Deep Fake Crimes

Dark side of AI & ML revolutions

By Urvi Chheda @ THM,Mumbai

$whoami

- Consultant
- Lead Auditor
- Certified Ethical Hacker
- HTB SME
- Security Researcher

in Urvi Chheda

What are Deep Fakes?

- A deepfake is a type of artificial intelligence (AI)-generated content that uses machine learning algorithms to create hyper-realistic videos, audio, or images.
- These digital forgeries are so convincing that they can make it appear as though someone is saying or doing something they never actually did.
- The main ingredient in deep fakes is machine learning, which has made it possible to produce deep fakes much faster at a lower cost.
- To make a deep fake video of someone, a creator would first train a neural network on many hours of real video footage of the person to give it a realistic “understanding” of what he or she looks like from many angles and under different lighting.
- Key developments in the advancement of deep fakes:

  - 2014: Introduction of GANs (Generative Adversarial Networks).

  - 2017: Emergence of deep fake tools like FakeApp.

  - 2020: Increased use in misinformation and political manipulation

WHAT ARE DEEPFAKES

Deepfakes, as the name suggests, are deceptive images or audiovisuals created using AI and 'deep learning', where an algorithm is trained on original material and it then learns to produce something that closely resembles that original content. A picture or video of a person, for example, can be seamlessly doctored to look like a different person. Even audio can be manipulated to mimic a person's voice

How Deep Fakes are created

Training with Autoencoders

Autoencoder Structure: At the heart of a deepfake is an autoencoder. An encoder takes an image and reduces it to a small number of variables. The decoder then takes this compact representation and expands it into the original image.

Training Algorithm: The training algorithm feeds another image to the network, and the whole process repeats. It can take thousands of iterations to produce an autoencoder that does a good job of reproducing its own input.

Switching Decoders

Encoding and Decoding: Once you’ve trained a pair of autoencoders, you switch decoders. You encode a photo of, say, A and B, but then use the decoder of A in the decoding of B. The result is a reconstructed photo of B, but with the same head pose and expression as in the original photo of A.

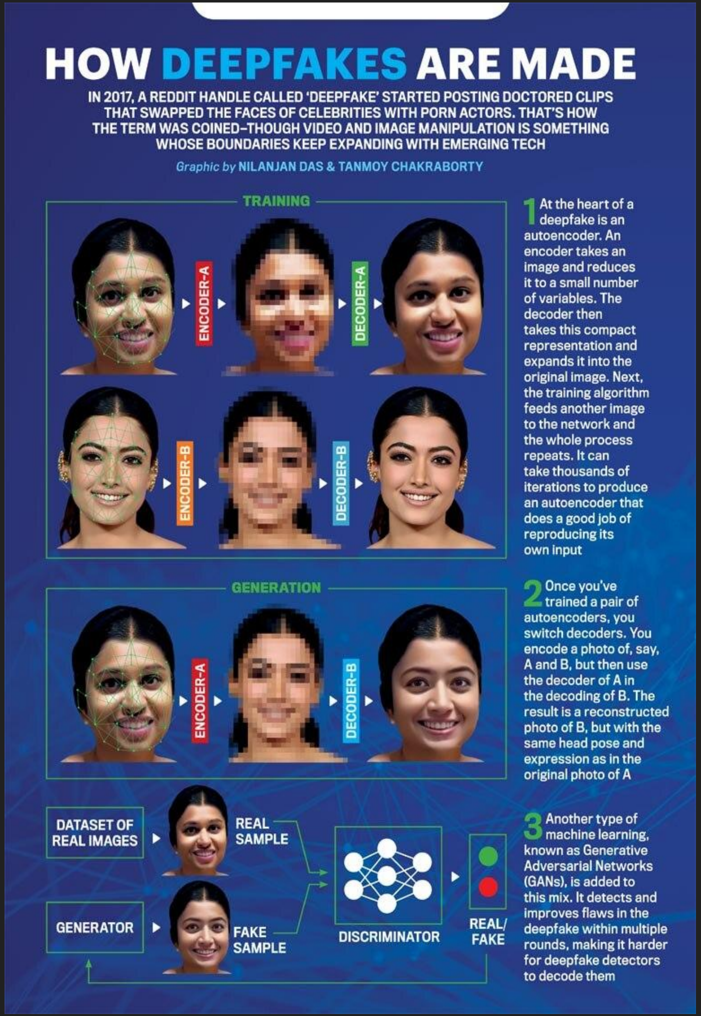
Training:

Encoder A → Compressed Image → Decoder A → Original Image

Encoder B → Compressed Image → Decoder B → Original Image

Generation:

Encoder A → Compressed Image → Decoder B → Reconstructed Image of B with Pose of A



HOW DEEPFAKES ARE MADE

IN 2017, A REDDIT HANDLE CALLED ‘DEEPFAKE’ STARTED POSTING DOCTORED CLIPS THAT SWAPPED THE FACES OF CELEBRITIES WITH PORN ACTORS. THAT’S HOW THE TERM WAS COINED–THOUGH VIDEO AND IMAGE MANIPULATION IS SOMETHING WHOSE BOUNDARIES KEEP EXPANDING WITH EMERGING TECH

Graphic by NILANJAN DAS & TANMOY CHAKRABORTY

1 At the heart of a deepfake is an autoencoder. An encoder takes an image and reduces it to a small number of variables. The decoder then takes this compact representation and expands it into the original image. Next, the training algorithm feeds another image to the network and the whole process repeats. It can take thousands of iterations to produce an autoencoder that does a good job of reproducing its own input

2 Once you’ve trained a pair of autoencoders, you switch decoders. You encode a photo of, say, A and B, but then use the decoder of A in the decoding of B. The result is a reconstructed photo of B, but with the same head pose and expression as in the original photo of A

3 Another type of machine learning, known as Generative Adversarial Networks (GANs), is added to this mix. It detects and improves flaws in the deepfake within multiple rounds, making it harder for deepfake detectors to decode them

Adding GANs: Another type of machine learning, known as Generative Adversarial Networks (GANs), is added to this mix. It detects and improves flaws in the deep fake within multiple rounds, making it harder for deep fake detectors to decode them.

GANs:

Dataset of Real Images → Generator → Fake Sample

Real Sample and Fake Sample → Discriminator → Real/Fake

Tools

1.DeepFaceLab: One of the most popular and advanced deepfake creation tools.Supports face swapping, de-aging, and many other effects. It provides detailed tutorials and an active community.

2.Faceswap: An open-source tool for creating deepfakes.Allows for face swapping, and includes tools for face detection, alignment, training, and conversion.

3.Zao: A mobile app that became popular for its easy-to-use interface for creating deepfakes.Allows users to swap their face with celebrities in video clips with just a few clicks.

4.FakeApp: One of the original deepfake creation tools, though less advanced than newer options.Provides basic face swapping functionality.

5.JigSaw (formerly Lyrebird AI): A platform for creating synthetic voices, which can be combined with deepfake video tools for more realistic deepfakes.Voice cloning, customizable voice parameters.

6.Deepfakes Web : Deepfakes Web is probably the best online deepfake maker. It uses a machine-learning algorithm to face-swap the subject in a video with another.

Criminal Applications of Deep Fakes

1. Identity Theft and Fraud

- **Financial Scams**: Deepfake technology can be used to create videos or audio of individuals, such as CEOs or financial officers, to deceive employees into transferring money or sharing confidential information.
- **Social Engineering**: Deepfake videos or audio can be used to impersonate individuals, gaining unauthorized access to secure systems or manipulating people into disclosing sensitive information.

2. Disinformation and Political Manipulation

- **Fake News**: Deepfakes can create realistic-looking videos of politicians or public figures making false statements or engaging in compromising activities, which can be used to spread misinformation and influence public opinion.
- **Election Interference**: Manipulated videos can be used to undermine candidates or political parties, affecting election outcomes.

3. Extortion and Blackmail

- **Compromising Videos**: Deepfakes can be used to create fake explicit content featuring victims, which can then be used to extort money or other favors by threatening to release the fake content publicly.
- **Threats**: Creating threatening messages or actions attributed to individuals can be used to coerce them into compliance with demands.

4. Defamation and Harassment

- **Reputation Damage**: Creating and distributing deepfake content that portrays individuals in a negative light can severely damage their personal and professional reputation.
- **Cyberbullying**: Deepfake technology can be used to create embarrassing or harmful content to harass or bully individuals online.

5. Corporate Espionage

- **False Communications**: Deepfakes can be used to impersonate executives in video calls or voice messages, manipulating employees into divulging trade secrets or other sensitive corporate information.
- **Sabotage**: Misleading videos or audio can be used to create internal discord or confusion within a company, potentially leading to operational disruptions.

6. Fraudulent Activities

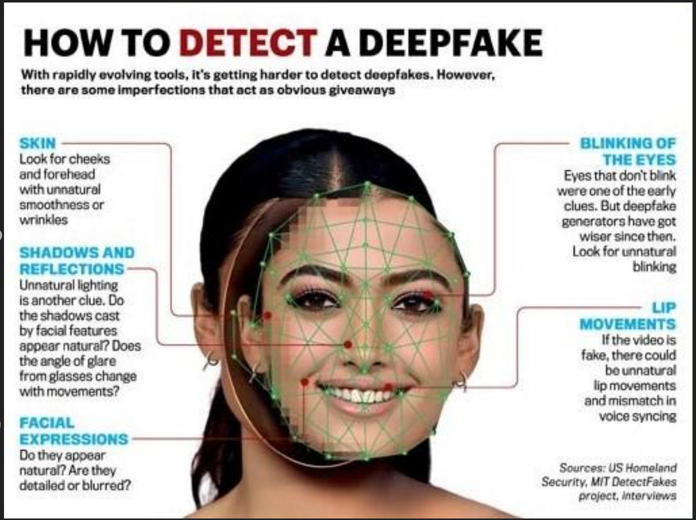
- **Insurance Fraud**: Creating fake videos of accidents or incidents to claim insurance money.
- **Fake Evidence**: Using deepfakes to create false evidence in legal cases, potentially impacting the outcomes of trials and investigations.

7. Fake Celebrity Content

- **Unauthorized Use**: Deepfakes can create fake endorsements or advertisements featuring celebrities without their consent, misleading consumers and violating intellectual property rights.
- **Pornographic Content**: Creating explicit content featuring celebrities, which can harm their reputation and violate their privacy.

How to identify Deep fake

1. Pay attention to the face. High-end DeepFake manipulations are almost always facial transformations.
2. Pay attention to the cheeks and forehead. Does the skin appear too smooth or too wrinkly? Is the agedness of the skin similar to the agedness of the hair and eyes? Deep Fakes are often incongruent on some dimensions.
3. Pay attention to the eyes and eyebrows. Do shadows appear in places that you would expect? Deep Fakes often fail to fully represent the natural physics of a scene.
4. Pay attention to the glasses. Is there any glare? Is there too much glare? Does the angle of the glare change when the person moves? Once again, Deep Fakes often fail to fully represent the natural physics of lighting.
5. Pay attention to the facial hair or lack thereof. Does this facial hair look real? Deep Fakes might add or remove a mustache, sideburns, or beard. But, Deep Fakes often fail to make facial hair transformations fully natural.
6. Pay attention to facial moles. Does the mole look real?
7. Pay attention to blinking. Does the person blink enough or too much?
8. Pay attention to the size and color of the lips. Does the size and color match the rest of the person's face?



HOW TO DETECT A DEEPFAKE

With rapidly evolving tools, it's getting harder to detect deepfakes. However, there are some imperfections that act as obvious giveaways

SKIN
Look for cheeks and forehead with unnatural smoothness or wrinkles

SHADOWS AND REFLECTIONS
Unnatural lighting is another clue. Do the shadows cast by facial features appear natural? Does the angle of glare from glasses change with movements?

FACIAL EXPRESSIONS
Do they appear natural? Are they detailed or blurred?

BLINKING OF THE EYES
Eyes that don't blink were one of the early clues. But deepfake generators have got wiser since then. Look for unnatural blinking

LIP MOVEMENTS
If the video is fake, there could be unnatural lip movements and mismatch in voice syncing

Sources: US Homeland Security, MIT DetectFakes project, interviews

Detection Strategies

1. **Deep Learning-Based Detection**
    - **Convolutional Neural Networks (CNNs)**: These models can be trained to distinguish between real and fake images or videos by learning from large datasets of both types.
    - **Recurrent Neural Networks (RNNs)**: Particularly effective for analyzing video sequences, as they can capture temporal inconsistencies that may indicate deep fake content.
2. **Forensic Techniques**
    - **Image Forensics**: Analyzing image artifacts, such as inconsistencies in lighting, shadows, or reflections, which may reveal signs of tampering.
    - **Video Forensics**: Examining frame rates, compression artifacts, and inconsistencies in motion or lip-syncing that are indicative of deepfake manipulation.
3. **Biometric Analysis**
    - **Facial Recognition**: Comparing facial features and expressions with known genuine images to identify anomalies.
    - **Voice Analysis**: Using spectrograms and other audio analysis tools to detect synthetic audio artifacts that are typical in deep fake audio.
4. **Blockchain and Digital Watermarking**
    - **Blockchain**: Recording the provenance of digital content on a blockchain to verify its authenticity and detect any alterations.
    - **Digital Watermarking**: Embedding invisible watermarks in videos and images that can be checked to verify the authenticity and detect tampering.
5. **Generative Adversarial Networks (GANs) for Detection**
    - **GAN-Based Detection**: Training GANs specifically to detect deepfakes by generating both real and fake examples and learning to distinguish between them.

Deep Fake Detector Tools

1. HyperVerge
2. Sentinel
3. Sensity
4. Oz Liveness
5. WeVerify
6. Intel's FakeCatcher
7. Microsoft Video AI Authenticator
8. Deepware
9. Phoneme-Viseme Mismatches
10. DuckDuckGoose

10 Best AI DeepFake Detector Tools
Sentinel
Oz Liveness
intel's FakeCatcher
Deepware
DuckDuckGoose
HyperVerge
Sensity
WeVerify
Microsoft Video AI Authenticator
Phoneme-Viseme Mismatches

Deep Fake Legal and Regulatory Measures in India

Existing Legal Framework

1. **Information Technology Act, 2000 (IT Act)**
   - **Section 66D**: Punishes cheating by personation using computer resources, which can be applied to cases involving deep fakes.
   - **Section 67**: Prohibits publishing or transmitting obscene material in electronic form, which can include deep fake pornography.
   - **Section 69**: Empowers the government to intercept, monitor, or decrypt any information generated, transmitted, received, or stored in any computer resource.
2. **Indian Penal Code (IPC), 1860**
   - **Section 463 and 465**: Deals with forgery and punishment for forgery, applicable to deep fake videos created with the intent to deceive.
   - **Section 469**: Addresses forgery for the purpose of harming reputation, which can include deep fakes.
   - **Section 500**: Deals with defamation, which can be applied to deep fakes that harm an individual's reputation.

Regulatory Measures and Initiatives

1. **Personal Data Protection Bill, 2019 (PDP Bill)**
   - This proposed bill aims to protect the privacy of individuals' data and includes provisions that can indirectly address issues related to deep fakes by regulating the processing of personal data.
2. **Intermediary Guidelines and Digital Media Ethics Code Rules, 2021**
   - These rules under the IT Act mandate social media intermediaries and digital media platforms to exercise greater diligence in moderating content, which includes addressing manipulated media and deep fakes.
   - **Due Diligence Obligations**: Intermediaries are required to remove or disable access to unlawful information, including deep fakes, upon receiving actual knowledge or on being notified by the appropriate government or its agency.
3. **Fact-Checking Initiatives**
   - Several government and private initiatives are in place to combat misinformation and deep fakes. Organizations like the Press Information Bureau (PIB) and various fact-checking portals work to verify and debunk deep fakes.
4. **Awareness and Training Programs**
   - The government and various organizations conduct awareness campaigns and training programs to educate the public, law enforcement, and other stakeholders about the dangers of deep fakes and ways to identify and counter them.

Strategies to Protect Yourself from Deep Fakes

Share with Care

- Limit personal information shared online.
- Adjust social media settings to restrict who can see your content.
- Only trust and accept follow/friend requests from known individuals.

Enable Strong Privacy Settings

- Use privacy settings to control access to your content.
- Restrict visibility of photos, videos, and sensitive data.
- Minimize publicly available material.

Watermark Photos

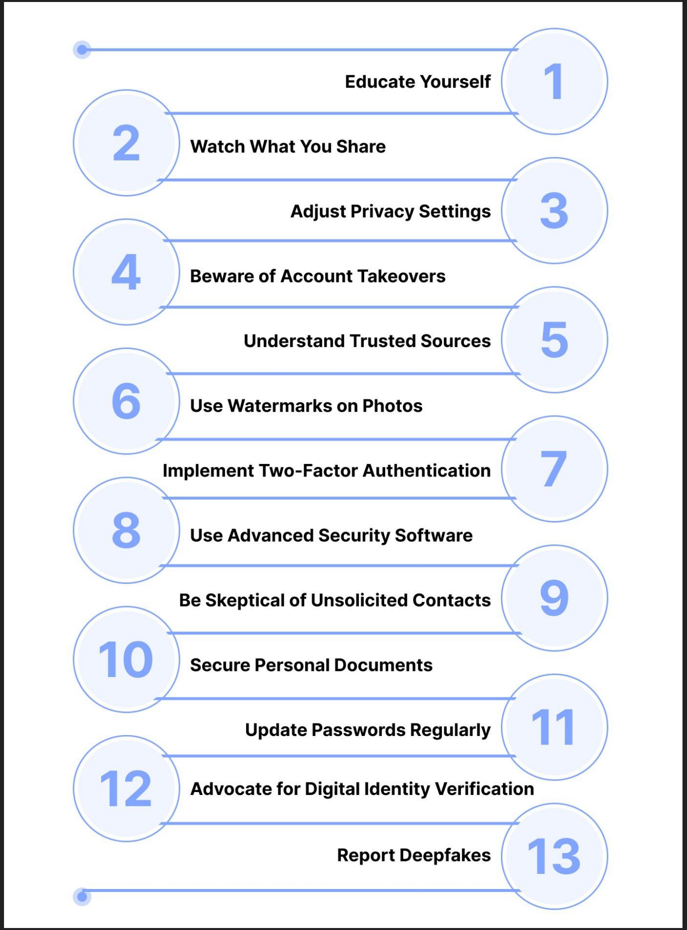
- Use digital watermarks on shared images/videos to discourage misuse.

Learn About Deep Fakes and AI

- Stay informed about AI developments.
- Recognize potential red flags in suspicious content.

Use Multi-Factor Authentication (MFA)

- Implement MFA for all accounts for extra security.
- Examples include facial scans, codes, or authentication apps.



1. Educate Yourself
2. Watch What You Share
3. Adjust Privacy Settings
4. Beware of Account Takeovers
5. Understand Trusted Sources
6. Use Watermarks on Photos
7. Implement Two-Factor Authentication
8. Use Advanced Security Software
9. Be Skeptical of Unsolicited Contacts
10. Secure Personal Documents
11. Update Passwords Regularly
12. Advocate for Digital Identity Verification
13. Report Deepfakes

1. **Use Strong, Unique Passwords**
   - Create passwords at least 16 characters long with a mix of characters.
   - Use a password manager with MFA enabled.
2. **Keep Software Up to Date**
   - Regularly update devices and software with the latest security patches.
   - Enable automatic updates.
3. **Avoid Phishing**
   - Be cautious with communications from unknown sources.
   - Verify sender identity and avoid clicking suspicious links.

If You Suspect You’re a Victim of a Deep fake

1. **Report Deepfake Content**
   - Report to the platform hosting the content.
   - Inform federal law enforcement.
2. **Consult Legal Advice**
   - Seek help from cybersecurity and data privacy legal experts.
   - Advocate for stronger deep fake prevention laws with elected representatives.

THANK
YOU