DEEPFAKES

from fake nudes to fake news

Max Benjamin Mayer Rizzuto, B.S.
Science, Technology, and Society
Thesis Advisor: **Dr. Alex Wellerstein**

Abstract

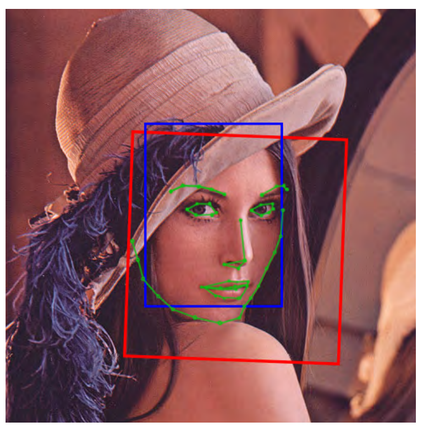
Recent developments in the area of machine learning have made it possible to manipulate video into photorealistic and eerily natural fabrications of human speech and movement with nothing more than a computer and the conviction to do so. Known as deepfakes, this computer generated form of video forgery has threatened the very integrity of video evidence, which has long maintained the gold standard for substantiating truth. Since the code was made freely available in late 2017, the process of producing deepfakes has been developed, proliferated, and democratized by a highly skilled anonymous community online. While many extrapolate the technology's effect to issues of politics and national security, deepfakes impression on identity has already been realized in the creation of what has come to be known as involuntary pornography. The purpose of this paper was to explore deepfakes peculiar origin and development and consider the impact of this unruly technology at the intersection of ethics, identity, and truth.

understanding training data

Deepfake programs produce arrays of images similar to the graphic above. The columns, shown in groups of three, visualizes the neural network's learning process. The from the left, the first image is the the computer's recreation of Wolf Blitzer's face, the second is the computer's attempt at recreating my own face, and the third picture is Wolf Blizer's face with my facial expression. **None of the images shown above are real.**

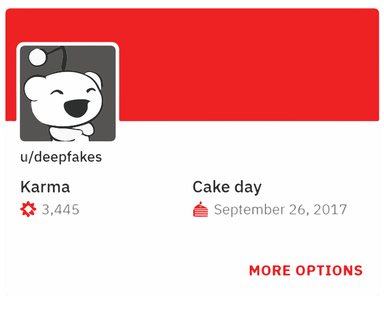
pornography and computer science

The use of female models in computer programming dates back to the field's inception. Some of the first computers were known to have printed pinup girls from punchcard programs. Lenna (pictured right with deepfacelab debug overlay) is a prime example of the trend as her spread in playboy became a staple in the computer vision and face detection fields.



Research and Creating a Fake

Studying deepfakes presented a number of research challenges somewhat unique to the area of inquiry. Made popular by the creation of face-swapped pornography, the development and discourse related to the field was hosted anonymously online. Because of the unethical nature of the videos deepfakes produced, communities were silenced and secluded to less trafficked online spheres. Lacking any specific geography or centralized authority meant that my research methods would have to adapt. Furthermore, the emerging technology lacked the scholarly commentary that often accompanies a field of significant societal affect. Limited academic resources were exclusively technical in nature, outlining how to produce and combat deepfakes. The news media reported on the emergence starting in December of 2017, but failed to debate the consiquences of deepfakes in depth. In the end, creating a deepfake seemed like a logical progression, working with the technology allowed me to get first hand insight, learn from community members, and become my own primary source.



u/deepfakes

Although the first deepfakes were technically produced by Nvidia researchers, Reddit user u/deepfakes popularised the technology with the creation of the first NSFW deepfake swapping the face of a porn actress with that of Gal Gadot. The face-swap format of fakery through deep learning.

Deepfakery and Post-Truth Societies

The fight against deepfakes has been waged thus far by shutting down community forums and developing tools to identify fake videos through the use of neural networks. While much of the media coverage takes comfort in the alleged success of these precautions, this thesis argues their effect is temporary and that they fail to address the underlying issues that deepfakes pose.

- It is not enough to only identify a deepfake.
- The repercussions of a deepfake do not hinge on authenticity.
- What recourse do we have to protect society from advanced and wide spread disinformation?

Reality Fuzzing: Akin to fuzzing in the conventional sense, where invalid or unexpected inputs are used to debug code, deepfakes may produce a similar affect wherein the information one is exposed to becomes so saturated with fakery that it becomes difficult (and perhaps impossible) for the average media consumer to identify truth.

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®