A Framework Analysis of Deepfakes: Using SWOT and FMEA to Calculate the Risk Posed by Deepfakes

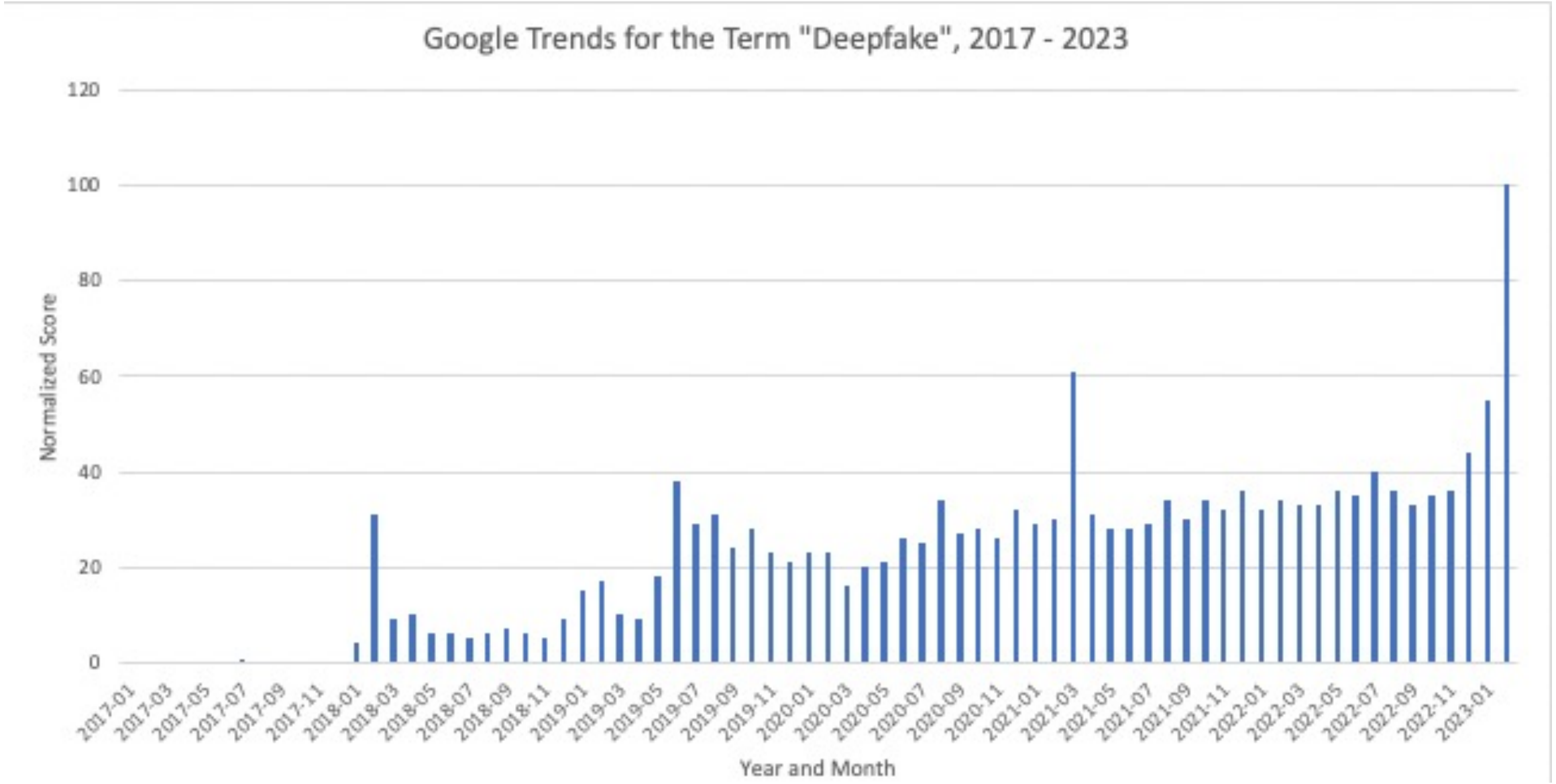
Kastur Koul
kastur.koul@vanderbilt.edu

March 27, 2023

Deepfakes Overview

- Definition
  - Piece of digital media in which the face or movement of one actor is replicated on another actor using deep learning algorithms
  - Combination of “deep learning” and “fake”
  - Name comes from Reddit user “deepfakes”
    - Posted face swapped deepfake of actress Gal Gadot in a pornographic video
    - Reported by *Motherboard* in December 2017
    - Starting point of public concern
- Types of deepfakes
  - Head puppetry: Source head and upper shoulder movement replicated on target
  - Face Swapping: Face of the source swapped with face of target
  - Lip Synching: Manipulation of lip region of the target
- Interest in deepfakes started and has grown since 2017

Google Trends



Deepfake Creation & Detection

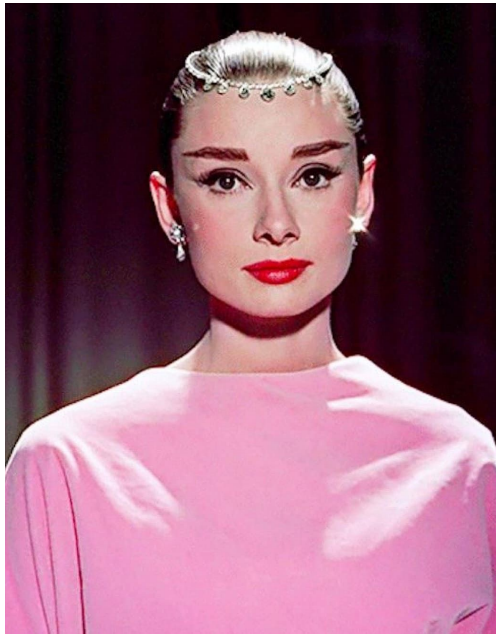
Creation	Detection

Generative Adversarial Networks (GANs)
• Generator
  • Creates fake images
• Discriminator
  • Learns to tell difference between real and fake images
	Convolutional Neural Networks (CNNs)
	• Extracts features from source images

Encoder/Decoder Networks
• Encoder
  • Learns facial features from sources
• Decoder
  • Reconstructs faces for target
	Long Short-Term Memory Networks (LSTMs)
	• Frame-by-frame sequence analysis (for videos)

	Human eye • Spotting visual artifacts

I am a movie star



Previous Risk Analysis Work

- Ali *et al.*
  - Potential malicious uses of deepfakes, especially in politics
  - No risk analysis frameworks
  - “primarily just a technical tool with more positive uses than negatives”
    - No further elaboration on positive uses
- Gamage *et al.*
  - Research questions provide framework
  - Reddit community conversations about deepfakes and societal implications
  - “double edged sword”
- Pew Research Center
  - Survey on views about altered images and video
  - Generally negative
  - 77% adults in the U.S. say “steps should be taken to restrict altered videos and images that are intended to mislead”

SWOT Analysis

- SWOT Analysis Method
  - Analysis method used to define the strengths, weaknesses, opportunities, and threats of an organization or system

- **S**trengths
  - Currently exhibited positive attributes the technology displays in its use by an end user
- **W**eaknesses
  - Currently exhibited negative attributes of the technology that might hinder the experience for end users
- **O**pportunities
  - Potential positive uses of deepfakes in various fields
- **T**hreats
  - Potential negative uses of deepfakes in various fields

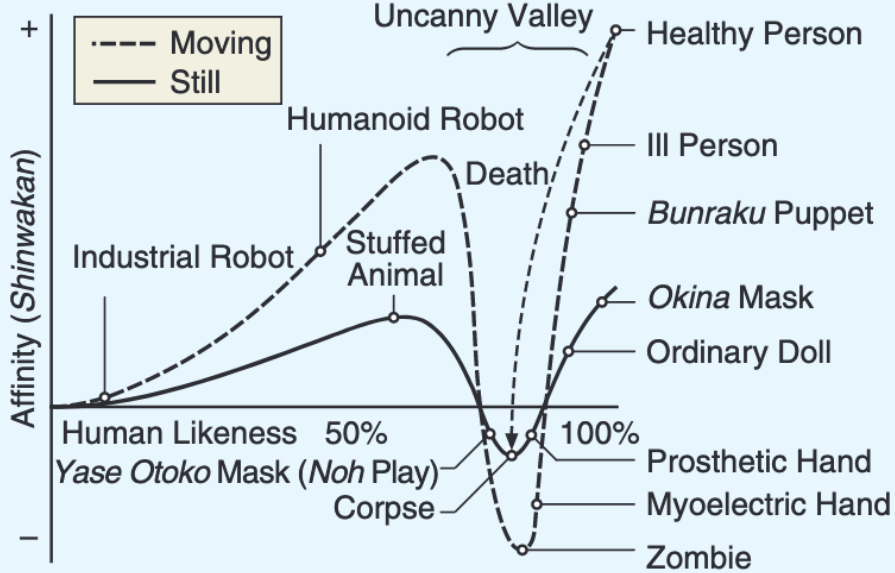
Strengths (S)	Weaknesses (W)	Opportunities (O)	Threats (T)
• Saves **time and money** for small companies as they can be cheaply made. • Uses **deep learning** techniques. • Innovation in digital **realism**.	• Most deepfakes have **detectable differences** that make them too easy to spot. • **Uncanny Valley** effect can spoil the experience of using a deepfake. • Availability and quality of **training data** might not be enough to make a deepfake of any random person.	• Use in the **Fashion** industry to introduce accessibility. • Use in **Entertainment** to recreate popular actors faces and accurate lip dubbing. • Use in **Education** to create interactive learning tools. • Uses **AR/VR** to create memorable experiences. • Uses in **Video Games** to assist in the development process. • Create opportunities for **Trustworthy AI** by being transparent, accessible, and diverse. • Embodied chatbots for **telehealth and teletherapy**	• Use in **social engineering** through impersonation. • Ethical concerns about the creation of deepfakes using someone's image without their **consent**. • Difficult to regulate and use of deepfakes as evidence in cases of **law**. • **Online harassment** (such as blackmail and impersonation) resulting in a lack of privacy and security. • Use of deepfakes to create **nonconsensual pornography**. • Spread of **misinformation** to bolster personal agenda. • Use of deepfakes in **politics**.
Total: 3	**Total: 3**	**Total: 7**	**Total: 7**

Strengths

- Time and Money
  - Machine learning algorithms are faster and cheaper than humans at certain tasks
  - Resources saved can be spent on other areas
- Deep Learning
  - Research reports successful uses of deepfake models
  - Mentioned above, can be faster and cheaper than human at the same task
- Realism
  - Speaks to the quality of the deepfake
  - Better experience for users through realistic deepfakes

Weaknesses

- Detectable differences
  - Visual digital artifacts
    - Irregular eye movement, out of sync with audio, mismatched features on a face
- Uncanny Valley
  - Increasing affinity people have for human-looking robots until that affinity comes to a valley
  - Detectable differences make deepfakes fall into Uncanny Valley
- Training Data
  - Algorithms require a lot of good quality data
    - Quality of data directly leads to quality of output
  - Consent of the people whose image is used to train the model



FMEA Risk Analysis

- Failure Means and Effects Analysis
  - Used to identify and address potential problems and the effects on the system
  - Provides an insight into what the behavior of a system will be given a single point of failure
- Probability
  - How likely is the use of the deepfake to occur
- Detection
  - How likely is the deepfake to be detected
- Severity
  - How damaging or beneficial the emotional, financial, or societal effects of the use if the deepfake is for an individual
    - S_D – The severity of detected uses of deepfakes
    - S_U – The severity of undetected uses of deepfakes
- Risk Priority Number (RPN)
  - Expected risk of the use of the deepfake

$$RPN = P * ((D * S_D) + ((1 - D) * S_U))$$

FMEA Risk Analysis

Probability Scale

Probability	Description
0.2	Not likely; Most likely the use will not occur in the future.
0.4	A little likely; The use might happen, but the chance is unlikely.
0.6	Likely; The use might happen.
0.8	Very likely; High chance the use will happen.
1	Guaranteed; This use will happen in the future.

Detection Scale

Detection	Description
0	Impossible to detect; The deepfake will not be caught.
0.2	Difficult to detect; The deepfake very likely not get caught.
0.4	A little difficult to detect; The deepfake might not get caught.
0.6	A little easy to detect; the deepfake might be caught.
0.8	Easy to detect; The deepfake will very likely be caught.
1	Guaranteed detection; The deepfake will be caught.

Severity Scale

Severity	Description
-4	Extremely severe; Irreparable and guaranteed negative emotional, financial, or societal damage.
-3	Very severe; Guaranteed negative emotional, financial, or societal damage that is potentially but not easily reparable with time.
-2	Severe; Negative emotional, financial, or societal damage that is reparable with time.
-1	A little severe; Mild negative emotional, financial, or societal damage that is easily reparable.
0	Neutral; Neither positive nor negative consequences.
1	Few benefits; Mild positive emotional, financial, or societal outcomes.
2	Some benefits; Positive emotional, financial, or societal outcomes
3	Many potential benefits; Long lasting positive emotional, financial, or societal outcomes.
4	Guaranteed benefits; Life-changing positive emotional, financial, or societal outcomes

Opportunities - Entertainment

- Accurate dubbing
  - Use of lip dubbing to change actor's mouth movement to match language
  - Pros
    - Make movies/shows accessible for audiences in different countries
    - More entertaining experience for audiences
  - Cons
    - Errors can fall into Uncanny Valley
- Use popular actors faces
  - Current use: Luke Skywalker in *The Mandalorian*
  - Pros
    - Likeness for commercial purpose protected by law
    - Potential to generate more revenue
    - Retain the likeness of popular actors
  - Cons
    - Misuse of actor's likeness without their consent

- Severity (Detected): 3
- Severity (Undetected): -1
- Probability: 0.8
- Detection: 0.8
- RPN: 1.76

Opportunities - Education

- Students interact with historical figures
  - Learn physics with Isaac Newton, or read Shakespeare's plays with the Bard himself
  - Pros
    - Immersive and memorable experience for students
    - New educational tools for teachers
    - No need for personal data collection
  - Cons
    - Potential spread of misinformation might harm educational process
    - Affordability and access to technology needed

- Severity (Detected): 3
- Severity (Undetected): -2
- Probability: 0.6
- Detection: 0.6
- RPN: 0.6

Opportunities - Fashion

- Accessible shopping
  - GAN-made full body models of shoppers
  - Customers unable to try on clothes can do so virtually
  - Pros
    - Saves travel time
    - Potentially saves money (no needless spending)
    - Increased market
    - Does not remove option of going to the store
  - Cons
    - Hacking can lead to loss of personal data

- Severity (Detected): 2
- Severity (Undetected): -2
- Probability: 0.8
- Detection: 0.6
- RPN: 0.32

Opportunities – AR/VR

- Realistic models
  - Quick and easy creation of models for AR/VR purposes
  - Tie to education: additional immersive experience
    - Learn with Shakespeare at the Globe, Newton under the apple tree
    - Current example: Georgia Peanut Commission Education Center
  - Easier communication
    - Bring people from around the world to the same room
  - Pros
    - Faster and cheaper ways of creating models
    - Immersive and memorable experiences for users
  - Cons
    - Misuse of person's image can lead to harm
      - Example: Deepfaked pornography

Figure 1: Frame by frame video of the image animated portrait

- Severity (Detected): 2
- Severity (Undetected): -3
- Probability: 1
- Detection: 0.4
- RPN: -1

Opportunities – Video Game Development

- Realistic models for games
  - Create character models with ease
  - Assist in the process of development
  - Pros
    - Saves time and money for developers
    - Better experience for players
  - Cons
    - Potential to trigger players

- Severity (Detected): 2
- Severity (Undetected): 0
- Probability: 0.8
- Detection: 1
- RPN: 1.6

Opportunities – Telehealth/Teletherapy

- Embodied chatbots
  - Deepfakes chatbots to provide telehealth or teletherapy services for doctors who might be busy
  - Pros
    - Reduce travel and wait times for patients
      - Potential to save lives
    - Reach more patients in less time
    - Help patients process thoughts and emotions
  - Cons
    - Overreliance on the technology
      - Anxiety, depression, suicide
    - Potential to replace the need for human company
      - Social isolation

- Severity (Detected): 3
- Severity (Undetected): -4
- Probability: 0.6
- Detection: 0.4
- RPN: -0.72

Opportunities – Trustworthy AI

- Create trust in AI Technology
  - Design with trustworthy AI principles in mind
    - Accountable
      - Creators take responsibility for the deepfake and its effects
    - Transparent and Explainable
      - Discussion of how deepfakes are created and what data they use
    - Human-centered Values
      - Accessibility – make tools and services easier for audiences to use
      - Diversity – generating diverse outputs by having diverse input data
  - Pros
    - Increase use of the technology
    - Better understanding of the technology by the public and its effect by researchers and developers
  - Cons
    - Undetected deepfakes that cause damage can ruin trust in the technology

- Severity (Detected): 3
- Severity (Undetected): -2
- Probability: 0.4
- Detection: 0.8
- RPN: 0.8

Threats – Misinformation

- Spread of misinformation under guise of reputable sources
  - “Wolf News” from Spamouflage
    - Deepfakes promoting interests of Chinese Communist Party
  - Confirmation bias
    - People not using critical thinking when analyzing material, favoring evidence that bolsters their own views
  - Pros
    - None
  - Cons
    - Damaging to person’s reputation
    - Easy to create and spread
    - Can lead to dangerous actions based on incorrect information

- Severity (Detected): -1
- Severity (Undetected): -4
- Probability: 0.8
- Detection: 0.4
- RPN: -2.24

Threats – Politics

- Deepfaked politicians
  - Some are humorous or satirical
    - Jordan Peele's Barack Obama
      - Not hiding the fact that it is a deepfake
  - Some are harmful
    - President Volodymyr Zelenskyy deepfake
      - Caught and exposed
  - Pros
    - Humorous uses can be entertaining if known
  - Cons
    - Potential to start or end wars unfavorably
    - Viewers have to stay constantly vigilant
    - Politicians can lose credibility

- Severity (Detected): -1
- Severity (Undetected): -4
- Probability: 0.8
- Detection: 0.8
- RPN: -1.28

Threats – Social Engineering

- Puppet fraud
  - Deepfakes made to look like an employee from a credible source (“puppet”)
  - Gather personal information from customers under guise of reputable companies (“fraud”)
  - Pros
    - Companies can build their cybersecurity
  - Cons
    - Loss of customer’s financial security and privacy
    - Damage to company reputation

- Severity (Detected): -1
- Severity (Undetected): -4
- Probability: 0.4
- Detection: 0.6
- RPN: -1.68

Threats – Deepfakes and the Law

- Use of deepfakes as evidence in law
  - Guilty defendant can get away with crime
    - Deepfaked alibi
  - Vengeful plaintiff can get defendant sentences
    - Deepfaked evidence
  - Knowingly submitting falsified evidence is a felony
  - Pros
    - Laws in place that try to mitigate the use of deepfakes
      - Ex.: CA AB 730 (2019), Texas Penal Code 33.07 from (2021)
  - Cons
    - Undetected uses can lead to innocent people going to jail
    - Cast doubt on innocence of defendant

- Severity (Detected): -1
- Severity (Undetected): -4
- Probability: 0.2
- Detection: 0.8
- RPN: -0.32

Threats – Online Harassment

- Blackmail and impersonation
  - Deepfakes can be made of an individual saying or doing something bad
  - Can be used as blackmail to extort money or services
  - Pros
    - None
  - Cons
    - Emotional, financial, and interpersonal damage for target
      - Depression, anxiety, no way to prove deepfake is false
      - Disruption from everyday life

- Severity (Detected): -1
- Severity (Undetected): -4
- Probability: 0.8
- Detection: 0.4
- RPN: -2.24

Threats – Deepfake Pornography

- Use of deepfakes to create pornographic material
  - This was the first use of a “deepfake”
  - Threat comes from the nonconsensual use of someone’s image to create pornography
  - Pros
    - Protection offered by some state laws
      - Ex. CA AB 602 (2019)
  - Cons
    - Nonconsensual use of someone’s image can be emotionally distressing for them
      - Recent case of Twitch streamer Sweet Anita
    - Copies can be spread even if original is taken down
    - Little protection for non-public figures

- Severity (Detected): -2
- Severity (Undetected): -4
- Probability: 1
- Detection: 0.4
- RPN: -3.2

Threats – Consent

- Obtaining consent from the person whose image is used
  - The big question: how often does this happen?
  - How is the deepfake/dataset used
    - Commercial purposes not allowed
    - Fair use – protects satire and parody
    - Celebrities are generally the targets of deepfakes
  - Pros
    - Liability for creators and ease of mind for image provider
  - Cons
    - Reputation damage from misuse of image
      - Can cause further emotional, financial, or interpersonal harm
    - Defamation can lead to lawsuits

Want to see a magic trick? Tom Cruise impersonator Miles Fisher (left) and the deepfake Tom Cruise created by Chris Ume (right). Image: Chris Ume

- Severity (Detected): -1
- Severity (Undetected): -3
- Probability: 0.8
- Detection: 0.2
- RPN: -2.08

Results - Opportunities

Breakdown of RPNs for the Opportunities of Deepfakes

Opportunities	Severity (Detected)	Severity (Undetected)	Probability	Detection	RPN
Fashion	2	-2	0.8	0.6	0.4
Entertainment	3	-1	0.8	0.8	1.76
Education	3	-2	0.6	0.6	0.6
AR/VR	2	-3	1	0.4	-1
Video Games	2	0	0.8	1	1.6
Trustworthy AI	3	-2	0.4	0.8	0.8
Telehealth	3	-4	0.6	0.4	-0.72
Average	2.57	-2.00	0.71	0.66	0.49

Results - Threats

Breakdown of RPNs for the Threats of Deepfakes

Threats	Severity (Detected)	Severity (Undetected)	Probability	Detection	RPN
Social Engineering	-1	-4	0.6	0.4	-1.68
Consent	-1	-3	0.8	0.2	-2.08
Law	-1	-4	0.2	0.8	-0.32
Online Harassment	-1	-4	0.8	0.4	-2.24
Pornography	-2	-4	1	0.4	-3.2
Misinformation	-1	-4	0.8	0.4	-2.24
Politics	-1	-4	0.8	0.8	-1.28
Average	-1.14	-3.86	0.71	0.49	-1.86

Results - Average

Summary of Average Scores

SWOT	Severity (Detected)	Severity (Undetected)	Probability	Detection	RPN
Opportunities	2.57	-2.00	0.71	0.66	0.49
Threats	-1.14	-3.86	0.71	0.49	-1.86
Average	0.72	-2.93	0.71	-0.69	-0.52

Conclusion

- Deepfakes are viewed negatively
  - Popular media describes them with more negativity
- Different ways to create and detect deepfakes
- There are strengths, weaknesses, opportunities, and threats
  - Opportunities capitalize on strengths
  - Threats exploit weaknesses

- Deepfakes do pose a risk
  - Final analysis shows that they pose a nonnegligible risk
  - Should inform caution for future development
  - Does not mean there should be no future development

Questions