Don’t Believe Your Eyes (Or Ears):

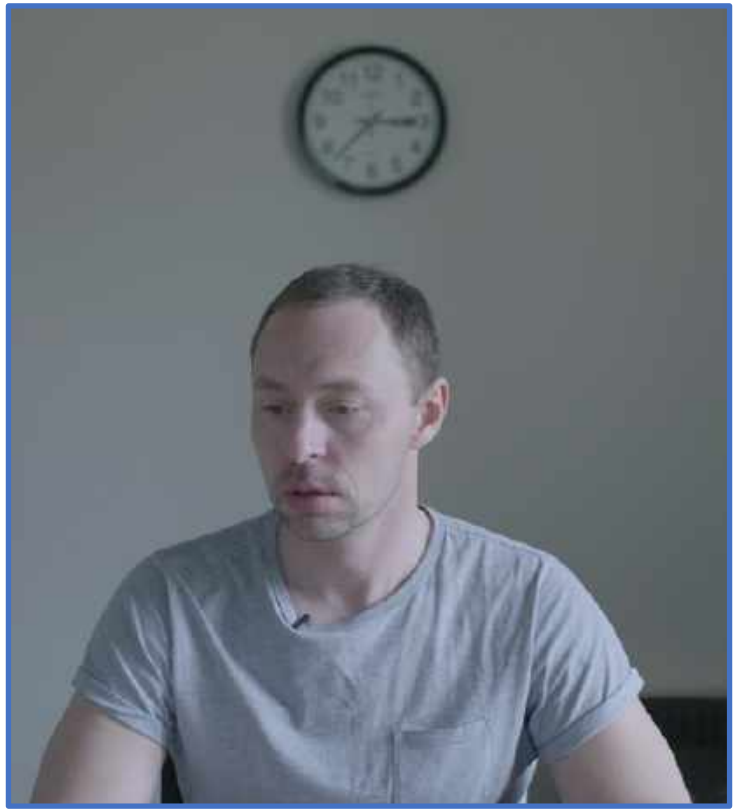
The Weaponization of Artificial Intelligence, Machine Learning, and Deepfakes

Joe Littell

Agenda

- Introduction
  - What is A.I.?
  - What is a DeepFake?
  - How is a DeepFake created?
- Visual Manipulation
- Audio Manipulation
- Forgery
- Data Poisoning
- Conclusion
- Questions

Introduction



Deniss Metsavas, an Estonian soldier convicted of spying for Russia's military intelligence service after being framed for a rape in Russia. (Picture from Daniel Lombroso / The Atlantic)

What is A.I.?
...and what is it not?

General Artificial Intelligence (AI)

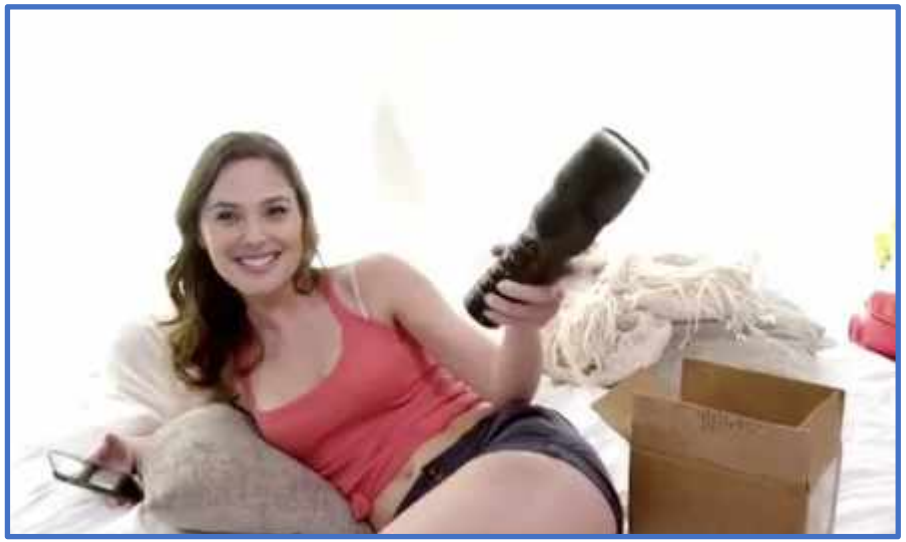
- Machine (or Statistical) Learning (ML) is a subset of AI

- ML works through the probability of a new event happening based on previously gained knowledge (Scalable pattern recognition)

- ML can be supervised, leaning requiring human input into the data, or unsupervised, requiring no input to the raw data.

What is a Deepfake?

faceswap



- Deepfake is a mash up of the words for deep learning, meaning machine learning using a neural network, and fake images/video/audio.
  - Taken from a Reddit user name who utilized faceswap app for his own ‘productions.’

- Created by the use of two machine learning algorithms, Generative Adversarial Networks, and Auto-Encoders.

- Became known for the use in underground pornography using celebrity faces in highly explicit videos.

How is a Deepfake created?

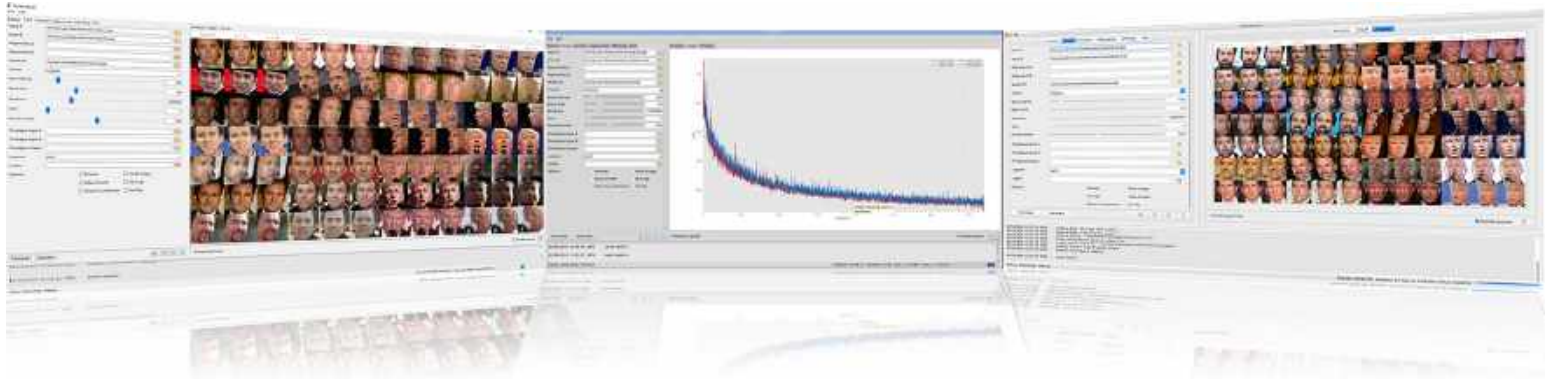
- Deepfakes are generated using Generative Adversarial Networks, and Auto-Encoders.

- These algorithms work through the uses of competing systems, where one creates a fake piece of data and the other is trained to determine if that datatype is fake or not

- Think of it like a counterfeiter and a police officer.

Visual Manipulation

faceswap



Facesswap

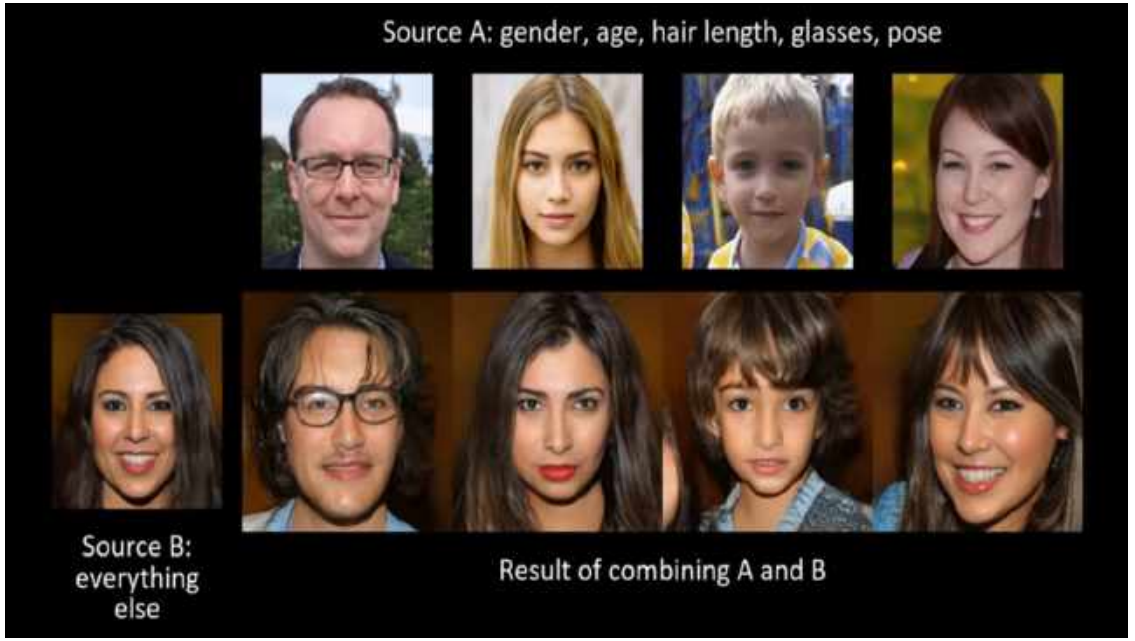
- Created by u/DeepFake on Reddit in early 2018.

- Utilized Auto-Encoders to remove face of a target video and replace it with another in the same position and expression. Similar technology is used in social media filters like Facebook Messenger, SnapChat, and Instagram.

- Most commonly used of the technologies, most videos on Youtube that are Deepfakes created with the faceswap app.

- Opensource and available on Github

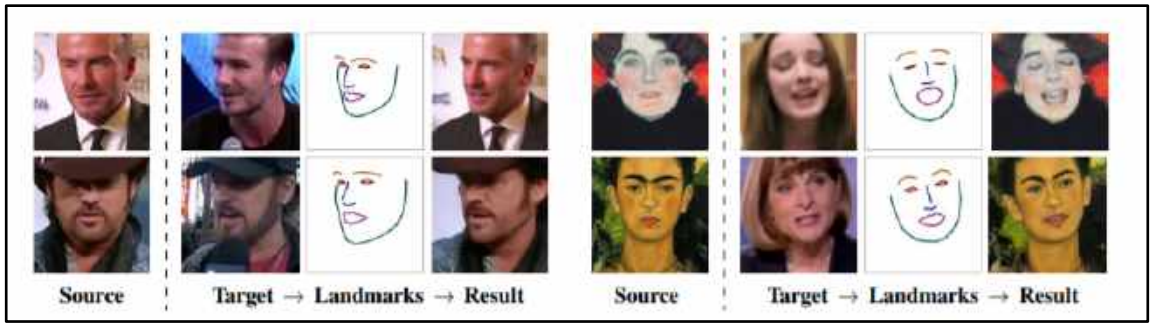
Visual Manipulation

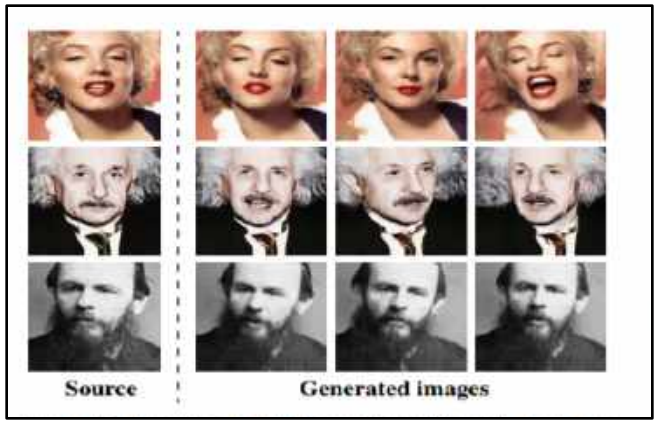


StyleGan

- Created by nVidia Corps in 2018.
- Highly accurate portraits of humans in natural positions based off of baseline images
- Already being used to create realistic fake profiles to infiltrate LinkedIn networks. https://www.apnews.com/bc2f19097a4c4fffaa00de6770b8a60d [October 2019]
- Opensource and available on Github

Visual Manipulation





Few Shot Talking Head Model

- Created by Samsung-Moscow in 2019.
- Overlays target video with a limited number of still images, without deformations associated with older face swapping apps.
- Uses facial landmarks to identify correct movement and features
- Currently proprietary, however the technical paper address the background needed to recreate

Audio Manipulation

Voco

- Created by Adobe in 2016

- Capturing only 20 minutes worth of conversational speaking allows for recreation of a targets voice through text

- After concerns with the potential risk, the project was shelved and remains proprietary for Adobe.

Audio Manipulation

WaveNet

- Created by DeepMind (Alphabet and Google Subsidiary) in 2016
- Similar to Adobe Voco at the time of release, but has surpassed its cababilities
- Much of the code and paper are available opensource, the end product was Integrated with Google Assistant.

Audio Manipulation

Dessa

- Created by AI startup of the same name in 2019
- Created a viral video of Comedian and UFC Announcer Joe Rogan
- Proprietary

Forgery

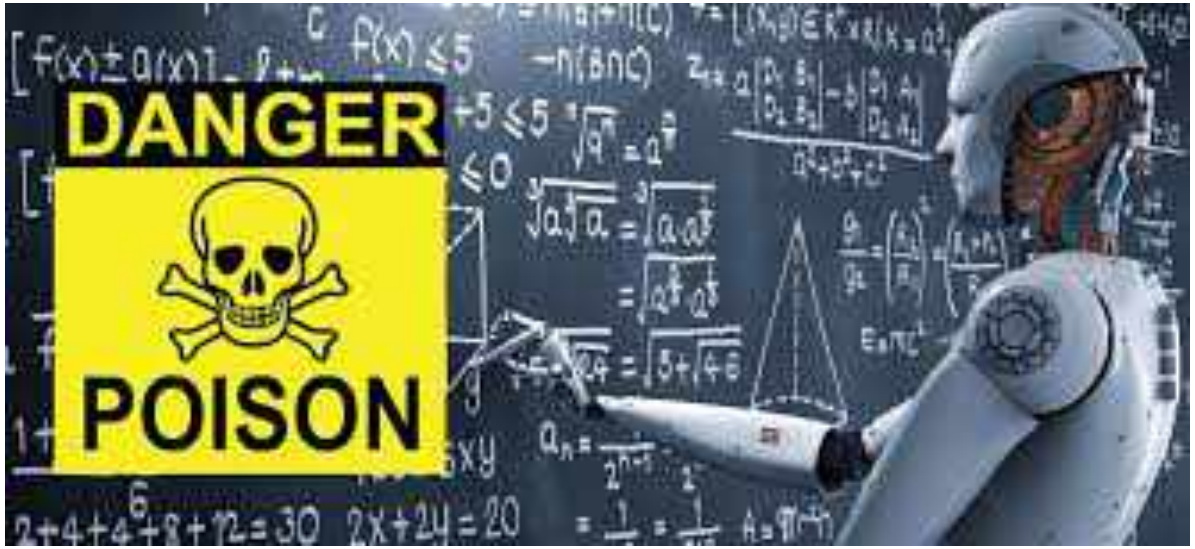
Forgery

- Although much focus of late with regards to deepfakes is on images, video, and audio, creating highly accurate and realistic administrative documents are also a real threat

- While the United States has numerous safeguards to prevent forgeries, Many counties in which US personal operated in do not.

- Many opensource methods already exist

Data Poisoning



Data Poisoning

- With a high use of sensor data for collection, and expansion of the Internet of Things (IoT), the risk to data poisoning is very real
- GANs can be used to create raw input data to directly affect machine learning models that are self learning.
- Many opensource methods already exist

Conclusion

Way Forward

- Currently expertise in the various fields that could combat the new threat of Machine Learning are spread across the DoD, IC, and National Labs.
- Certain groups have established pipelines but no methodology or best practices are shared across the defense enterprise.
- In order to fully leverage American ingenuity, outreach to academia and industry are needed.

Questions?

Joe Littell
Email: Joseph.Littell@duke.edu
Joseph.P.Littell.mil@mail.mil
LinkedIn: https://www.linkedin.com/in/joe-littell-6a374516/