Detecting deepfakes: how can we ensure that generative AI is used for good?

The field of generative artificial intelligence (AI) is advancing at an astronomical pace. As a result, deepfakes – manipulated pieces of media using generative AI technology and designed to trick their viewers – are becoming more convincing, prevalent and problematic. **Professor Siwei Lyu**, based at **University at Buffalo, The State University of New York** in the US, is determined to halt the advance of deepfake media and ensure that generative AI is used for the good of society.

Professor Siwei Lyu

SUNY Empire Innovation Professor, Department of Computer Science and Engineering, University at Buffalo, The State University of New York, USA

Fields of research
Media forensics, generative AI, computer vision, machine learning

Research project
Developing techniques for detecting and mitigating deepfake media

Funder
US National Science Foundation (NSF)

This work is/was supported by the NSF of the USA, under award number SaTC-2153112. The contents are solely the responsibility of the authors and do not necessarily represent the official views of NSF.

Talk like a ...

media forensics researcher

Artefacts — irregularities or uncommon features in manipulated or synthesised media content that expose their unauthentic nature

Artificial intelligence (AI) — computer algorithms that can learn from data and experience to perform tasks which would usually require a human level of intelligence

Deepfake — digital media content created or manipulated using generative AI, most commonly used for images and videos

Disinformation — false information that is deliberately fabricated and spread to deceive or mislead people

Generative AI — a type of artificial intelligence capable of creating a wide variety of data, such as images, videos, audio, text and 3D models

Media forensics — the subfield of AI that aims to detect and mitigate manipulated or synthesised media using AI algorithms

How would you feel if you were scrolling through Instagram or TikTok and came across a video of yourself that you had not created or were even aware of? What if the video showed you doing something you would never do or saying something that you disagree with?

In March 2022, just a month after Russia launched its invasion of Ukraine, a video was released showing the Ukrainian president, Volodymir Zelensky, calling for his troops to surrender. Upon closer inspection, the video was quickly identified as a deepfake, a piece of media modified with generative AI models and designed to trick the viewer.

Luckily, the deepfake of Zelensky was fairly easy to spot as the image of his head and his voice were clearly inauthentic. However, just a year and a half after its release, advancements in generative AI have made deepfake videos much harder to identify. The improving quality of deepfake media is making many people nervous about the potential harm that the technology could cause.

Professor Siwei Lyu, a computer scientist at University at Buffalo, The State University of New York, is a leading expert in the field of generative AI and media forensics. Much of Siwei's research focuses on the detection of deepfake media, although his expertise spans the whole field of generative AI.

What is generative AI?

In order to understand what generative AI is, we first need to clarify what we mean by the term 'AI'. Artificial intelligence, or AI, refers to algorithms and systems that can learn, predict and, in some cases, create. Algorithms and tools that have this creative ability are known as generative AI, and they can be used to fabricate media such as images, videos and audio.

Recent advancements in the field mean that generative AI systems can now learn by analysing photos, audio and videos that are widely available on the internet. As a result, it has become cheaper and

© ozrimoz/Shutterstock.com

easier for users to create convincing fake media, created to deceive the viewer, in large quantities. "Since 2015, generative AI technologies have experienced astronomical development, with new methods showing up on a monthly or even weekly basis," says Siwei.

Although generative AI is often used to create deepfake media, researchers in the field are convinced that the technology can also be used for good. For example, a new generative AI algorithm is being used to help rehabilitate stroke patients who have lost their ability to speak. The algorithm translates the patient's brain activity into simulated speech that resembles their old voice. Such technology will have a hugely positive impact on the patient's everyday life.

What are the dangers of deepfakes?

Deepfakes can pose significant threats to society. "Personal security risks emerge as deepfakes can be used to fabricate convincing yet false representations of individuals," explains Siwei. "This can lead to reputational damage and psychological distress for anyone whose personal images or videos can be accessed by people using AI to produce deepfakes."

At a wider scale, deepfakes can impact our democratic processes by spreading **disinformation**. "By creating illusions of an individual's presence and activities that did not occur in reality, deepfakes can influence our opinions or decisions." This can become a particular issue around election time, when people are trying to decide who to vote for based on things they see and read online.

As the field of generative AI continues to develop, deepfakes will become more powerful but so will our methods of detecting them. Siwei compares the situation to developing vaccines to combat a virus. As vaccines improve, viruses evolve new ways of overcoming them, so scientists have to work even harder to develop new vaccines.

How can we detect deepfakes?

Although deepfakes are becoming more convincing, it is possible to detect them simply by looking closely to find **artefacts**. The generative AI models that create deepfakes are trained on enormous amounts of data, but they have no understanding of the laws of physics or how the human body works. This means that they often make mistakes, which attentive observers can pick up on.

For example, AI models are often trained on thousands of images of human faces downloaded from the Internet however, almost all of the people in these images will have their eyes open. As a result, the simulated people in many deepfake videos do not blink, which is something that viewers can spot relatively easily. Other errors to look out for are missing teeth when someone is talking, hands that have the wrong number of fingers, and the reflections/light flecks in eyes pointing in different directions.

Unfortunately, as generative AI models become more sophisticated, deepfakes are becoming harder to spot, even if you know what to look for. To address this, Siwei has developed tools that can detect deepfakes that even a trained human would struggle to identify.

Whilst genuine videos come directly from the real world, deepfakes have been stitched together from a variety of sources. The process of stitching the deepfakes together is not perfect and often leaves minor inconsistencies or visual errors, known as artefacts or noise, that do not appear in the original media. Siwei likens his detection tools to X-ray scanners that can see inside deepfakes and uncover artifacts that might not be visible to the human eye.

How else can we stop deepfakes?

Detection methods can be effective at spotting deepfakes, but what if we could prevent deepfakes from being created in the first place? "Unlike detection methods, the pre-emptive approach directly obstructs the training or generation of deepfakes," explains Siwei. "One idea is to 'poison' the would-be training data by adding specially designed patterns."

These patterns disrupt the generative AI's training process and cause the AI models to create low-quality deepfakes. For example, one of Siwei's methods deliberately adds unnoticeable patterns into the original media, making it hard for generative AI models to detect faces. This could prevent one of the most common methods of creating deepfakes, which involves inserting someone's face into a video that they did not actually appear in.

What does the future of generative AI hold?

"Looking into the future, we will continue to see accelerated development of deepfake technology," says Siwei. He predicts that we will see new generative AI models that can create deepfakes that look more realistic and have fewer artifacts. "In the not-too-distant future," continues Siwei, "ordinary users may have access to more easy-to-use and ready-made tools to manipulate media in the same way that they use Photoshop to edit images today."

Although deepfake technology will become more accessible, Siwei is not too worried about the consequences. "I don't think it will necessarily lead to a future of dooming dystopia or an apocalypse!" he says. "Human brains are unbelievably flexible and as we are seeing more synthetic media created with generative AI, our ability to cope with them will evolve." For example, in the early 2000s, spam emails were becoming a huge problem until we learnt how to spot them and developed technologies to help protect us.

"We must continue developing deepfake forensic methods that are more effective, efficient, robust and explainable," says Siwei. In doing so, he, and other researchers like him, will help to halt the progress of deepfake technologies and ensure that generative AI is used for the good of society.

© Rroselavy/Shutterstock.com

About *media forensics*

The field of generative AI research is growing rapidly as these tools become increasingly powerful and sophisticated. We are always learning more about the potential impacts of generative AI, both good and bad, so there is always new research to be done.

Rewards

"The most rewarding experience of my research is when I see that I can have a real impact on society," explains Siwei. "The unique nature of media forensics puts me in the frontier of one of the most vexing challenges of this century." Siwei often meets with journalists, government officials and members of the public to help them understand the situation and provide expert advice.

"On a more personal level, I like puzzles and games as pastimes," says Siwei, who sometimes imagines that his research in media forensics is a game of cat and mouse that he plays against the people who make deepfakes. "The intention to beat the other party is one important motivation for me to continue exploring my research work," he says.

Opportunities

"The current situation with AI, and particularly with generative AI, has taught us that technologies must be developed with adequate consideration to their potential impacts on humanity and society," explains Siwei. "These tools are going to be woven into the fabric of our lives."

Over the coming decades, we will begin to see the development of 'responsible AI' and 'trustworthy AI'. These new generations of AI will focus on human and social impacts, instead of solely focusing on the technology. "This will create many challenging and exciting research opportunities for the younger generation of computer scientists," says Siwei.

Pathway from school to *media forensics*

- Siwei says a solid foundation in mathematics is critical. In particular, calculus, linear algebra, and probability are vital to computer science.
- More advanced topics such as signal processing, optimisation, and statistical analysis are stepping stones to understanding AI and machine learning.
- Programming is another indispensable skill. You must be able to write code in various programming languages, such as Python and Java, and think using programming terms.

Explore careers in *media forensics*

- Societies such as the Institute of Electrical and Electronics Engineers (**www.ieee.org**), the Association for Computer Machinery (**www.acm.org**), the International Association for Pattern Recognition (**www.iapr.org**), and the Association for the Advancement of AI (**www.aaai.org**) provide education and careers resources and often have student memberships.
- Joining these societies and taking advantage of their seminars, tutorials and conferences can be a great way to keep up with the latest developments in the field. They are also good for building a professional network and making connections.
- Contact generative AI researchers and scientists and ask them questions, both about their research and their career progression.

Meet *Siwei*

When I was in high school more than three decades ago, I had a classmate who had a CASIO programmable calculator. It had some primitive graphics, and you could program simple games on it, like canon ball shooting. I was immediately fascinated and would work on it for hours, exchanging playing cards and other goodies just to be able to play with it longer. That was my first interest in computer science, and it made me want to write computer programs.

Focusing on media forensics happened by chance. When I started my graduate study at Dartmouth College in 2001, my attention was caught by a lecture given by Professor Hany Farid, who is a pioneer in media forensics. In the lecture, Hany discussed the research that he was working on at the time, which needed a classification algorithm. I happened to be studying a more recent classification method, so, after the lecture, I went to his office and offered to try my method on his problem. Fortunately, my method worked and improved the performance significantly. That was the basis of my first academic paper and the starting point of a long and continuous collaboration with Hany, who supervised my PhD thesis.

My encounter with deepfakes started at a conference I attended with Hany in 2018. Deepfakes were a new phenomenon at the time, and I was instantly interested. Detecting deepfakes falls into my research areas of machine learning and media forensics.

My proudest achievement is that I was the first researcher who understood the significance of combatting deepfakes. I developed the world's first dedicated detection method of deepfake videos by capturing an artifact in 2018.

My career goals are to continue research in mitigation technologies of deepfakes and to train the next generation of researchers to work in this crucially important and exciting research area. I will also put more emphasis on public outreach and education programmes on this topic, to improve society's overall awareness and resilience to the new challenges posed by generative AI.

Siwei's *top tip*

Stay curious — research in computer science can be a lengthy, difficult process. The curiosity to solve a problem is often the only driving force that helps us find the answers.

Meet *Shan*

Dr Shan Jia is a research scientist in Siwei's lab.

Funder
US National Science Foundation (NSF)

I participated in a research project on image restoration during my junior year in college. I really admired the magic of using computer science to solve real-world problems. That experience inspired me to delve deeper into the field.

Participating in an exchange student programme at West Virginia University had a profound impact on me. This experience shaped the research direction of my career, motivating me to contribute to the fields of information security and media forensics.

One of my eureka moments came during my visit to Madame Tussauds Hong Kong. The highly realistic wax figure faces made me realise the potential threat of such 3D face models as face presentation attacks to face recognition systems. This insight led me to research the threat posed by 3D face spoofing attacks and to develop novel detection methods to identify these spoofed faces.

My primary role in Dr Lyu's lab involves conducting research in media forensics and taking charge of media synthesis research projects. Additionally, I provide guidance to graduate students and lead new research initiatives as an assistant lab director.

One of the primary technical challenges I face is ensuring the performance and reliability of our media forensics methods in practical scenarios. This demands enhancing the generalisation ability of our algorithms and effectively addressing potential biases in detection. The field of computer science is evolving rapidly, so ensuring that I'm up-to-date with the latest trends and technologies is both a challenge and a necessity.

I have authored/co-authored 11 academic papers within this role. I have served as a reviewer for over 10 international journals and conferences and co-supervised more than 4 students.

In the short term, I'm focused on contributing to top-tier publications. In the long run, I hope to leverage my expertise to provide effective information security solutions for practical scenarios.

Shan's *top tip*

Nurture authentic curiosity. It fosters strong motivation and critical thinking, which are crucial for doing research.