Cyber Threat Intelligence Report

Threats from Disruptive Technologies: AI Personas & Simulacra

Cybersecurity Division
Compiled by Sarah Hunt & Arjun Dhaliwal

Date: 28 July 2023

Introduction

Over the past few years, the field of artificial intelligence [AI] has witnessed significant progress, leading to its widespread adoption and application across various industries. AI encompasses several key benefits that have contributed to its popularity and success. These advantages include:

Automation & Efficiency: AI has the ability to automate repetitive and mundane tasks, freeing up human resources to focus on more complex and strategic activities.

Enhanced Decision Making, Accuracy, & Precision: AI systems excel in processing and analyzing large volumes of data swiftly and accurately, enabling data-driven decision making with minimal errors caused by human factors like fatigue, emotions, or distractions.

24/7 Availability & Reliability: AI systems can operate continuously without experiencing fatigue, making them particularly valuable in applications such as customer support where round-the-clock availability is essential.

Big Data Handling: With the exponential growth of data, AI algorithms and techniques enable the extraction of valuable insights and meaningful conclusions from massive datasets.

Personalization & Recommendation: AI algorithms can analyze user behavior, preferences, and historical data, enabling personalized recommendations and tailored experiences.

Innovation, Creativity, & Augmentation Capabilities: AI technologies, such as machine learning and deep learning, can generate novel ideas, designs, and solutions by recognizing and learning from patterns in existing data. Additionally, AI can augment human capabilities by providing intelligent assistance in various tasks.

Risk Mitigation: AI systems can analyze risks and predict potential failures or hazards, facilitating proactive measures to mitigate risks and prevent adverse events.

While AI has experienced remarkable advancements in recent years, it is crucial to consider important factors such as ethics, transparency, cybersecurity, and responsible deployment to ensure its beneficial and fair use in society. This perspective on responsible utilization is an essential aspect considered in this threat report.

What are AI Personas and Simulacra?

AI personas and simulacra are very similar in execution; however, the minor differences between them result in different implicit threats.

An AI persona is a completely fictional AI avatar, often generated as a composite representation of a group (*delve.ai, 2023*). Assistive chatbots on websites are a good example of the potential use of an AI persona, where the persona would be that of a personable expert in the field. Personas can be broad—for example, someone who lives in a certain province—or very specific—for example, a hardware store's website might have a help bot who has the persona of a 43-year-old Albertan single male who works in manufacturing and owns their own business. The specificity depends on the purpose the persona is being created for.

How an AI simulacra differs from a persona is in that it is an AI avatar based off an individual that actually exists in the real world. An example of an AI simulacra would be the 2018 video of Barack Obama warning of the dangers of AI generated deepfakes. It was not actually Obama making this statement, but a deepfake created to raise awareness using the former president's voice and likeness (*Silverman, 2018*).

AI Persona
An avatar created by artificial intelligence to mimic human-like behaviours and characteristics. Usually based on a composite representation of a group or demographic.

With advancements in natural language processing, machine learning, and deep neural networks, personas and simulacra are becoming increasingly sophisticated, enabling them to better generate realistic responses and adapt to context. While this means that AI personas and simulacra are showing greater promise for successful application in various sectors—including customer service, entertainment, and technology—they also introduce significant risks and threats, particularly in the realms of cybersecurity, privacy, and social engineering.

AI Simulacra
An avatar created by artificial intelligence to mimic a specific individual's behaviours and characteristics.

Why are AI Personas and Simulacra Such a Threat?

The nature of cyberspace is evolving rapidly. Not so long ago the security perimeter used to be physical, with literal buildings and security personnel. Then with the prevalence of the internet, the perimeter moved to be digital, with things such as firewalls protecting it. Now with the movement towards hybrid and remote work, the security perimeter has again moved. The new perimeter includes identity. Authentication and authorization of identity to access certain digital systems, devices, and services is how organizations are protecting their information. This makes technologies—such as AI personas or simulacra—that can create or mimic identities a threat to organizational security. Specific ways these technologies can be a threat are discussed later in this paper; however, to better understand the threat a brief overview of how these technologies work will be examined first.

Legitimate Uses of AI Personas & Simulacra

Ⓟ = AI persona example Ⓢ = AI simulacra example

Information Sharing

- Ⓟ Creating more *engaging lessons* to keep student's attention
- Ⓢ Allowing *trusted news anchor* to provide information in breaking news situations

Entertainment

- Ⓢ *Val Kilmer's voice*, *Top Gun: Maverick*
- Ⓢ *Resurrecting Paul Walker*, *Fast & Furious* franchise
- Ⓟ *Zo*, Microsoft AI Teen Chatbot

Assistive Chat Bots

- Ⓟ *IBM Watson Assistant*
- Ⓟ *Alexa/ Siri*

Advertising & Marketing

- Ⓟ *Netflix recommendations*
- Ⓢ *Gun violence advertisement* using high school student killed in a school shooting

Identity Protection

- Ⓢ *Protecting Russian LGBTQ2S+ refugees' likenesses* in the documentary *Welcome to Chechnya*

Malicious Uses of AI Personas & Simulacra

Ⓟ = AI persona example Ⓢ = AI simulacra example

Manipulative Advertising or Campaigning

- Ⓢ *Socialistische Partij Anders, a Belgian political party, created a video of Donald Trump appearing to offer advice to Belgium to manufacture attention to get people to side with their party*

Generation of Mis,Dis, and Mal-information

- Ⓟ *Fictional news anchors* promoting untrue view of China's role in an international summit meeting

Digital Impersonation

- Ⓢ Ukrainian President Volodymyr Zelenskyy "*surrendering*" near the start of the Russian invasion

Fraud

- Ⓢ *By-passing bank biometric authentication*

Non-consenting Depictions of Real Individuals

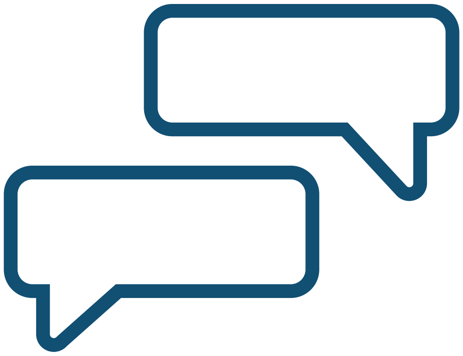
- Ⓢ *Face swaps* of celebrities onto adult performers bodies

Disclaimer

How AI Personas & Simulacra Work

Text-Based

Chatbots have recently been storming the internet, showcasing a wide array of ways they can be applied. Depending on one's objective, chatbots can be employed to facilitate diverse tasks, including customer service, entertainment, or information retrieval. As most businesses interested in chatbots do not usually have the expertise to code it themselves, it is often created on third-party chatbot creation websites, such as Facebook Messenger. After a chatbot is created it also needs to be trained to understand situations it may face, such as answering frequently asked questions. This usually consists of feeding the chatbot sample interactions to train the bot to understand human language. For in-depth human conversations a process called sequence-to-sequence modeling is used. For example, this is the same type of modeling used by Google Translate. This allows for the generation of a large number of conversational logs to be formed and used to train the bot with a variety of datasets. Iterative testing and fine-tuning are necessary to improve the chatbot's performance and address any constraints or errors. Monitoring regularly is essential to ensure the chatbot remains up to date and evolves with the users needs. Despite regular updates and diligent monitoring, chatbots still encounter numerous challenges, particularly when aiming to achieve a high level of sophistication. These challenges include the complexities of accurately discerning and understanding customers' emotions and intentions, an inability to comprehend uncommon use cases, and the significant costs associated with creating and maintaining chatbots and their associated systems (*Discover.Bot, 2019*).



Voice-Based

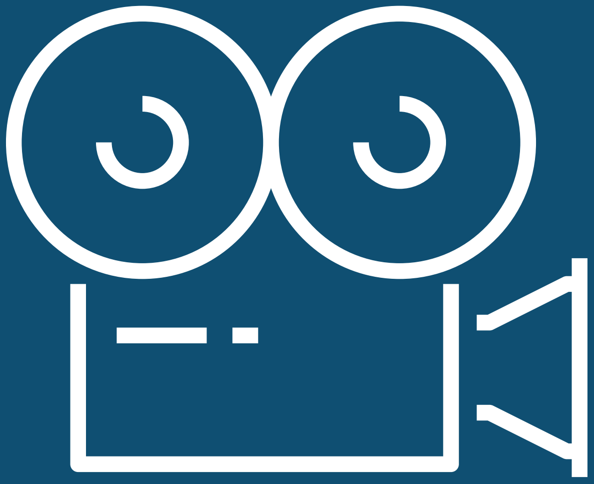
Another intriguing technological advancement lies in the capability to synthesize speech that closely mimics the voice of a specific individual. By utilizing audio clips and text inputs as reference, an AI model can generate highly realistic speech patterns, effectively replicating the vocal characteristics of any person. A specific example of this would be the neural codec language model, VALL-E created by Microsoft. With only a three second audio clip, VALL-E has the ability to synthesize high quality personalized speech while maintaining the speaker's emotional expression and the auditory context associated to the original audio recording (*Microsoft, n.d.*). Essentially, VALL-E analyzes the vocal characteristics of how a person sounds and breaks it down into individual elements referred to as tokens. Then using up to 60,000 hours of English language audio from close to 7,000 different speakers on an audio library called LibriLight, VALL-E speech-synthesis abilities are trained to detect similarities between the tokens in the library-based audio and the original clip (*Edwards, 2023*). Using this data, VALL-E constructs how that voice would sound when speaking different phrases outside of the original three second sample.

Image-Based

One of the most interesting recent technological innovations is the means to generate images based off of text alone. Around a year ago, a new AI program named DALL-E was created with an ability to create highly realistic and creative images from text depictions. The way DALL-E does this is through the interactions of multiple AI models. In a combined effort DALL-E is able to use hundreds of millions of images and captions filed in its memory to accurately depict a relation between a language snippet and a visual concept. Based on this relationship, an image is accurately encoded and generated to effectively represent the text written (*O'Connor, 2022*). DALL-E offers a wide range of applications and has the potential to significantly enhance various business operations and personal hobbies for a multitude of people. Some of these include the ability to enhance 3D design, such as architects beginning to incorporate AI generated designs into their products and plans. Similarly, the design and planning of different physical products is exponentially easier and accelerates innovation for retail and interior designers around the world. With only a few words, any concept or idea can be displayed for you in a matter of seconds. Finally, the ability to quickly generate custom, unique visuals, and images based on your own distinctive ideas streamlines the creation of both brand identities and personas (*Kane, 2023*). In a remarkably short period of time, DALL-E has facilitated diverse needs of individuals and provided invaluable assistance with its expansive range of applications.

Video-Based

Deepfakes are an advanced technology with the capabilities to project almost any individual onto a photograph or video in which they did not actually appear in. Although they have been around for the past 10-20 years the quality and ability of deepfakes have progressed rapidly, reaching a stage where distinguishing between reality and fabrication can become challenging. To generate deepfakes takes technical skill, specialized software, and a substantial amount of computing power. There are two main ways deepfakes are made. The first and easiest is to use AI to superimpose faces of a target subject onto a video of another subject, this is also commonly referred to as face swapping. This can be done with any videos found on the internet or by home videos created by the individual. The second major way deepfakes are made also uses AI via a process called deep learning. An AI is first fed a multitude of videos, images, and audio clips of the target individual they want to impersonate. Once the program has this information it creates millions of models to capture unique facial features, expressions and movements (*Somers, 2020*). After this process the AI rarely generates a perfect product; however, through a complex process—using facial landmark detection, texture blending, and image alignment—a creator is able to perfect the deepfake. With each continued iteration of learning and training the AI can steadily improve the quality and ability of the deepfake to replicate the target individual.



Threats Posed by AI Personas & Simulacra

Improvements to the technology responsible for the generation of AI personas and simulacra has brought with it increased threat exposure. The exact threats are wide ranging; however, they can be broken down into the following broad categories.

Manipulation of Information

One of the single biggest threats associated with personas and simulacra starting to become indistinguishable from reality is that it allows for information to be manipulated in such a way that it cannot be easily distinguished from reality. A trusted source for news, such as the Prime Minister or a news anchor, may have their likeness usurped by an AI simulacra and be used to spread disinformation. In a relatively harmless example, in 2018 Barack Obama appeared to warn of the dangers of deepfakes (*Silverman, 2018*). A more harmful version of this was seen shortly after Russian invaded Ukraine in 2022, when a video appeared online apparently showing the Ukrainian president Volodymyr Zelenskyy surrendering in the fight against the Russians (*Allyn, 2023*). This was a deepfake and was quickly debunked; however, it served its purpose to sow confusion and panic. AI persona's can also be used to this end, with campaigns where entirely false but believable news anchors spread disinformation (*Satariano & Mozur, 2023*).

False narratives can be created by AI personas flooding comment sections of social media sites with fake information or an extreme opinion, creating an illusion of consensus or discord that does not exist in reality. To quote Joseph Gobbels, a Nazi propagandist and master of manipulation and disinformation, "[i]f you repeat a lie often enough, people will believe it." An example of this playing out in real life can be seen in the Federal Communication Commission's [FCC] comment section during the six-month period the public was able to voice their opinion on the proposal to repeal net neutrality protections in 2017. The comments were used to inform the FCC's final decision and saw approximately 17 million out of 22 million people support the repeal, which the FCC eventually did (*Confessore, 2018*). Unfortunately, an investigation after the fact revealed that only 800,000 comments—less that 4% of the total—were from real people and of those approximately 99% were against the repeal. This means that a few malicious actors utilizing a very basic AI persona resulted in the skewing of democratic expression as it was determined that 17 million of the comments originated from just 20 different bot campaigns (*Singel, 2018*).

In terms of generative AI (e.g., ChatGPT), hallucinations are responses provided by the AI that appear to be real and are given with the confidence that they are real; however, they are not based on sufficient data. In short, hallucinations are when the AI wants to provide an answer but does not have the information required to provide a correct answer so it makes up some or all of the response (*Weise & Metz, 2023*). In terms of AI personas and simulacra, hallucinations are most relevant to text-based generations that rely on machine learning to create their identities.

While this misinformation is problematic and can pose a threat to the propagation of accurate information, there is also another underlying issue that poses a threat. As these programs rely on machine learning, this misinformation gets cycled back into the AI's library, meaning that that data it pulls its information from is now poisoned with misinformation. If the same question is asked of the AI again it will likely respond with the same answer, creating a cycle that can see misinformation be considered the truth unless corrected (*Simonite, 2018*).

Social Engineering

Social engineering is when a malicious actor manipulates an individual into revealing private or confidential information to them, often to be used for malevolent purposes (*Malwarebytes, n.d.*). Some of the most common types of cyber social engineering include phishing—email-based social engineering—and vishing—phone-based social engineering. AI personas and simulacra have great potential to improve social engineering methods since they can allow for greater depth to the deception. Previously, if you received an email from your boss requesting you purchase gift cards for a conference, you would have been able to phone or video call your boss and find out if this was actually a phishing attempt. With deepfake voice and video, cybercriminals can now create a simulacrum of your boss, so when you call the number at the bottom of the email someone who sounds like your boss might tell you the email is legitimate.

A very similar situation to this played out in 2019 when a UK CEO was convinced to transfer €220,000 ($325,700 CAD) to who he thought was the CEO of his parent company. The reality of the situation was that this was a vishing call where the cybercriminals used an audio deepfake of the CEO of the parent company to trick the UK CEO into sending the money (*TrendMicro, 2019*).

The Grandparent Scam is a type of social engineering that has existed since the 1200s. Back in the 13th century, a family would receive a letter that a family member had been arrested and sent to a Spanish prison, but that the letter sender could get the person out of jail if the family sent them money. The reality, of course, was that no one was in jail and the letter was meant to panic the family into sending the money without thinking (*Burke, 2015*). The only difference between the Spanish Prisoner scam and the Grandparent Scam is that they no longer use letters, now preferring phone calls.

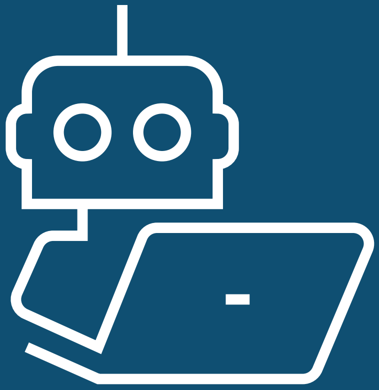
In the past year or two, AI simulacra have started playing an important role in the success of the Grandparent Scam. Before scammers would phone the family member—often a grandparent giving the scam its name—and just tell them their grandchild was in prison. Now with audio deepfakes the cybercriminals can mimic the grandchild they are claiming is in prison asking for help, which lends credence to the deception and makes the family member more likely to pay (*Puig, 2023*; *Cooke, 2023*).

It is not just AI simulacra that can be helpful with social engineering. A well-made AI persona adds credence to a help desk scam or other type of social engineering, as there can now be a voice, face, and personality behind the falsehood.

Fraud

A small peak behind the curtain, but the impetus behind this threat intelligence report was an article written by Joanna Stern for the Wall Street Journal. In the article, Stern details how she created an AI simulacra of herself using generative AI text, deepfake video, and deepfake audio. She then set her AI "clone" to complete tasks that she would regularly do during her day. One of these tasks was to call her credit card company, which uses biometric voice authentication to allow customers to access services. The audio deepfake passed the check and connected the simulacra with customer service, giving them access to make changes to Stern's account (*Stern, 2023*; *Wall Street Journal, 2023*).

While in this case the person using the AI simulacra was the legitimate user, the potential for misuse of this technology is great. A malicious user could have used the same technology to defraud credit card and banking customers whose organization's use similar authentication methods by accessing their accounts and changing information to benefit the cybercriminal.



Espionage

In 2019, there was a LinkedIn account accredited to Katie Jones. Jones purported to be in the employ of the Center for Strategic and International Studies with over 50 connections to high-profile Washington-based politicians. The catch? Katie Jones does not exist, she is—at least in part—an AI persona. The purpose of the profile is suspected to be state-sponsored espionage. Validity was granted to the account due to the profile photograph—which is heavily suspected to be generated using a visual deepfake tool. This is a fairly common tactic used by foreign espionage operators, as it allows for research on individuals and networks, giving them the information they need to spear phish or otherwise defraud protected information (*Vaas, 2019*). The inclusion of the AI generated image adds a level of believability to the profile, since if a person does a reverse image search nothing will come back as it is a unique image (*Hao, 2019*).

As the tools to develop more elaborate and harder to spot AI personas and simulacra improve, the potential for them to be used for spying increases. Whereas, Jones' photograph helped with initial infiltration, deepfake videos and audio have been proven to work to trick individuals into sharing information or resources, which could include protected information. While not espionage, the audio deepfake discussed in the *social engineering section* is a good example of how this could play out. The victim in that case was convinced that the voice on the other end of the line was the parent company's CEO even though it was an audio deepfake (*TrendMicro, 2019*). While it was money that was transferred that time, it is easy to see how a similar phone or video call from an AI simulacra could convince someone to share corporate or government secrets.

Defamation & Extortion

The idea of someone else being able to make you say or do something against your beliefs is a terrifying concept for most people. However, it is now a possible reality with AI simulacra. For example, a voice deepfake service was used in early 2023 by a malicious community online to develop and propagate audio clips of celebrities espousing often offensive opinions contrary to their actual beliefs. One of these clips has a simulacra of Emma Watson—a United Nation's spokesperson—reading Mein Kampf and spewing racist views (*Cox, 2023*). Clips like this can work towards discrediting an individual, by showing them to be a hypocrite or offensive. With current deepfake voice requiring as little as a five second clip of a person's voice (*Christ, 2020*) it is not just public figures that can be discredited this way. Accessibility is making this a threat vector that any person who wishes to harm the reputation of another can use.

Another very common use of generative AI simulacra has been shown to be the creation of adult-content. According to SensityAI, 90-95% of all deepfake videos tracked online show nonconsenting depictions of people—primarily women—in adult situations (*Sensity Team, 2021*). In these instances, benign images of the person—typically from public social media sites—are deepfaked onto the bodies of adult film performers.

Victims of this type of cyber attack often report feeling humiliated, powerless, and frightened but also worried about the impact such videos being online might have on their family or vocational life should the simulacra of themselves in compromising positions be discovered (*Hao, 2021*). This results in them being placed in a prime position to be extorted. In June 2023, the Federal Bureau of Investigation [FBI] released a public warning about generative AI programs being used to create explicit photos and videos of individuals. These AI simulacra are then being used to harass the victim into paying money or providing real explicit videos or photographs, or else the cybercriminal will send the video or images to the person's family or friends (*FBI, 2023*).

Privacy & Information Breaches

AI personas and simulacra pose a significant threat in privacy and information breaches, as they have the potential to deceive and manipulate individuals on a massive scale. Malicious threat actors can utilize AI technology to create highly convincing videos or audio recordings, impersonating someone else by seamlessly replacing their face or voice, which could lead to a range of privacy infringements and breaches of sensitive information. For instance, an AI persona poised to impersonate a position of power could be employed to give credence to an information request, similar to how the Katie Jones persona was able to gain access to powerful individual's LinkedIn profiles. AI simulacra have the potential to be even more dangerous, as cybercriminals could impersonate the face and/or voice of a person who has access to sensitive material, potentially allowing them to unauthorized access to the protected information.

The nature of generative AI used in chat bots can also have impacts on privacy and information, as these programs tend to use machine learning. Machine learning relies on data being input to learn and keeps learning and drawing connections from the data (*Brown, 2021*). This means that input data becomes part of the program, which can cause privacy and information breaches if protected information or personal information is input to create an AI persona. In the first half of 2023, multiple companies including Amazon had proprietary information showing up in generative AI prompts for non-related persons from employees using these services, and it has become a concern about how data is removed from machine learning programs if it should not be there (*Al-Sibai, 2023*). The issues with the inability or difficulty in removal has led to concerns should protected or identity information enter machine learning data.

Undermining Trust

AI personas and simulacra have emerged as potent technological tools capable of undermining trust in all forms of information. By blending reality and unreality into a single persona or simulacra, these AI generated individuals can deceive people into believing something happened, was said, or is widely believed that is not true. This issue extends beyond the spread of mis and disinformation as it poses a threat to the nature of trust that society is built on. Seeing used to be believing but now, as AI iterates, improves, and becomes capable of generating more sophisticated and difficult to detect personas and simulacra, that is not the case. The mere existence of such technology casts doubts on the authenticity of any content or information presented, leaving individuals questioning the legitimacy of everything they encounter, undermining the very foundation of trust in information as it becomes difficult or incommodious to distinguish truth from fiction.

This undermining of trust also has the ability to delegitimize genuine discourse, videos, or audio, leaving the potential for real information to be considered false as individuals can deflect the truth—for example, if someone was caught on video saying something disparaging they may be able to avoid consequences by claiming it was a faked version created by AI (*Benton, 2021*). Even just the potential for information to be falsified can create reasonable doubt in the minds of people. Questions are being raised by lawyers currently about the potential for both audio and video deepfakes to be submitted as false evidence during trials. As the technology improves and becomes more accessible, they argue, it will become harder for the prosecution to prove whether evidence is real or fabricated, and the defence could raise reasonable doubt by suggesting the possibility of an AI simulacra (*Surovell Isaacs & Levy PLC, 2020*). This has the potential to undermine legal systems unless a clear way to distinguish a simulacra and reality are defined (*Finger, 2022*).

Threat Actor Profiles

Nation-States
State-sponsored actors with the resources and expertise to create and spread digital misinformation or disinformation on a large scale.

Political Organizations
Political groups or campaigns looking to manipulate public opinion in their favour or discredit their opponents.

Cybercriminals & Organizations
Individuals or groups looking to profit from manipulating or scamming individuals.

Hacktivists
Individuals or groups motivated by political or social causes to use digital misinformation and manipulation as a means of achieving their goals.

Disgruntled Individuals
Individuals or groups with a personal vendetta against an individual, event, nation state, organization, etc. who may create and spread disinformation to create chaos, manipulate the public's perception, or otherwise cause disruptions.

Trolls
Individuals, sometimes collected into loose groups, who intentionally create inflammatory, crude, or disturbing content online with the goal to create a disruptive response or to detract from the intended topic of discussion.

Mitigations

The unfortunate reality is that people are not very good at spotting AI personas or simulacra, and the ability to detect them is becoming more difficult as the technology improves. An example to showcase how good people are at detecting AI personas and simulacra can be seen in a 2019 experiment run by Max Weiss. In the experiment, Weiss flooded an American government forum on healthcare reforms with comments generated by a chat bot trained on a rudimentary AI persona based on the type of people who regularly comment on such forums. Even using older technology, the experiment was a success and all 1,001 AI generated comments were accepted by the system. After the initial submission, Weiss polled a random selection of the public to see if they could determine which comments were written by a bot and which were written by people. The results of his polling showed that humans were only able to accurately determine which was a real or bot generated comment 49.63% of the time, indicating they were no better than if they had been guessing (*Weiss, 2019*).

Using an AI persona in such a way is a fairly basic and easy example of how this technology can be used and highlights the need for mitigations to better protect against the threats posed by these disruptive technologies. Some of the possible mitigations include:

Governance & Oversight

Governance and oversight should play an important role in mitigating the threats implicit with AI personas and simulacra. The organization's responsible for creating AI technologies primary motivation for what they do is financial. With the security, safety, and privacy concerns created by the malicious use of AI personas and simulacra, it is important that governance and oversight includes considerations to protect people.

In addition to oversight being required for the general use of AI, AI simulacra specifically present another area where oversight will need to evolve: privacy. Due to the inherently false nature of a simulacra, current privacy legislation may not adequately address the challenges created with the creation and dissemination of AI simulacra (Tseng, 2018).

Blockchain

One of the biggest issues facing cybercrime in general, including the malicious use of AI personas and simulacra, is that it can be exceptionally difficult to determine who has perpetrated the crime. Location and access are no longer limiting features as they would be in physical crimes and the rapidly evolving nature of the digital space has meant that keeping on top of the technology needed to detect and capture digital evidence has been a Sisyphean task (*Cheikosman, Hewett, & Gabriel, 2021*). The lack of nonrepudiation—or the ability to prove the source, legitimacy, and veracity of data—in the creation of AI personas and simulacra are one of the biggest limiting factors in being able to prosecute against them, as several of the crimes they are used for (e.g., espionage, fraud) have formal legislation already.

A proposed solution to this issue is using blockchain systems to create a record for digital content similar to that of a provenance document for physical artwork (*Hasan & Salah, 2019*). While the blockchain is often considered to be synonymous with cryptocurrency, the reality is that it is just a type of technology used for record keeping, not unlike a spreadsheet. Where it differs from a spreadsheet is that it is immutable or unable to be changed which makes it an ideal way to accurately track transactions involving any type of assets, including video, audio, and text-based assets (*IBM, 2023*).

There are three elements often associated with blockchain technology that make it an ideal tool for ensuring an audit trail for digital content, including AI personas and simulacra (Cheikosman, Hewett, & Gabriel, 2021):

Hashes

Hashing on the blockchain involves the creation of a unique digital "fingerprint" for an asset. It is represented by an alphanumeric string whose length is set depending on the specific hashing algorithm used. The element that makes hashes useful is that they change if the content they are generated from changes. This makes it similar to a tamper-proof seal on the asset. In the blockchain, hashes for one record are referenced in the next record created, linking the assets on a metadata level. This allows for people to see if a video, audio file, or text file has been modified which is useful with AI personas and simulacra as it allows for modifications—such as those used to create deepfake audio and video—to be tracked (*Rhodes, 2022*).

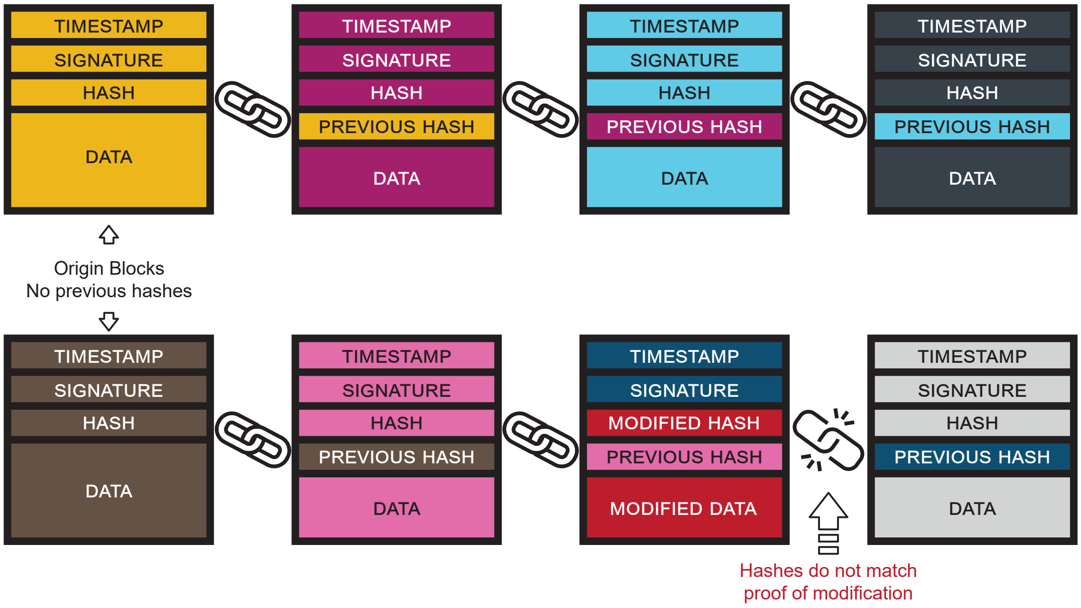
Cryptographic Signatures

Another important tool that when paired with the record keeping prowess of the blockchain can help create nonrepudiation for AI personas and simulacra are cryptographic signatures. A cryptographic signature is a way for the integrity of data to be proven as a creator "signs" the asset with a signature that is unique to them and can be validated as such. This is crucial for nonrepudiation as it allows for the creator of the content to be tracked and verified, allowing for consequences to befall individuals using AI personas and simulacra maliciously while protecting the integrity of the victims (*Pine et al., 2022*).

Timestamps

The final element that the blockchain can bring to help mitigate the malicious use of AI personas and simulacra is timestamping. Combined with blockchain's inability to change records, timestamps create a useful track record to confirm the date and time of the creation or modification of an asset and can show who accessed the asset at a particular time. This creates verifiable evidence of when a persona or simulacra was created, allowing for victims to present proof it could not be them, as well as showing a timeline for who might have modified a video, audio file, etc. to create the persona or simulacra (*Cheikosman, Hewett, & Gabriel, 2021*). An example of timestamps being used in this manner can be seen with Truepic, an organization focused on bringing the same level of transparency to digital assets as physical assets. They have utilized a combination of notarized metadata and timestamps documented on the blockchain to establish an auditable chain of custody for digital assets, including video and audio (*Truepic, 2023a*). Recently, Truepic made news by creating an authenticated deepfake video. This video is the first of its kind in the world and includes verified data from the moment of creation to publication to allow for a transparent and verifiable chain of custody (*Truepic, 2023b*).

How the Blockchain Works



Zero-Trust Maturity Model

When it comes to AI personas and simulacra, the mitigation of associated risks becomes crucial. Many cybersecurity experts feel that to ensure top level security and safety of networks and platforms, a zero-trust approach must be enforced. The zero-trust model assumes that no user or device should automatically be trusted. By verifying and continuously validating every user, device, and system, all access requests are authenticated and authorized to enable access and guarantee everything is secure. Zero-trust succeeds on the mantra to always verify and never trust.

When using a zero-trust maturity model to minimize risk, there are typically various types of capabilities that define the level of security maturity. The higher the level of maturity, the more sophisticated and thorough the security measures are.

Some of the main types of capabilities that define a zero-trust model include:

Identity and Access Management [IAM]:
By implementing multi-factor authentication to verify and confirm the identity of legitimate persons, assurance can be gained to the individual's identity.

Least Privilege Access:
By applying the concept of least privilege, it grants individuals (and legitimate AIs) only the necessary authorization and access rights to complete their tasks. By avoiding providing excessive access, there is a decrease in the potential for misuse and compromise.

Continuous Monitoring:
By continuously surveilling and analyzing the behaviours and activities of users anything that is unusual or unauthorized, such as might be expected by a breach from an AI persona or simulacra, can be noticed and handled quickly.

Incident Response:
With clear roles and regular drilling exercises it will be easier to identify and contain threats, making it important to have an effective incident response plan to handle any security incidents—including those involving AI personas and simulacra—established. Also, after incidents have transpired, remediations should be applied as soon as practicable to prevent further occurrences.

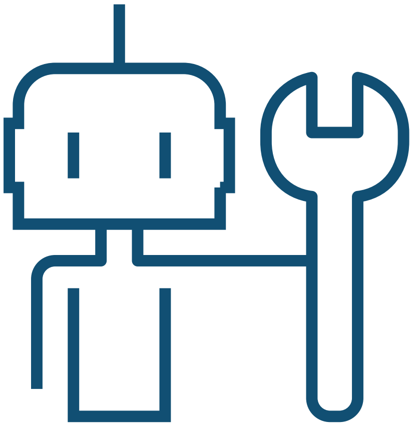
Source: (*Thomas, 2023*)

An area of study which could help with enforcing the zero-trust model is investigative digital ethnography. Digital ethnography involves studying and analyzing online communities, behaviours, and interactions. It typically relies on collecting data from various digital sources, such as social media platforms, online forums, and websites (*Twohig, 2022*). By analyzing digital ethnographic data, organizations can better understand how users are interacting with the network, pinpoint patterns that indicate any suspicious activities or unauthorized access attempts, and improve overall security measures. This information can be used to strengthen access controls, enhance authentication mechanisms, and proactively respond to security incidents within the zero-trust framework (*Friedberg, 2020*).

AI & Machine Learning-Based Tools

Strides have been made in many directions to utilize the power of artificial intelligence and machine learning [ML] to fight the threats associated with AI personas and simulacra; in short, fighting fire with fire. Both AI and ML excel at pattern recognition so the theory behind using them to detect AI personas and simulacra is to use them to detect patterns that either exist or should exist but do not in real versus generated content. These patterns can then be used to develop advanced detection algorithms specifically designed to identify AI-generated synthetic media. The algorithms can analyze various aspects of the content—such as visual or auditory artifacts, inconsistencies, and anomalies—that may indicate manipulation. By using machine learning to continuously train and update the algorithms using large datasets of both genuine and manipulated media, their accuracy can be improved over time.

The specifics of what patterns are looked for depend on the tool being used and what form of AI persona or simulacra it is being used against. With its prevalence in media in recent years a lot of attention has been focused on developing AI assisted digital forensics tools to detect deepfake videos and images. Part of the reason behind this may be because there are a lot of visual tells that indicate the potential of AI videos or image compared to other types of synthetic media. Some of these anomalies are things that the human eye can pick up on—such as unusual blinking, blurred ears, overly smooth skin textures, unusual hands, or uncanny mouth movements—which are elements AI programs can be easily trained to recognize too. Other information that AI can see but the human eye cannot is also being used to create patterns to detect AI generated videos and images. Fake Catcher, a digital forensics imaging tool, utilizes photoplethysmography to pinpoint synthetic visuals. Photoplethysmography detects minute colour changes in the skin which have a specific pattern in legitimate videos or images but tend to be inconsistent or unchanging in synthetic videos and images (Ciftci, Demir, & Yin, 2020).



Tool marks, or evidence on how the asset was generated, from the programs used to create the media can also be analyzed to create algorithms for detection. For example, FaceForensics++, another digital forensic imaging tool, notes that certain deepfake video programs use a specific manipulation method when creating their videos and therefore all videos created by the program have the same visual artifacts or inconsistencies. This allows for FaceForensics++ to train an AI program to look for these specific artifacts in videos known to have been created by this program. With that training, the program can then be used on videos of unknown provenance to determine if they have the same inconsistencies, indicating that the video has a high likelihood of having been created by that specific deepfake program (Rössler et al., 2019). Other indicators such as differences in frequency domains or analysis of the data behind the visual are also commonly used by algorithms to detect synthetic media.

While synthetic visuals seem to be the current primary focus of research in the field of detecting AI personas and simulacra, similar work is being completed for both text and audio-based generated content. Patterns such as short sentences, repetitive language, audio artifacting, unnatural speech patterns, inconsistent volume changes, in addition to metadata analysis can be used to detect deepfake audio or AI chat bots.

While any one of these patterns can be used by AI to detect AI generated media the strength in using the technology is that it can combine and discover new patterns to improve the likelihood that a piece of media has been accurately labeled as either real or AI generated. Due to their complexity and AI's propensity to develop biases based on skewed data, these AI models need to be loosely supervised or have intermediate human validation—to ensure that biases do not creep into the algorithms distorting the results (Silva et al., 2022).

Individual Mitigations

General awareness and training are key for individuals to be able to question and recognize the potential for AI personas or simulacra. Constant vigilance and zero trust is imperative when considering content being ingested.

Guidance on steps and indicators that individuals can use to determine if something has been AI generated include:

General Tips

Unusual Metadata

Information about the persona or simulacra, such as its creation date, file size, and author, is inconsistent with the purported source of the data.

Unusual Origin

The origin of the persona or simulacra, such as the website or platform from which it was obtained or the person who created it, is not consistent with who would have that information. The source may also be spoofed, such as an account set up to look like a specific person or brand.

Glitches

The persona or simulacra may contain glitches or missing data, such as distorted or missing portions of the image or audio.

Text-Based

AI Text Check

Use apps and websites (e.g.,*OpenAI Classifier*, *CopyLeaks*, etc.) to check on a balance of probability if text is AI generated.

Technical Writing

Look for technical writing and syntax clues, such as:

- shorter sentences
- pattern based writing style (e.g., repeated words or phrases, paragraphs and sentences that are all approximately the same length)

Writing Context

Look at how the information is presented, does it:

- lack analysis or complex thought
- Provide inaccurate information/data

Audio-Based

Codewords

If you personally know the person talking:

- ask a question that only the real person would know
- ask for a previously established codeword to indicate it is the real person

Source Verification

Question the source:

- Is it a known number?
- Is it something you'd expect the person to say?

Validation

If the person is saying something strange or unusual validate from another news source or contact the person via another method (e.g., another phone number, email, in person, etc.)?

Unusual Requests

Watch out for manipulative requests or scenarios where it seems like the caller wants you to panic or they are creating a sense of urgency.

Unusual Payment

Watch out for the caller asking for payment in a hard to trace or recover manner such as wire transfers, cryptocurrency, or gift cards.

Unusual Speech

Unnatural or abrupt changes in speech patterns, such as changes in volume, pitch, or speed.

Image-Based

Unusual Visuals
Anomalies in the image, such as strange teeth or hands.

Inconsistent Light
The light source in the image may be inconsistent with where the shadows and highlights are in the picture.

Overly Smooth
The subject may be overly smooth looking, as if they do not have any pores or imperfections.

Video-Based

Visual Irregularities
Anomalies in the video, such as blinking eyes, blurred ears, unnatural skin textures, or strange mouth movements.

Unusual Movement
Unnatural movements or expressions, such as exaggerated facial expressions or slightly off/unnatural movements.

Unusual Lighting
Inconsistent lighting or shadows, as the subject may be lit differently from the background or other objects in the scene.

Unusual Background
The background or setting of a video may change or be inconsistent with the subject's movements or actions.

Out-of-Sync Audio
The audio and video may not match perfectly. For example, the subject's mouth may move out of sync with the spoken words.

Conclusion

AI technologies, such as the ones being used to develop AI personas and simulacra, are constantly evolving and improving. This means that the risks and threats associated with them may also change and become more sophisticated over time. The mitigation strategies that were effective initially may become outdated or less effective as AI systems become more advanced. As a lot of the programs used to develop these technologies are iterative and many use machine learning, there is also the potential that mitigating the threats will help improve the systems to the point where the mitigations are ineffective.

A core tenet of successfully mitigating the threat posed by AI personas and simulacra is human involvement and awareness. Technological solutions can never completely fix technological threats. Even the AI mitigations presented involve human oversight.

Ultimately, AI—and the personas and simulacra it can develop—is a tool. A tool that can be used for good or ill depending on the hands that wield it. As such, as advancements continue in the world of AI, we must remain vigilant and proactive in developing robust defense mechanisms and ethical frameworks that ensure the tool is used for good. By fostering collaboration among researchers, policymakers, and industry leaders, AI can be harnessed for positive and transformative purposes, ensuring that it becomes a force for good in the ever-evolving digital landscape. However, it is our collective responsibility to steer the course of AI development and safeguard against its misuse, paving the way for a future where technology enhances our lives and empowers us to create a more secure, inclusive, and prosperous world.

Definition Guide

Term	Definition
AI Persona	An avatar created by artificial intelligence to mimic human-like behaviours and characteristics. Usually based on a composite representation of a group or demographic.
AI Simulacra	An avatar created by artificial intelligence to mimic a specific individual’s behaviours and characteristics.
Artificial Intelligence [AI]	A broad field in computer science that involves the simulation of human or human-like knowledge, reasoning, synthesis, or inference by computers.
Cryptographic Signature	A unique digital signature that can be applied to a digital asset to allow for nonrepudiation and validation of the integrity of the asset.
Deep Neural Networks	A subsect of machine learning which focuses on developing algorithms which mimic the way the human brain processes information.
Deepfake	A colloquial term for the process of using an AI algorithm to manipulate existing media (e.g., video, photos, audio) to create a new piece of media which depicts events that did not take place or took place differently.
Disinformation	When false information is shared which is intended to cause harm.
Hallucinations	A response given by AI that is given with confidence but whose accuracy cannot be validated based on the information it was trained on. The answer may be true but given in ignorance, partially true and based off inference from its training, or completely false but seemingly plausible based on the information the AI was trained on.
Hash	A fixed length alphanumeric code algorithmically generated based on a digital asset. The same digital asset run through the same algorithm will produce the same hash so long as the asset has not been modified, allowing it to act as a form of validation for the integrity of the asset.
Machine Learning	A subsect in the field of artificial intelligence which involves computers learning and adapting without explicit human instructions through the analysis of patterns and then drawing inferences from that analysis.
Malinformation	When true information is shared with the intent to cause harm. This is often accomplished by misrepresenting the context of the information.

Term	Definition
Misinformation	When false information is shared (potentially unknowingly) but it is not shared with the intent to cause harm. Harm may still be caused by misinformation, but the individual who shared the information did not intend it.
Natural Language Processing	A subsect of the field of artificial intelligence, which involves the computers analysing and mimicking natural human language and speech.
Nonrepudiation	When an individual cannot deny actions (e.g., modification, access, deletion, etc.) they took relating to a specific digital asset.
Phishing	A type of social engineering which involves cybercriminals using emails to trick individuals into providing access to protected information, credentials, etc.
Social Engineering	A broad category of behaviours which involve using psychology to gain access to protected information, locations, etc. Social engineering can take place in the physical world but is also commonly used by cybercriminals.
Timestamp	A digital record of when a particular event—such as access, modification, movement, etc.— involving a digital asset took place.
Vishing	A type of social engineering which involves cybercriminals using phone calls to trick individuals into providing access to protected information, credentials, etc.
Zero-Trust Model	A security model which requires all individuals regardless of role to be authenticated and authorized before being granted or keeping access to networks, systems, assets, applications, information, etc.

References & Further Reading

Adee, S. (2020, April 29). What are deepfakes and how are they created? *IEEE Spectrum*. https://spectrum.ieee.org/what-is-deepfake

Alavi, S., Charleston, E.S., Du Perron, S., El Khoury, D-N., Freedman, B., Gauthier, J.M., Ghignone, R., Gratton, É., Hémond, A., Henry, E., Jarvie, M., Joli-Coeur, F., Labasi-Sammartino, C., Michaluk, D.J., Morganstein, S., Nagy, A., Stanger, K.M., & Windt, D. (2022, June 21). Canada's Consumer Privacy Protection Act (Bill C-27): Impact for businesses. *BLG*. https://www.blg.com/en/insights/2022/06/canadas-consumer-privacy-protection-act-bill-c27-impact-for-businesses

Allyn, B. (2022, March 16). Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn. *NPR*. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

Al-Sibai, N. (2023, January 25). Amazon begs employees not to leak corporate secrets to ChatGPT: It might already be too late. *The Byte*. https://futurism.com/the-byte/amazon-begs-employees-chatgpt

Benton, J. (2021, January 15). Yes, deepfakes can make people believe in misinformation — but no more than less-hyped ways of lying. *NiemanLab*. https://www.niemanlab.org/2021/01/yes-deepfakes-can-make-people-believe-in-misinformation-but-no-more-than-less-hyped-ways-of-lying/

Berglind, N., Fadia, A., & Isherwood, T. (2022, July 25). The potential value of AI—and how governments could look to capture it. *McKinsey & Company*. https://www.mckinsey.com/industries/public-sector/our-insights/the-potential-value-of-ai-and-how-governments-could-look-to-capture-it

Brown, S. (2021, April 21). Machine learning, explained. *MIT Management Solan School*. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Burke, J. (2015, February 17). History of scams: Nothing new under the sun. *CNBC*. https://www.cnbc.com/2015/02/17/scams-hacking-spanish-prisoner.html

Carbone, C. (2019, September 18). AI can't offer protection from 'deepfakes,' new report says. *Fox News*. https://www.foxnews.com/tech/ai-cant-protect-deepfakes-report-claims

Change the Ref. (2020, October 2). *UnfinishedVotes.com* [Video]. YouTube. https://www.youtube.com/watch?v=m6I_wEetSck

Cheikosman, E., Hewett, N., & Gabriel, K. (2021. October 12). Blockchain can help combat the threat of deepfakes. Here's how. *World Economic Forum*. https://www.weforum.org/agenda/2021/10/how-blockchain-can-help-combat-threat-of-deepfakes/

Christ, M. (2020, August 19). The dark side of AI – Part 1: Cyberattacks and deepfakes. *Computer Science Blog @ HdM Stuttgart*. https://blog.mi.hdm-stuttgart.de/index.php/2020/08/19/ai-cyberattacks-deepfakes/

Ciftci, U.A., Demir, I., & Yin, L. (2020). FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2020.3009287

Civil Code of Québec, 1991, C.64. https://www.legisquebec.gouv.qc.ca/en/document/cs/ccq-1991

Confessore, N. (2018, October 16). New York Attorney General expands inquiry into Net Neutrality comments. *The New York Times*. https://www.nytimes.com/2018/10/16/technology/net-neutrality-inquiry-comments.html

Cooke, R. (2023, March 22). How scammers likely used artificial intelligence to con Newfoundland seniors out of $200K. *CBC*. https://www.cbc.ca/news/canada/newfoundland-labrador/ai-vocal-cloning-grandparent-scam-1.6777106

Cox, J. (2023, January 30). AI-Generated voice firm clamps down after 4chan makes celebrity voices for abuse. *Vice*. https://www.vice.com/en/article/dy7mww/ai-voice-firm-4chan-celebrity-voices-emma-watson-joe-rogan-elevenlabs

Debusmann Jr., B. (2021, March 8). Deepfake is the future of content creation. *BBC*. https://www.bbc.com/news/business-56278411

DeLallo, D. (2020, May 26). Why governments need an AI strategy: A conversation with the WEF's head of AI. *Quantum Black AI by McKinsey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-governments-need-an-ai-strategy

delve.ai. (2023). AI generated persona. *delve.ai*. https://www.delve.ai/blog/ai-generated-persona

Dickson, B. (2019, August 17). The fight against deepfakes. *VentureBeat*. https://venturebeat.com/ai/the-fight-against-deepfakes/

Discover.Bot. (2019, May 7). Facebook Messenger bots 101. *Discover.Bot*. https://discover.bot/bot-talk/guide-to-messaging-apps-chatbot/fb-messenger/

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Canton-Ferrer, C. (2019). The deepfake detection challenge (DFDC) preview dataset. *arXiv*. https://doi.org/10.48550/ARXIV.1910.08854

Drapkin, A. (2023, May 2). How to Avoid the Latest AI Voice Cloning Scam. *tech.co*. https://tech.co/news/ai-voice-cloning-scams

Edwards, B. (2023, January 9). Microsoft's new AI can simulate anyone's voice with 3 seconds of audio. *Ars Technica*. https://arstechnica.com/information-technology/2023/01/microsofts-new-ai-can-simulate-anyones-voice-with-3-seconds-of-audio/

Engler, A. (2019, November 14). Fighting deepfakes when detection fails. *Brookings*. https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/

Federal Bureau of Investigation. (2023, June 5). *Malicious Actors manipulating photos and videos to create explicit content and sextortion schemes.* https://www.ic3.gov/Media/Y2023/PSA230605

Fernick, J. (2021, December 31). On the malicious use of large language models like GPT-3. *NCCGroup*. https://research.nccgroup.com/2021/12/31/on-the-malicious-use-of-large-language-models-like-gpt-3/

Finger, L. (2022, September 8). Deepfakes - The danger of artificial intelligence that we will learn to manage better. *Forbes*. https://www.forbes.com/sites/lutzfinger/2022/09/08/deepfakesthe-danger-of-artificial-intelligence-that-we-will-learn-to-manage-better/?sh=6352651d163a

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. *arXiv*. https://doi.org/10.48550/arXiv.2003.08685

Friedberg, B. (2020, October 20). Investigative digital ethnography: Methods for environmental modeling. *The Media Manipulation Handbook*. https://mediamanipulation.org/sites/default/files/2020-10/Investigative_Ethnography_v1.pdf

Gluska, J. (2023, June 6). How to check if something was written with AI. *Gold Penguin*. https://goldpenguin.org/blog/check-for-ai-content/

Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv*. https://doi.org/10.48550/arXiv.2301.04246

Goodin, D. (2023, June 6). FBI warns of increasing use of AI-generated deepfakes in sextortion schemes. *Ars Technica*. https://arstechnica.com/information-technology/2023/06/fbi-warns-of-increasing-use-of-ai-generated-deepfakes-in-sextortion-schemes/

Government of Canada. (2023a, March 13). *Artificial Intelligence and Data Act.* https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act

Government of Canada. (2023b, March 13). *The Artificial Intelligence and Data Act (AIDA) – Companion document.* https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

Government of Canada. (2023c, April 25). *Responsible use of artificial intelligence (AI)*. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html

Hao, K. (2019, June 14). Deepfakes may be a useful tool for spies. *MIT Technology Review*. https://www.technologyreview.com/2019/06/14/134934/deepfakes-spies-espionage/

Hao, K. (2021, February 12). Deepfake porn is ruining women's lives. Now the law may finally ban it. *MIT Technology Review*. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/

Hasan, H.R., & Salah, K. (2019, April 12). Combating deepfake videos using blockchain and smart contracts. *IEEE Access* (7), 41596-41606. https://www.doi.org/10.1109/ACCESS.2019.2905689

Heikkilä, M. (2022, December 19). How to spot AI-generated text. *MIT Technology Review*. https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/

Henry, J. (2021, April 13). Disinformation and deepfakes fuel growing mistrust. *Catalyst*. https://catalyst.iabc.com/Articles/disinformation-and-deepfakes-fuel-growing-mistrust

Hsu, T. & Myers, S.L. (2023a, April 8). Can we no longer believe anything we see? *The New York Times*. https://www.nytimes.com/2023/04/08/business/media/ai-generated-images.html

Hsu, T. & Myers, S.L. (2023b, May 18). Another side of the A.I. boom: detecting what A.I. makes. *The New York Times*. https://www.nytimes.com/2023/05/18/technology/ai-chat-gpt-detection-tools.html

Hsu, T. (2022, November 4). Worries grow that TikTok is new home for manipulated video and photos. *The New York Times*. https://www.nytimes.com/2022/11/04/technology/tiktok-deepfakes-disinformation.html

IBM. (2023). What is blockchain technology? *IBM*. https://www.ibm.com/topics/blockchain

IBM. (n.d.). AI chatbot that's easy to use. *IBM*. https://tinyurl.com/ykd3jn3r

Jansen, J. AI personas in the media. The Persona Blog. https://persona.qcri.org/blog/ai-personas-in-the-media/

Johnson, D., & Johnson, A. (2023, June 15). What are deepfakes? How fake AI-powered audio and video warps our perception of reality. *Insider*. https://www.businessinsider.com/guides/tech/what-is-deepfake

Juneau, J. (2022, June 1). Val Kilmer spoke in 'Top Gun: Maverick' with assistance of artificial intelligence voice models. *People*. https://people.com/movies/val-kilmer-talked-in-top-gun-maverick-with-help-of-ai-voice-models/

Kane, R. (2023, May 1). 6 examples of real businesses using DALL·E for visual content. *Zapier*. *https://zapier.com/blog/dall-e-examples/*

Kerry, C.F. (2020, February 10). Protecting privacy in an AI-driven world. *Brookings*. https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/

Lawton, G. (2023a). How to prevent deepfakes in the era of generative AI. *TechTarget*. https://www.techtarget.com/searchsecurity/tip/How-to-prevent-deepfakes-in-the-era-of-generative-AI

Lee, S., Tariq, S., Kim, J., & Woo, S.S. (2021). TAR: Generalized Forensic Framework to Detect Deepfakes using Weakly Supervised Learning. *arXiv.* https://doi.org/10.48550/arXiv.2105.06117

Lees, D. (2022, November 4). Deepfakes are being used for good – here's how. *The Conversation*. https://theconversation.com/deepfakes-are-being-used-for-good-heres-how-193170

Malwarebytes. (n.d.). Social engineering. *Malwarebtyes*. https://www.malwarebytes.com/social-engineering

Marr, B. (2022, January 11). Deepfakes – The good, the bad, and the ugly. *Forbes*. https://www.forbes.com/sites/bernardmarr/2022/01/11/deepfakes--the-good-the-bad-and-the-ugly/?sh=154419474f76

McKay, D. (2021, July 23). How deepfakes are powering a new type of cyber crime. *How-To Geek*. https://www.howtogeek.com/devops/how-deepfakes-are-powering-a-new-type-of-cyber-crime/

Microsoft. (n.d.). VALL-E (X). *Microsoft*. https://www.microsoft.com/en-us/research/project/vall-e-x/

Neekhara, P., Dolhansky, B., Bitton, J., & Canton-Ferrer, C. (2020). Adversarial threats to deepfake detection: A practical perspective. *arXiv*. https://doi.org/10.48550/arXiv.2011.09957

Ngoc, N.H., Chan, A., Binh, H.T.T., & Ong, Y.S. (2022). Anti-forensic deepfake personas and how to spot them. *International Joint Conference on Neural Networks*, 1-8. https://doi.org/10.1109/IJCNN55064.2022.9892357

O'Connor, R. (2022, April 19). How DALL-E 2 actually works. *AssemblyAI*. https://www.assemblyai.com/blog/how-dall-e-2-actually-works/

O'Leary, L. (2023, March 24). The next A.I. scam is here—and it could cost you thousands. *Slate*. https://slate.com/technology/2023/03/ai-voice-scams-protect-yourself.html

Palmer, L.C. (2023, June 11). The future of AI ethics: Why government oversight + industry collab + public pressure are key. Dr. Lisa AI. https://www.drlisa.ai/post/the-future-of-ai-ethics-why-government-oversight-industry-collab-public-pressure-are-key

Pandey, K. (2021, August 26). Risks Posed by Deepfake Technology and How to Combat Them. *Jumpstart*. https://www.jumpstartmag.com/risks-posed-by-deepfake-technology-and-how-to-combat-them/

Pine, D., Sharkey, K., Wagner, B., Dykstra, T., Next Turn, Wenzel, M., Victor, Y., Li, S., Sherer, T., Schonning, N., Potapenko, M., Aymeric, A., Maddock, C., xaviex, Jones, M., Latham, L., Pratt, T., & yishengjin1413. (2022, August 10). Cryptographic signatures. *Microsoft Learn*. https://learn.microsoft.com/en-us/dotnet/standard/security/cryptographic-signatures

Potluck, M. (2023, May 5). AI voice scams: Report shares 77% of victims lose money, how common it is, and how to protect yourself. *9to5Mac*. https://9to5mac.com/2023/05/05/ai-voice-scams-how-to-prevent/

Puig, A. (2023, March 20). Scammers use AI to enhance their family emergency schemes. *Federal Trade Commission Consumer Advice*. https://consumer.ftc.gov/consumer-alerts/2023/03/scammers-use-ai-enhance-their-family-emergency-schemes

Rhodes, D. (2022, April 29). Cryptographic hash functions explained: A beginner's guide. *Komodo*. https://komodoplatform.com/en/academy/cryptographic-hash-function/

Roose, K. (2023, April 4). How does ChatGPT really work?. *The New York Times*. https://www.nytimes.com/2023/03/28/technology/ai-chatbots-chatgpt-bing-bard-llm.html

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019, October 27 - November 2). FaceForensics++: Learning to detect manipulated facial images. *2019 IEEE/CVF International Conference on Computer Vision,* 1-11. https://doi.org/10.1109/ICCV.2019.00009

Rundle, J. (2023a, April 26). AI's effects on cybersecurity concern U.S. Officials. *WSJ Pro Cybersecurity*. https://www.wsj.com/articles/u-s-officials-raise-concerns-about-ais-cybersecurity-implications-f32496ed

Rundle, J. (2023b, May 25). Cybersecurity Chiefs Navigate AI Risks and Potential Rewards. *WSJ Pro Cybersecurity*. https://www.wsj.com/articles/cybersecurity-chiefs-navigate-ai-risks-and-potential-rewards-9138b76d?mod=djemCybersecruityPro&tpl=cy

Sadekov, K. (2023, April 11). Types of chatbots: Rule-based chatbots vs AI chatbots. *MindTitan*. https://minditan.com/resources/guides/chatbot/types-of-chatbots/

Sample, I. (2020, January 13). What are deepfakes – and how can you spot them?. *The Guardian.* https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

Satariano, A., & Mozur, P. (2023, February 7). The people onscreen are fake. The disinformation is real. *The New York Times*. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

Schwartz, O. (2018, November 12). You thought fake news was bad? Deep fakes are where truth goes to die. *The Guardian*. https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

Sensity Team. (2021, February 8). How to detect a deepfake online: Image forensics and analysis of deepfake videos. *Sensity*. https://sensity.ai/blog/deepfake-detection/how-to-detect-a-deepfake/

Silva, S.H., Bethany, M., Votto, A.M., Scarff, I.H., Beebe, N., & Najafirad, P. (2022). Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, (4). https://doi.org/10.1016/j.fsisyn.2022.100217

Silverman, C. (2018, April 17). How to spot a deepfake like the Barack Obama–Jordan Peele video. *Buzzfeed*. https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed

Simonite, T. (2018, March 9). AI has a hallucination problem that's proving tough to fix. *Wired*. https://www.wired.com/story/ai-has-a-hallucination-problem-thats-proving-tough-to-fix/

Singel, R. (2018). Filtering out the bots: What Americans actually told the FCC about Net Neutrality repeal. *The Center for Internet and Society*. https://tinyurl.com/5db9taps

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Wook Kim, J., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). Release strategies and the social impacts of language models. *OpenAI*. https://d4mucfpksywv.cloudfront.net/papers/GPT_2_Report.pdf

Somers, M. (2020, July 21). Deepfakes, explained. *MIT Management Solan School*. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Statt, N. (2018, January 24). Fake celebrity porn is blowing up on Reddit, thanks to artificial intelligence. *The Verge*. https://www.theverge.com/2018/1/24/16929148/fake-celebrity-porn-ai-deepfake-face-swapping-artificial-intelligence-reddit

Stern, J. (2023, April 28). I cloned myself with AI. She fooled my bank and my family. *The Wall Street Journal*. https://www.wsj.com/articles/i-cloned-myself-with-ai-she-fooled-my-bank-and-my-family-356bd1a3

Streets, J. (2021, December 2). 10 examples of AI in customer service. *TechTarget*. https://www.techtarget.com/searchcustomerexperience/feature/10-examples-of-AI-in-customer-service

Stupp, C. (2019, August 30). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *WSJ Pro Cybersecurity*. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Surovell Isaacs & Levy PLC. (2020). Prevalence of 'deepfakes' causes courts to question validity of evidence. *Surovell Isaacs & Levy PLC.* https://surovellfirm.com/criminal-law/prevalence-of-deepfakes-causes-courts-to-question-validity-of-evidence/

The Telegraph. (2022, March 17). *Deepfake video of Volodymyr Zelensky surrendering surfaces on social media* [Video]. YouTube. https://www.youtube.com/watch?v=X17yrEV5sl4

Thomas, K. (2023, June 12). Understanding AI risks and how to secure using zero trust. *AT&T Business*. https://cybersecurity.att.com/blogs/security-essentials/understanding-ai-risks-and-how-to-secure-using-zero-trust

TrendMicro. (2019, September 5). Unusual CEO fraud via deepfake audio steals US$243,000 from UK company. *TrendMicro*. https://tinyurl.com/3w5e8dap

Truepic. (2023a). Truepic Display. *Truepic*. https://truepic.com/truepic-display/

Truepic. (2023b). What happens if real is actually fake?. *Truepic*. https://truepic.com/revel/

Tseng, P. (2018, March 29). Canada: What can the law do about 'deepfake'? *Mondaq*. https://www.mondaq.com/canada/copyright/687716/what-can-the-law-do-about-deepfake

Twohig, I. (2022, March 31). Using digital ethnography as a research tool for identifying user personas. *Indeemo*. https://indeemo.com/blog/ethnography-user-persona

Vaas, L. (2019, June 17). I'd like to add you to my professional network of people to spy on. *NakedSecurity by Sophos*. https://nakedsecurity.sophos.com/2019/06/17/id-like-to-add-you-to-my-professional-network-of-people-to-spy-on/

Wall Street Journal. (2023, April 28). *I challenged my AI clone to replace me for 24 hours* | *WSJ* [Video]. YouTube. https://www.youtube.com/watch?v=t52Bi-ZUZjA&t=4s

Weise, K., & Metz, C. (2023, May 1). When A.I. chatbots hallucinate. *The New York Times*. https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html

Weiss, M. (2019, December 17). Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions. *Technology Science*. https://techscience.org/a/2019121801/

Wilhelm, P. (2016, December 5). Meet Zo, Microsoft's new and improved chatbot. *TechRadar*. https://www.techradar.com/news/meet-zo-microsofts-new-and-improved-chat-bot