fake videos and real videos. Students will research unethically sound news. Students will recognize unreliable and biased sources and spot deep fakes.

Vocabulary:

- Deepfake - According to TechTarget, a deepfake is created using artificial intelligence to swap someone's face with another or manipulate audio and make it look like they said or did something they didn't.

- Misinformation - According to UNESCO, misinformation is misleading information created or disseminated without manipulative or malicious intent. Both are problems for society, but disinformation is particularly dangerous because it is frequently organized, well-resourced, and reinforced by automated technology.

Duration: One activity could take place over one day but could be extended over multiple days.

Procedure:

1.Students will discuss with a group what knowledge they have regarding media, AI, Deepfakes, Misinformation, etc.

2. Students will participate in activities that fit their classroom, i.e. online news/journalism classes could write an article. Yearbook students could create a mod based on deepfake/misinformation. Broadcast Media could create a video about how to spot deepfakes, etc.

Activity 1: In the *Sydney Morning Herald* article, four real images have been duplicated to include false or misleading information. Have students look at the images to try and determine which details have been added; answers are included.

Activity 2: Poynter's MediaWise Teen Fact-Checking Network has a 4-minute video on "Unreal Keanu Reeves," outlining exactly how to tell a deepfake from real news. Students could take the information they learn and write articles about how their readers can spot deepfake or AI-generated videos or content.

Activity 3: Share the two Best of SNO-winning student editorials on deepfake technology.

Materials for Activity 1:

- Full article
  https://www.smh.com.au/world/middle-east/can-you-spot-a-deepfake-gaza-the-latest-outlet-for-ai-deception-20231129-p5enrd.html
  - This article contains an interactive of four real images that have been photoshopped to include fake details. Don't swipe until you are ready to reveal the original image. Answers are included.
- Full article
  https://fortune.com/2023/12/04/deepfakes-israel-hamas-war-ai-detection-tech-startups/

Additional Reading

- Best of SNO student editorial
  ://archeroracle.org/100201/voices/__trashed-11/
- Best of SNO student editorial
  https://edinazephyrus.com/what-the-israel-hamas-war-means-for-journalism/

ARTICLE 1: Can you spot a deepfake?

Can you spot a deepfake? Gaza the latest outlet for AI deception
by David Klepper and James Lemon
Sydney Morning Herald
December 12, 2023 — 10.31am

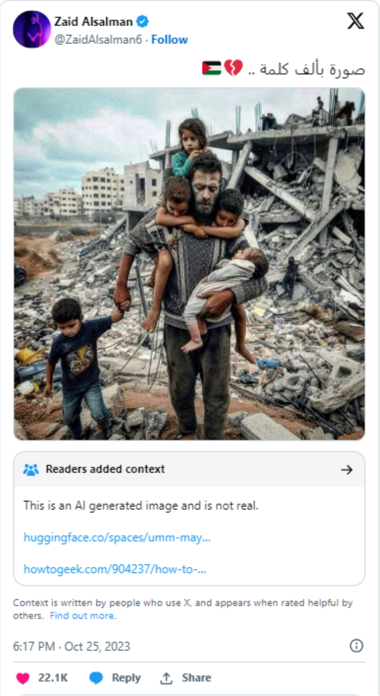
Listen to this article

Washington: Among images of the bombed out homes and ravaged streets of Gaza, some stood out for the utter horror: bloodied, abandoned infants.

Viewed millions of times online since the war began, these images are deepfakes created using artificial intelligence. If you look closely you can see clues: fingers that curl oddly, or eyes that shimmer with an unnatural light – all telltale signs of digital deception.

The outrage the images were created to provoke, however, is all too real.

The viral image below was generated by AI



Zaid Alsalman
@ZaidAlsalman6 · Follow

صورة بألف كلمة .. 💔🇵🇸

Readers added context

This is an AI generated image and is not real.

huggingface.co/spaces/umm-may...

howtogeek.com/904237/how-to-...

Context is written by people who use X, and appears when rated helpful by others. Find out more.

6:17 PM · Oct 25, 2023

Pictures from the Israel-Hamas war have vividly and painfully illustrated AI's potential as a propaganda tool, used to create lifelike images of carnage. Since the war began last month, digitally altered ones spread on social media have been used to make false claims about responsibility for casualties or to deceive people about atrocities that never happened.

While most of the false claims circulating online about the war didn't require AI to create and came from more conventional sources, technological advances are coming with increasing frequency and little oversight. That's made the potential of AI to become another form of weapon starkly apparent, and offered a glimpse of what's to come during future conflicts, elections and other big events.

What is a deep fake and how can you spot one?
"It's going to get worse – a lot worse – before it gets better," said Jean-Claude Goldenstein, CEO of CREOpoint, a tech company based in San Francisco and Paris that uses AI to assess the validity of online claims. The company has created a database of the most viral deepfakes to emerge from Gaza. "Pictures, video and audio: with generative AI it's going to be an escalation you haven't seen."

In some cases, photos from other conflicts or disasters have been repurposed and passed off as new. In others, generative AI programs have been used to create images from scratch, such as one of a baby crying amidst bombing wreckage that went viral in the conflict's earliest days.

Other examples of AI-generated images include videos showing supposed Israeli missile strikes, or tanks rolling through ruined neighbourhoods, or families combing through rubble for survivors.

In many cases, the fakes seem designed to evoke a strong emotional reaction by including the bodies of babies, children or families. In the bloody first days of the war, supporters of both Israel and Hamas alleged the other side had victimised children and babies; deepfake images of wailing infants offered photographic "evidence" that was quickly held up as proof.

The propagandists who create such images are skilled at targeting people's deepest impulses and anxieties, said Imran Ahmed, CEO of the Centre for Countering Digital Hate, a non-profit that has tracked disinformation from the war. Whether it's a deepfake baby, or an actual image of an infant from another conflict, the emotional impact on the viewer is the same.

'Detection and trying to pull this stuff down is no longer the solution. We need to have a much bigger solution.'

David Doermann, AI expert
The more abhorrent the image, the more likely a user is to remember it and to share it, unwittingly spreading the disinformation further.

"People are being told right now: look at this picture of a baby," Ahmed said. "The disinformation is designed to make you engage with it."

Around the world a number of start-up tech firms are working on new programs that can sniff out deepfakes, affix watermarks to images to prove their origin, or scan text to verify any specious claims that may have been inserted by AI.

How AI is leading the fight against disinformation in Ukraine, Gaza
"The next wave of AI will be: how can we verify the content that is out there? How can you detect misinformation? How can you analyse text to determine if it is trustworthy?" said Maria Amelie, co-founder of Factiverse, a Norwegian company that has created an AI program that can scan content for inaccuracies or bias introduced by other AI programs.

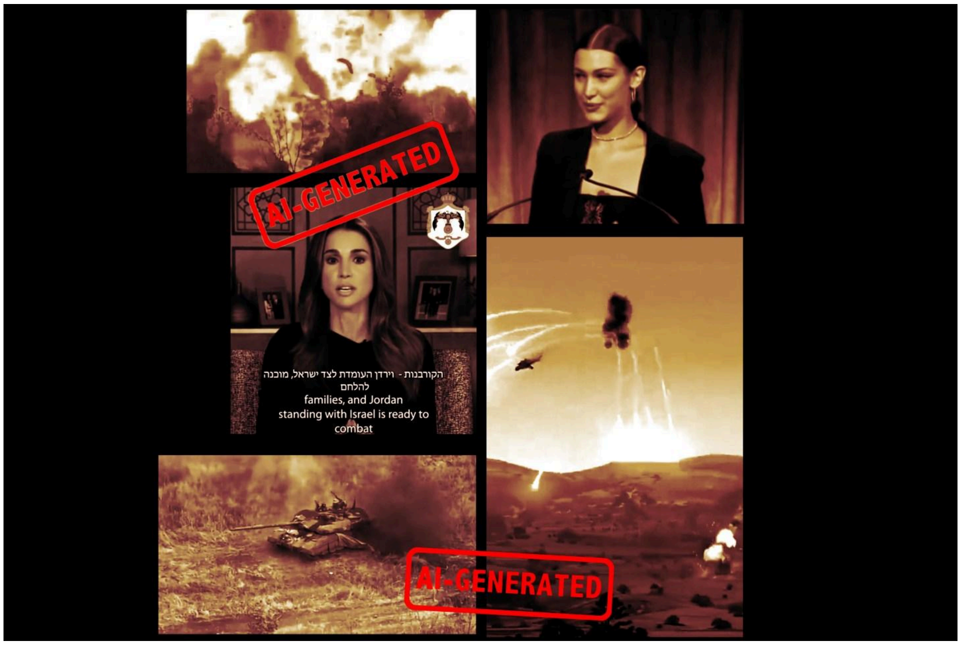
Such programs would be of immediate interest to educators, journalists, financial analysts and others interested in rooting out falsehoods, plagiarism or fraud. Similar programs are being designed to sniff out doctored photos or video.

While this technology shows promise, those using AI to lie are often a step ahead, according to David Doermann, a computer scientist who led an effort at the US Defence Advanced Research Projects Agency to respond to the national security threats posed by AI-manipulated images.

“Every time we release a tool that detects this, our adversaries can use AI to cover up that trace evidence,” said Doermann. “Detection and trying to pull this stuff down is no longer the solution. We need to have a much bigger solution.”

ARTICLE 2: Deepfakes are another front in the Israel-Hamas War

Deepfakes are another front in the Israel-Hamas war that risk unleashing even more violence and confusion in the future: ‘This is moving incredibly fast’
by Vivienne Walt
Fortune Magazine
December 4, 2023 at 4:15 PM EST


AI-GENERATED
הקורבנות - וירדן העומדת לצד ישראל, מוכנה
להלחם
families, and Jordan
standing with Israel is ready to
combat
AI-GENERATED

Deepfakes have played a starring role in the Gaza war, including manipulated videos featuring Jordan’s Queen Rania (middle left), and fashion model Bella Hadid (top right), along with repurposed battle footage from the Ukraine war (bottom left) and the video game Arma 3 (bottom right). SCREENSHOTS CLOCKWISE FROM TOP LEFT: @GLOOOUD/X, VERIFY/YOUTUBE, @AG_JOURNALIST/X, @GLOOOUD/X, THE QUINT/YOUTUBE

As dawn broke on Oct. 7, air-raid sirens blasted out across Tel Aviv, sending Michael Matias and his girlfriend catapulting out of bed and down to the bomb shelter in their apartment building. Inside, the messages on their cell phones revealed the horror that had set off the alarm: Hamas gunmen were waging mass slaughter on Israelis and seizing hundreds of hostages, less than an hour’s drive from where they huddled in safety.

In his stunned aftershock, Matias was struck by the implications for his tech startup. Late last year he had launched Clarity, an artificial intelligence company focused on detecting deepfakes in election campaigns; he believed such disinformation was an urgent threat to democracies. But with about 1,200 Israelis dead on Oct. 7 and the country at war, that mission would need to wait for now. He scrambled an emergency meeting that morning with Clarity's team, to trigger an action plan. "We said to each other, 'Our technology is going to be very meaningful here,' " says Matias, Clarity's CEO.

In the chaos following the massacres—the deadliest in the country's 75-year existence—Israel blocked Gaza's supplies of water, food, and electricity and dropped thousands of bombs on what it claimed were Hamas targets in the packed coastal enclave that's home to 2 million Palestinians. As the humanitarian crisis spiraled into a full-blown disaster, and more than 10,000 Palestinians killed, including civilians, in a month, macabre images flooded TV and phone screens, setting off spasms of rage and fraught protests worldwide. This was a war fought not only with munitions, but with information—both real and fake. And along with the outrage on both warring sides, some people raised questions about whether the gruesome scenes were even real, or whether the images were so-called deepfakes that had been created with the help of artificial intelligence.

In fact, AI-generated fakes have grown increasingly difficult to discern as the technology has improved. That has left startups like Clarity fighting "a cat and mouse game," Matias says.

The question is, will the cat or the mouse win in the end? Some fear the ease of generative technology will let malicious users outsmart far less nimble tech companies and governments attempting to rein them in. "This is moving incredibly fast," says Henry Ajder, a Cambridge, U.K.–based consultant on AI technologies for the British government along with businesses including Adobe and Facebook parent Meta. "What is it going to look like in 20 years, or 10 years?" Ajder says. "The tools are going to get more and more accessible. We could be fundamentally unprepared when we do see a lot of deepfakes."

Propaganda, the saying goes, is as old as war. But since OpenAI launched its first version of AI chatbot ChatGPT last year, the explosive popularity of generative artificial intelligence has turbocharged regular users' ability to create their own narratives. The implications are relatively trivial when it involves Pope Francis in a puffer jacket or Kim Kardashian as a bus driver. But in a war, deepfakes can be used to sow confusion with potentially life-and-death consequences.

Within days of Russia invading Ukraine in February 2022, setting off Europe's biggest land war in generations, Ukrainian President Volodymyr Zelensky appeared on Facebook, telling his soldiers to surrender. On Twitter days later, Russian President Vladimir Putin also commanded his forces to lay down arms. Twitchy mannerisms and blurry camerawork quickly exposed both videos as fake. But it signaled what could happen one day, once AI tools improve.

Now that day has come, with generative AI tools easily accessible, whether in a government office or at home. The technology has uncorked a slew of manipulated content that would previously have required skilled techies to produce. In Israel's war, it has come from all quarters, according to Layla Mashkoor of the Atlantic Council's Digital Forensic Research Lab, part of a U.S. think tank that closely tracks social media sites in the conflict. People on both sides have spread deepfakes, she says, citing a pro-Israel Instagram account that featured an AI-manipulated image of crowds of Israelis cheering soldiers from their balconies. To many people, the tsunami of information flashing by on phone screens has made everything seem unreliable. "For even authentic images, there is a counterclaim, so it's very difficult for people to find clarity," Mashkoor says.

When Matias chose the name Clarity for his AI startup, it was that exact problem he aimed to fix. With the outbreak of war, the team swung into action. As hundreds of videos and photographs hit the internet, some shot by Hamas during its attack, or by Gaza residents in devastated neighborhoods, international media organizations began emailing Clarity for help in weeding out deepfakes, according to Matias, who declined to let Fortune reveal his clients.

The war spawned a wealth of fakery online: Fashion model Bella Hadid apologizing for her pro-Palestinian sentiments was in fact synthesized audio. Jordan's Queen Rania saying her country was “standing with Israel” was AI-generated audio added atop an appearance she had made on CNN. A Hamas video supposedly showing its fighters destroying an Israeli tank was in reality from the Ukraine war. And another video, showing Hamas downing two Israeli helicopters? That was lifted from the Arma 3 video game.

Clarity, headquartered in Palo Alto, Calif., and Tel Aviv, has also partnered in the war with Israeli intelligence agencies to detect which war footage is real or fake. Almost all of Clarity’s team, including Matias, spent years honing their skills in the Israeli Defense Forces elite tech units during their compulsory national service. Matias himself served in the hyper-selective intelligence unit known as 8200, whose members have launched countless global startups after their military service.

With that pedigree, Matias has tapped a network of seasoned founders for both advice and capital. “Clarity is innovating on a new frontier,” says Udi Mokady, an angel investor in Clarity—and another 8200 veteran—who founded CyberArk, an identity-security company. Now CyberArk’s executive chairman, Mokady predicts the market for deepfake detection will rise sharply. “The awareness grew dramatically overnight with ChatGPT,” he says, “where people can create deepfakes from their homes.”

Rather than certifying content purely true or purely fake, Clarity instead uses AI to track a range of data points in a video, like a person’s facial tics or voice cadence, then uses neural networks to place them on a scale of certainty shown on a dashboard, from green to red. “It’s AI versus AI,” Matias says. Some, like the video showing the leader of the Iran-backed organization Hezbollah condemning the Oct. 7 attacks, are clearly deepfakes (“unfortunately,” quips Matias). But other videos are not so clear-cut and require human judgment, rather than AI-trained machines.

Clarity isn’t alone in its fight against deepfakes. Several other similar startups have launched over the past few years, including Reality Defender in New York and Sentinel in Estonia. Tech giants are also focusing on fakery: Beginning in January, Meta will require political campaigns to flag any AI-generated content on Facebook ads, while Google now includes data about the origins of images.

Governments are also trying to take on deepfakes. In October, President Joe Biden issued an executive order on “safe, secure, and trustworthy” AI, which delegated the Commerce Department to find ways to authenticate content as real, and watermark media created with AI. But it came without threats of sanctions for those who violate the rules.

In the same month, officials agreed to similar measures during a U.K. summit on AI of 28 countries, including China. And in the 27-nation European Union—where an AI Act that would enforce safety standards and fund startups is inching its way through the labyrinthine process—top officials want to slap giant fines on social media platforms that fail to crack down on disinformation.

“The consensus increasingly is that there are catastrophic risks to be posed by AI,” says Ajder, the AI consultant in the U.K. “They want to avoid a situation where the Wild West we’ve seen over the last 18 months is perpetuated, in a way where the stakes just get ever higher.”

The stakes could hardly be higher than Israel’s fierce war on Gaza, with U.S. warships stationed close offshore and Iran and Lebanon poised for possible battle. Clarity has been grappling with the implications through weeks of war, as the team analyzes hundreds of grueling videos and photos.

The effect of watching often horrifying videos on Clarity’s small team is clear enough, and Matias says that they will surely need post-traumatic therapy after the war ends. “We knew we were entering a war that is highly personalized,” he says, adding the work has left his team feeling “a deep emotional load, and an incredible sense of importance.”

Detective force

As deepfakes proliferate, both corporate giants and startups have raced to detect them. These are some of the companies involved:

DeepMedia

Launched in 2017 by Stanford and Yale University graduates specializing in AI. From its headquarters in Oakland, it works with the U.S. Department of Defense, the United Nations, and tech companies to spot fake content, using neural network processing.

Reality Defender

A Manhattan-based company founded in 2021 by former Goldman Sachs executive Ben Colman, who says it is crucial to stop a deepfake in its tracks before it goes viral. The company raised $15 million in an October funding round, and aims to roll out new tools that can spot manufactured voices in real time.

Intel

Introduced a deepfake detector, FakeCatcher, in 2022, which it says can analyze videos in real time and deliver 96% accurate results within milliseconds. One rare tool is analyzing the blood flow in the pixels, gauging whether the image depicts a live person.

A version of this article appears in the December 2023/January 2024 issue of Fortune with the headline, "Going to war against deepfakes."