(ISC)²
SECURITY CONGRESS
October 10-12, 2022
Caesars Palace, Las Vegas + Virtual

Are Deepfakes Really a Security Threat?

Dr. Thomas P. Scanlon
CERT Data Science
Software Engineering Institute
Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM22-0815

Carnegie Mellon University (CMU)

Pioneering discoveries that enrich the lives of people on a global scale

- Turning disruptive ideas into success through leading-edge research
- 2021 *U.S. News and World Report* rankings
  - #1 in computer engineering, artificial intelligence (AI), cybersecurity, and software engineering
  - #2 in overall computer science
  - #3 in data analytics/science

CMU Software Engineering Institute (SEI)

Bringing innovation to the U.S. Government

- A Federally Funded Research and Development Center (FFRDC) chartered in 1984 and sponsored by the Department of Defense (DoD)
- Leader in researching complex software engineering, cyber security, and AI engineering solutions
- Critical to the U.S. Government's ability to acquire, develop, operate, and sustain software systems that are innovative, affordable, trustworthy, and enduring

The CERT Division: The Birthplace of Cybersecurity

Trusted
Conducting research for the U.S. Government in a non-profit, public-private partnership

Valued
Collaborating with military, industry, and academia globally to innovate solutions

Relevant
Achieving technology and talent results for our mission partners

Can You Spot the Fake?

This Person Does Not Exist…

https://thispersondoesnotexist.com/

https://thisxdoesnotexist.com/

What is MDM?

DHS CISA defines MDM as information activities intended to cause chaos, confusion, and division.

Mis-, **D**is-, **M**al-information

- Misinformation: false information that is shared without intent to harm
- Disinformation: false information deliberately created to mislead of cause harm
- Mal-information: information based on truths but purposefully used out of context to mislead or cause harm

MDM Examples

Mis-, **D**is-, **M**al-information

- Misinformation: Betsy Ross sewed the first American flag
- Disinformation: Operation INFEKTION
- Mal-information: 80% of dentists recommend Colgate

Disinformation and Mal-information are often shared as misinformation

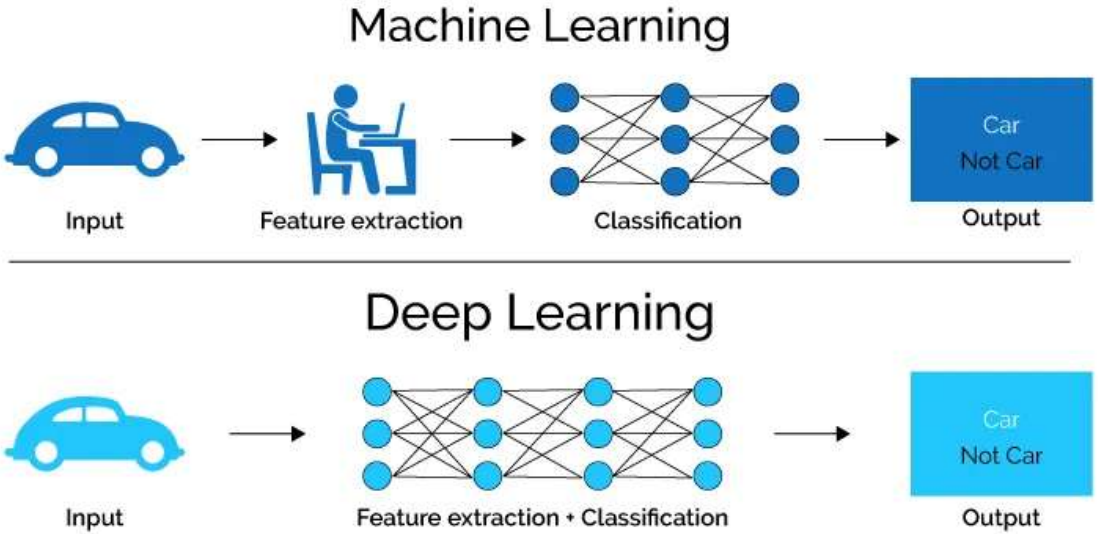
What Is a Deepfake?

- Deepfake = ‘deep-learning’ + ‘fake.’
- ‘deepfake’ originates from a Reddit user, who, in 2017, claimed to have created the method.
- A deepfake can be audio, video, an image, or multimodal.
- It is not the same as using Photoshop.
- Deepfakes are considered disinformation.
  - Or they are combined with disinformation (e.g., profile with deepfake images).

A deepfake is a media file, typically videos, images, or speech representing a human subject, that has been modified deceptively using deep neural networks to alter a person’s identity. Advances in machine learning have accelerated the availability and sophistication of tools for making deepfake content. As deepfake creation increases, so too do the risks to privacy and security.

Deep Learning

Deep learning is machine learning using a neural network.



https://semiengineering.com/deep-learning-spreads/

Deepfake Creation

Main Deepfake Types

- Face Swap
- Lip syncing
- Puppeteering
- Synthetic

Common Deepfake Techniques

- Auto-encoder
- GAN

Deepfake Creation Process

- Extraction
  - -Data collection (source data)
- Training
- Conversion / Generation

Deepfake Creation Process - Extraction

- As a practical matter, need to consider what data sources will provide this data
- A lot of training data is needed
- For images, thousands of images may be necessary
- Just a few video clips can replace thousands of images
- Extraction is the process of extracting individual frames from from video source, identifying faces and aligning them.
- Will need images of source (subject we want to embed) and destination (subject we want to override)

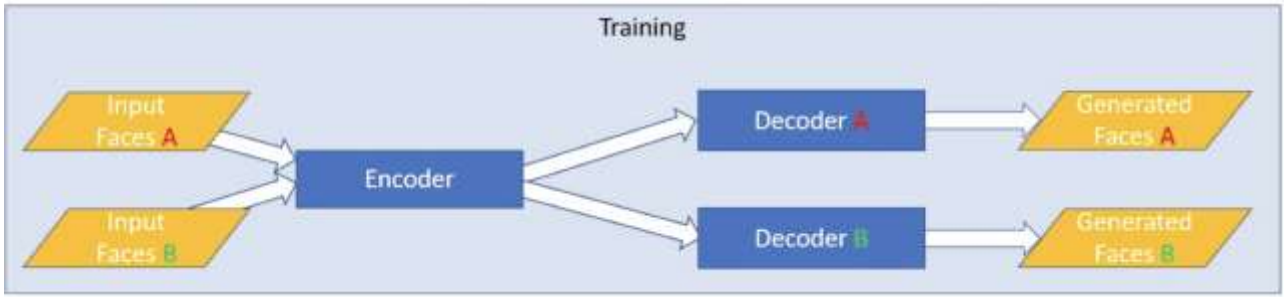
This is different from feature extraction

Deepfake Creation Process - Extraction

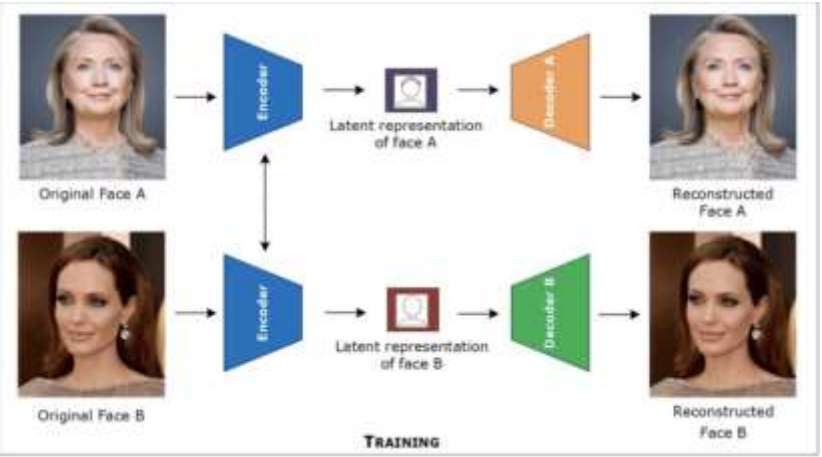
*resizing, normalization, augmentation, etc.

Deepfake Creation Process - Training

Autoencoders
Encoders and Decoders



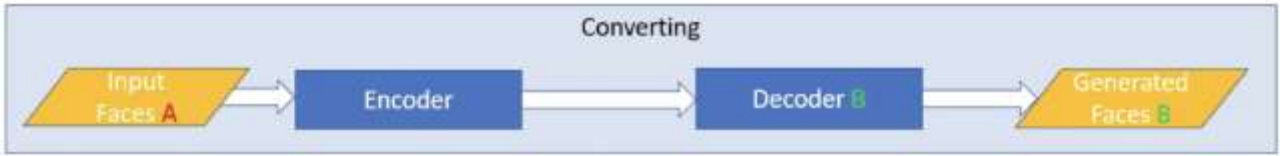
Deepfake Creation Process - Training



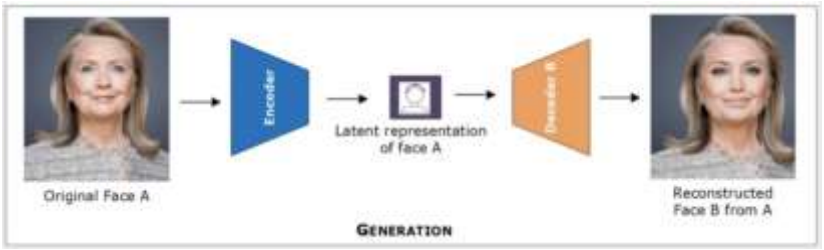
Masood, Momina & Nawaz, Marriam & Malik, Khalid & Javed, Ali & Irtaza, Aun. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward.

Deepfake Creation Process - Generating

Autoencoders
Encoders and Decoders

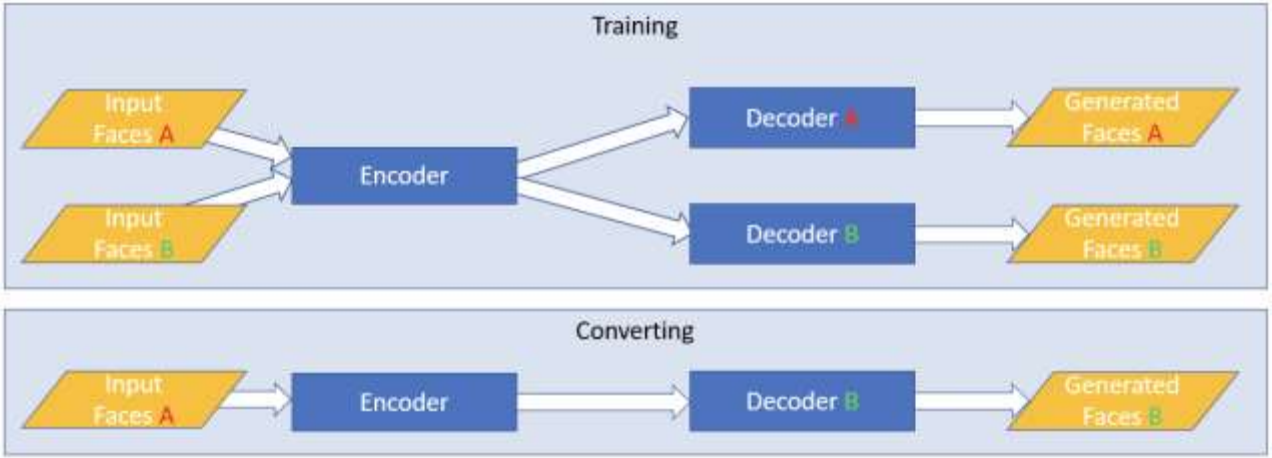


Deepfake Creation Process - Generating

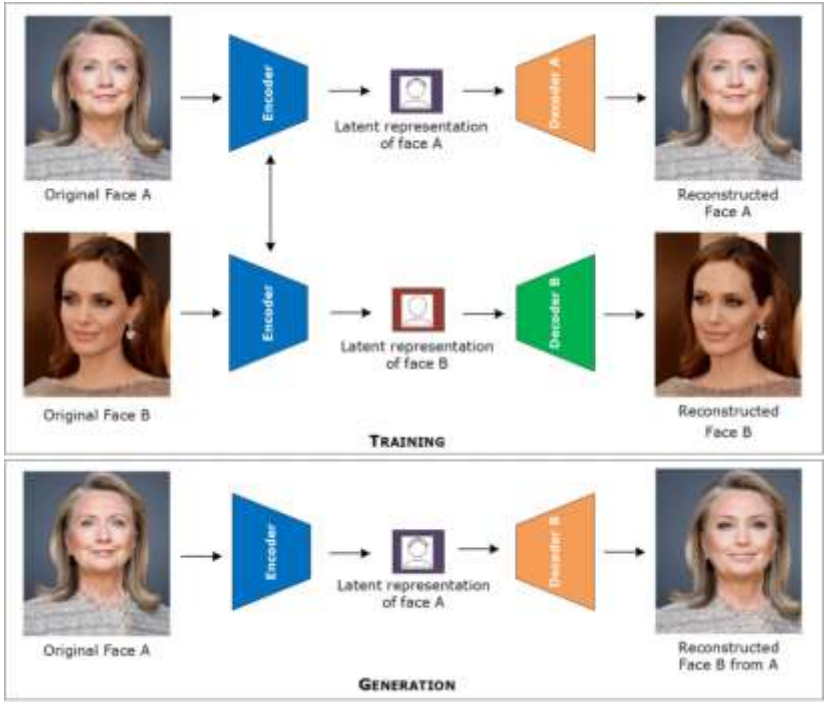


Masood, Momina & Nawaz, Marriam & Malik, Khalid & Javed, Ali & Irtaza, Aun. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward.

Deepfake Creation Process – Auto-encoder



Deepfake Creation Process – Auto-encoder



Masood, Momina & Nawaz, Marriam & Malik, Khalid & Javed, Ali & Irtaza, Aun. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward.

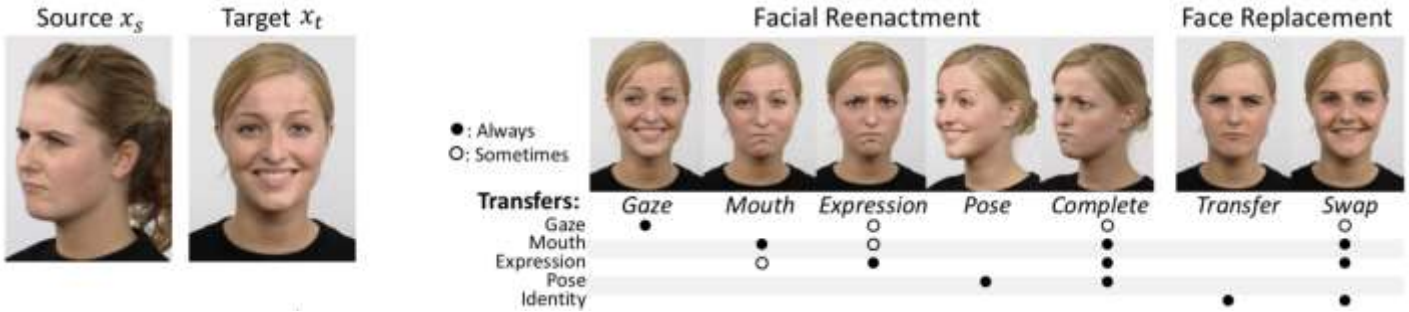
Common Deepfake Creation Activities

Reenactment

A reenactment deepfake is where x_s is used to drive the expression, mouth, gaze, pose, or body of x_t

Replacement

A replacement deepfake is where the content of x_t is replaced with that of x_s, preserving the identity of s.

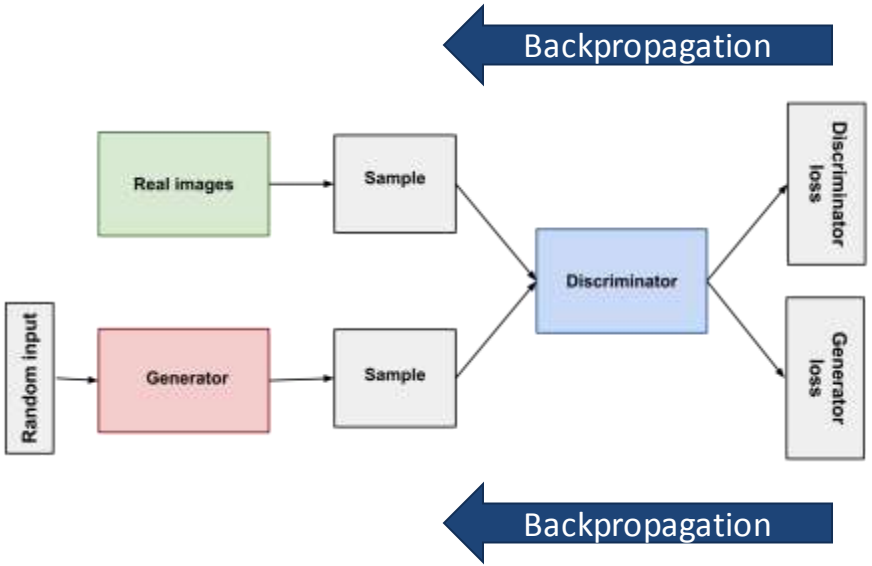


Yisroel Mirsky and Wenke Lee. 2020. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (December 2020), 41 pages

GAN For Deepfake Creation

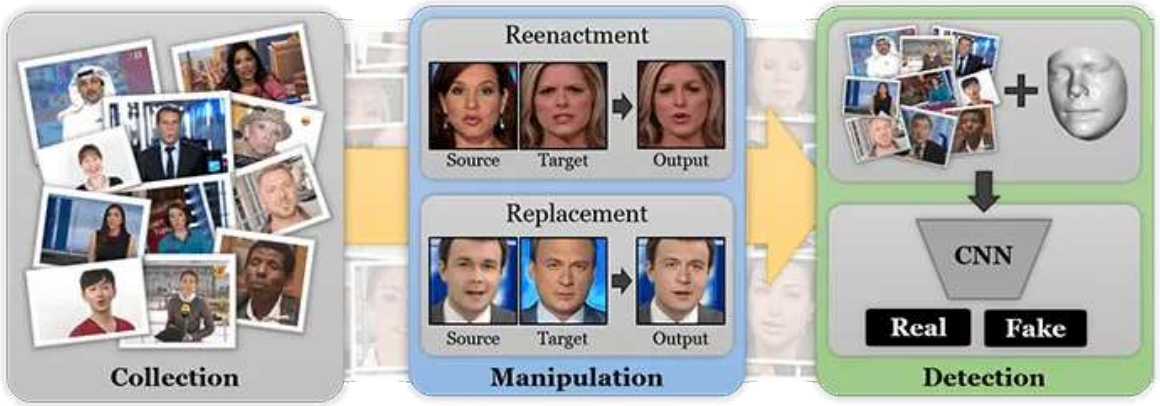
- Generative adversarial network (GAN) was introduced by Ian Goodfellow in 2014.
- GAN is a machine learning (ML) model in which two neural networks compete with each other to improve their predictions.
- There is a generator that tries to create fake images (“forger”) and a discriminator (“detective”) that tries to determine fake images from real images.
- The generator and discriminator are each using a deep neural network (DNN) and can be of same type or different.
- For deepfake creation, they are both often CNN.
- Frequently used for synthetic fakes

Generative adversarial network (GAN)



https://developers.google.com/machine-learning/gan/gan_structure

Deepfake Creation with GAN



https://deepware.ai

These Were Entertaining…

https://www.youtube.com/watch?v=cQ54GDm1eL0

https://www.tiktok.com/@deeptomcruise/video/6933305746130046214?is_copy_url=1&is_from_webapp=v1&lang=en

…but…

Deepfake Nefarious Uses

- scams & hoaxes
- social engineering
- fraud
- identity theft
- political/election manipulation
- forgery
- fake almost anything: pornography, rental ads, dating profiles, LinkedIn accounts, voicemail messages, etc.

Deepfakes for Malicious Use - Examples

- Malicious actors convinced a CEO to wire $243,000 to a scammer's bank account by using deep fake audio[1]
- Symantec reports they have observed at least 3 other deep fake audio cases involving CEOs and CFOs[2]
- Palestinian activists smeared by unknown, deepfaked identity[3]
- Politicians from the UK, Latvia, Estonia and Lithuania tricked by fake meetings with opposition figures [4]

1 - https://www.zdnet.com/article/forget-email-scammers-use-ceo-voice-deepfakes-to-con-workers-into-wiring-cash/
2 - https://www.bbc.com/news/technology-48908736
3 - https://www.reuters.com/article/us-cyber-deepfake-activist/deepfake-used-to-attack-activist-couple-shows-new disinformation-frontier-idUSKCN24G15E
4 - https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition

Deepfakes for Malicious Use – Examples cont.

- Deepfakes replace women on sextortion calls [1]
- Deepfake video of bank president offers false discount [2]
- Deepfakes used to Impersonate a Navy Admiral and Bilk Widow Out of Nearly $300,000 [3]
- AI app used to "undress" women [4]
- Zelensky (Ukraine) deepfaked [5]
- Mayo of Kyiv deepfaked [6]

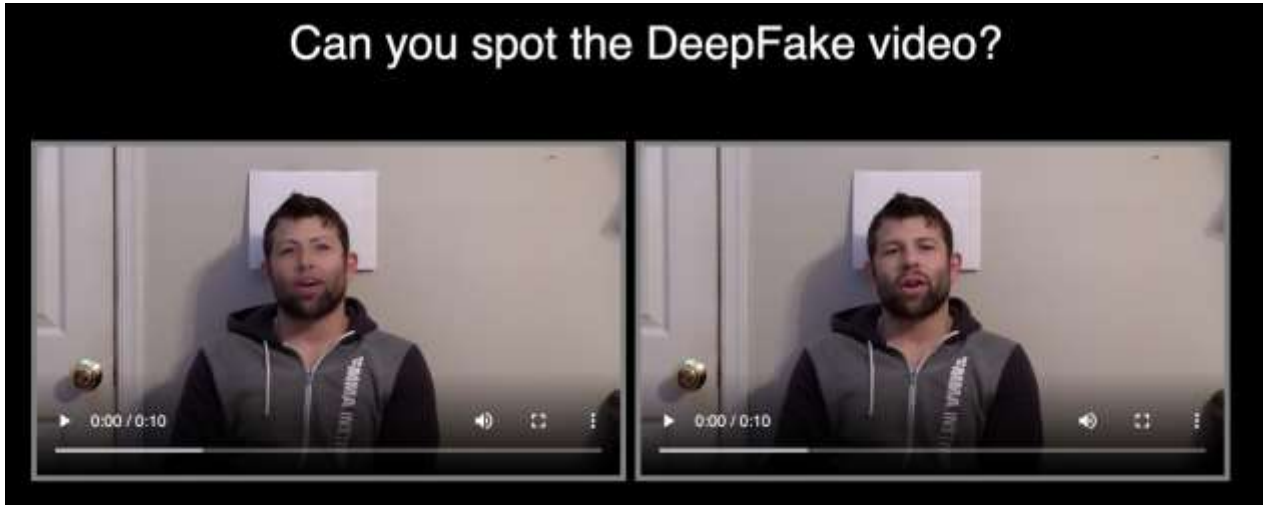
1 - https://timesofindia.indiatimes.com/city/ahmedabad/deepfakes-replace-women-on-sextortion-calls/articleshow/86020397.cms
2 - https://tekdeeps.com/fraudsters-created-a-deepfake-of-oleg-tinkov-dont-be-fooled-by-this-ad/
3 - https://www.thedailybeast.com/romance-scammer-used-deepfakes-to-impersonate-a-navy-admiral-and-bilk-widow-out-of-nearly-dollar300000
4 - https://www.technologyreview.com/2019/06/28/134352/an-ai-app-that-undressed-women-shows-how-deepfakes-harm-the-most-vulnerable/?truid=21defdb9a2d89523a2a6ea4c092cecca&utm_source=the_algorithm&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement&utm_content=10-08-2021
5 - https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia
6 - https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko

Fake deepfakes?

- Mother used deepfake to frame cheerleading rivals [1]
- How misinformation helped spark an attempted coup in Gabon [2]

1 - https://www.bbc.com/news/technology-56404038
2 - https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/

Detecting Deepfakes: The Eye Test



https://www.media.mit.edu/projects/detect-fakes/overview/

Detecting Deepfakes: Practical Cues

- Flickering
- Unnatural movements and expressions
- Lack of blinking
- Unnatural hair and skin colors
- Awkward head positions
- Appears to be lip-syncing
- Oversmoothed faces
- Double eyebrows; raised eyebrows at wrong time; one raised eyebrow
- Glare/lack of glare on glasses
- Realistic appearance of moles; consistent placement of moles
- Earrings—wearing only one or mismatched

Detecting Deepfakes Programmatically

1. Blending (spatial)
2. Environmental (spatial)
   - Lighting—background/foreground differences
3. Physiological (temporal)
   - Generated content lacks pulse, breathing; has irregular eye blinking patterns
4. Synchronization (temporal)
   - Mouth shapes and speech, “B-P-M” mouth closed failure
5. Coherence (temporal)
   - Flickering, predict next frame
6. Forensic (spatial)
   - Generative Adversarial Networks (GANs) leaving unique fingerprints, camera Photo-Response Non-Uniformity (PRNU)
7. Behavioral (temporal)
   - Video versus audio emotions; target mannerisms (> data)

https://dl.acm.org/doi/fullHtml/10.1145/3425780

Deepfake Detection Challenge (DFDC)

- AWS, Facebook, Microsoft, the Partnership on AI's Media Integrity Steering Committee, and other academics created the Deepfake Detection Challenge :

  https://www.kaggle.com/c/deepfake-detection-challenge

- 100,000 deepfake clips (created by Facebook using paid actors) for entrants to test their detectors.
- 2,000 participants from industry and academia, generated more than 35,000 deepfake detection models.
- The best model detected deepfakes from Facebook's collection about 82%of the time; when the same algorithm was run against previously unseen deepfakes, it detected about 65%/

Detecting Deepfakes – Tools

- Microsoft's Video Authenticator Tool

  -detects blending boundaries and grayscale elements that are undetectable to the human eye

- Facebook Reverse Engineering

  -detects digital fingerprints left behind by generative model

- Quantum Integrity

  -determines if images of videos have been manipulated, methods not well documented

DARPA Projects

- Semantic Forensics (SemaFor)

  -semantic detection algorithms, which will determine if multi-modal media assets have been generated or manipulated

  -attribution algorithms will infer if multi-modal media originates from a particular organization or individual

  -characterization algorithms will reason about whether multi-modal media was generated or manipulated for malicious purposes

- Media Forensics (MediFor)

  -developing technologies for the automated assessment of the integrity of an image or video and integrating these in an end-to-end media forensics platform

Current State of Deepfakes

- You don’t need to be a data scientist or AI researcher to create deepfakes; no code/low code options exist.
- Open source Python software such as Faceswap and DeepFaceLab are easy to use, and the deep learning can be treated as a “black box.”
- Motivated parties with more resources can produce fairly strong deepfakes.
- If you are in a cybersecurity role in your organization, there is a good chance that you will be asked about this technology.

Current State of Deepfakes cont.

- Good news: Even using tools that are already built (Faceswap, DeepFaceLab, etc.) it still takes considerable time and graphics processing unit (GPU) resources to create even lower quality deepfakes.*
  - Bad news: Well-funded actors can commit the resources to making higher quality deepfakes, particularly for high-value targets.

- Good news: Deepfakes are principally only face swaps and facial reenactments.
  - Bad news: That is good enough if you can find lookalikes, and eventually the technology capabilities will expand beyond faces.

- Good news: Advancements are being made in detecting deepfakes.
  - Bad news: Technology for deepfake creation continues to advance; it will likely be a never-ending battle similar to malware and anti-virus software.

*High quality deepfakes often require significant non-AI/ML post-processing

Deepfakes Organizational Concerns

Deepfakes are certainly a threat to personal privacy and security, but how can they harm organizations?

- Theft of money
- Fake contracts
- Corporate espionage
- Theft of intellectual property
- Fake virtual employees
- Fraudulent announcements
- Fraudulent advertisements
- Reputational harm

What Can Your Organizations Do?

- Understand the current capabilities for both creation and detection.
- Know what can be done realistically and learn to recognize indicators for fakes.
- Practical ways to defeat current deepfake capabilities – “turn your head”.
- Create a training and awareness campaign for your organization.
- Review business workflows for places deepfakes could be leveraged.
- Craft policies about what can be done through voice or video instructions.
- Establish out of band verification processes.
- Watermark media - literally and figuratively.
- Be ready to combat MDM.
- Use deepfake detection tools (eventually).

Rate the Security Threat

	Consequence				
Likelihood	Insignificant	Minor	Moderate	Major	Critical
Rare	LOW Accept the risk Routine management	LOW Accept the risk Routine management	LOW Accept the risk Routine management	MEDIUM Specific responsibility and treatment	HIGH Quarterly senior management review
Unlikely	LOW Accept the risk Routine management	LOW Accept the risk Routine management	MEDIUM Specific responsibility and treatment	MEDIUM Specific responsibility and treatment	HIGH Quarterly senior management review
Possible	LOW Accept the risk Routine management	MEDIUM Specific responsibility and treatment	MEDIUM Specific responsibility and treatment	HIGH Quartely senior management review	HIGH Quarterly senior management review
Likely	MEDIUM Specific responsibility and treatment	MEDIUM Specific responsibility and treatment	HIGH Quarterly senior management review	HIGH Quarterly senior management review	EXTREME Monthly senior management review
Almost certain	MEDIUM Specific responsibility and treatment	MEDIUM Specific responsibility and treatment	HIGH Quarterly senior management review	EXTREME Monthly senior management review	EXTREME Monthly senior management review

https://www.qgcio.qld.gov.au/information-on/ict-risk-management/ict-risk-matrix

Engage with Us

Download software and tools.

Participate in education offerings.

Attend an event.

Search the SEI Digital Library.

Read the SEI *Year in Review*.

Explore our research and capabilities.

Collaborate with the SEI on a new project.

Contact Us

Carnegie Mellon University
Software Engineering Institute
4500 Fifth Avenue
Pittsburgh, PA 15213
888-201-4479

info@sei.cmu.edu
www.sei.cmu.edu