Brace yourself!

Why Managers should adopt a Synthetic Media Incident Response Playbook in an age of Falsity and Synthetic Media

Accepted for publication in Business Horizons.

Lucas Whittaker[a]

Jan Kietzmann[b*]

Kate Letheren[c]

Rory Mulcahy[d]

Rebekah Russell-Bennett[e]

[a] Centre for Behavioural Economics, Society and Technology (BEST)/School of Advertising, Marketing and Public Relations, Queensland University of Technology, 2 George Street, Brisbane City, 4000, Queensland, Australia. Email: ll.whittaker@qut.edu.au

[b] Peter B. Gustavson School of Business, University of Victoria, Victoria, British Columbia, Canada. Email: jkietzma@uvic.ca

[c] Centre for Behavioural Economics, Society and Technology (BEST)/ School of Advertising, Marketing and Public Relations, Queensland University of Technology, 2 George Street, Brisbane City, 4000, Queensland, Australia. Email: kate.letheren@qut.edu.au

[d] USC Business School, University of the Sunshine Coast, Sippy Downs Drive, Maroochydore DC, 4558, Queensland, Australia. Email: rmulcahy@usc.edu.au

[e] Centre for Behavioural Economics, Society and Technology (BEST) School of Advertising, Marketing and Public Relations, Queensland University of Technology, 2 George Street, Brisbane City, 4000, Queensland, Australia. Email: rebekah.bennett@qut.edu.au

[*] Corresponding author

Brace yourself!

Why Managers should adopt a Synthetic Media Incident Response Playbook in an age of Falsity and Synthetic Media

Abstract

This article acts as a guide for managers to navigate through the looming threats presented by synthetic media (e.g., deepfakes and other media generated by artificial intelligence). It first offers a short overview of the evolution of media manipulation to provide context to the new era of synthetic media. Next, it presents the problems associated with synthetic media through the lens of veridicality and heuristics to illustrate how consumers have little choice but to believe what they see, read, and hear online. We outline the most likely and impactful types of synthetic media threats and attacks before we present a Synthetic Media Incident Response Playbook to inform managers about six specific phases to help them prepare, assess, detect, analyze, and recover from synthetic media incidents and to coordinate their lessons learned.

Keywords: synthetic media; deepfakes; generative adversarial networks; artificial intelligence; falsity; fake news; cybersecurity

Brace yourself!

Why Managers should adopt a Synthetic Media Incident Response Playbook in an age of Falsity and Synthetic Media

When Alan Turing devised the original "imitation game" in 1950 (Turing, 1950), he was interested in the potential for a computer to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. A human participant of this imitation game would send text-based questions to their conversation partner, and if this participant could not determine if the answers that appeared on the screen in front of them were produced by another human or by a machine, the machine was thought to be as "intelligent" as a person. This "Turing test" was as simple as it was powerful, and it has been discussed ever since in the debate on whether machines could truly be as intelligent as people.

The advancement of artificial intelligence (AI) has once again revived the discussion of the value of the Turing test. Some argued that passing the test should not be a valuable goal for AI (French, 2000), because AI should be designed to solve problems rather than make us believe in its humanness. While this might be true, the ability of AI to "act" human-like certainly becomes imminently relevant when the goal of the technology *is* the deception of people (Natale, 2021).

In our increasingly audio-visual rich world, AI imitators continue to become more adept at tricking us, not only by writing like us, but by looking, speaking, and moving like us too (Kietzmann et al., 2020). We have seen the meteoric rise of "synthetic media", or in other words, artificial media (such as deepfake audio and videos) that are *autonomously* generated through AI and forms of machine learning. While the results are quite impressive from a computing and media perspective, the fact that fake synthetic media cannot be told apart from authentic media is particularly concerning for all organizations, whether small or big, product or service oriented, or private or public sector. Predictions are dire and suggest that brands, governments, managers, employees, and customers will fall victim to synthetic media attacks (Bonfanti, 2020). Synthetic media attacks are orchestrated to extract financial resources, inflict harm to brand reputation, shake customer confidence, and gain access to IT resources. In anticipation of what is likely to come, how can managers brace themselves for malicious synthetic media that look and sound genuine, but are completely unsanctioned?

This article acts as a guide for managers to navigate through this new era of 'false' or 'fake' media. First, we provide a short overview of the evolution of media manipulation to provide context to the new era of synthetic media and discuss how these media are generated by AI. We then discuss synthetic media through the lens of veridicality and heuristics to illustrate how consumers have little choice but to believe what they see and hear online. The following section outlines the most likely and impactful types of synthetic media threats and attacks. Next, we argue managers stand to gain much by learning from cybersecurity practices. We present a Synthetic Media Incident Response Playbook to inform managers about six phases they should consider to prepare, assess, detect, analyze, and recover from synthetic media incidents and to coordinate lessons learned across their industry.

The Evolution of Media Manipulation

People have been manipulating the common types of media – text, audio, images, and video – throughout history to create a sense of falsity. We have come to know and accept – even adore – such manipulations, whether they exist in the arts (think of Andy Warhol's art), in entertainment (Star Wars), or in advertising (airbrushed ads). The fact that there is even a special term for the rare occasion that content is not manipulated ('unretouched' media) speaks volumes about the degree of manipulation around us.

It all started with *analog media manipulation* (Campbell et al., 2021). An early example was of Abraham Lincoln, who utilized printmakers and their engraving and lithography skills to manually improve his facial and bodily features and dress within prints. After his assassination, a famous print of him was created by superimposing his head onto John C. Calhoun to cut costs and meet public demand (Holzer, 1979). The later invention of photographic film allowed images to be retouched using chemicals, blades, pencils, and brushes dipped in ink or watercolor (Young, 2022). More recently, authoritarian rulers such as Stalin revised history by cropping and airbrushing individuals out of photographs when expedient (Swerzenski, 2021). Early filmmakers conducted analog video manipulation through camera angles, movement, and lens effects. Using models and specific angles, forced perspective was an optical illusion which altered the perception of scenes (Fabe, 2014). Films were edited by manually cutting and gluing celluloid film together before technological innovations helped mechanize this process (Landay, 2019). Analog manipulation examples exist for other media modalities (e.g., audio and text), which also involved significant technical skill and expertise to convincingly manipulate. Those who manipulated media had to have advanced artistic skills, but their ability to access the original media and the means to distribute the manipulated version was constrained by the physical nature of the media. As a result, analog media manipulation presented a relatively low risk to brands.

The imitation game between technology and humans grew in importance with digital tools and the Internet. *Digital media manipulation* involves using computers to alter and produce media (Campbell et al., 2021). Editing functions, such as those present in Adobe Photoshop, have widely facilitated how media could be altered on computers, but similar functions have also been integrated into apps like Snapchat and Instagram. Such functions allow users to easily retouch images and apply filters to enhance the perceptual quality of their images. These functions have even become automated within cameras themselves, being able to apply digital alterations automatically to enhance image aesthetics (Swerzenski, 2021). Digital video and audio manipulation tools, and the whole editing industry spawned by Adobe Photoshop, make it possible for prosumers and professionals alike to add digital effects and otherwise precisely alter captured media through computers. More and more fake content is the result, which decreases the control over manipulation and increases potential brand risk. However, a high-quality manipulation that can fool consumers still demands significant time and editing skills.

A new era of *synthetic media manipulation* started in 2017 when a Redditor with the name 'Deepfakes' showcased videos that they created by using open-source face-swapping technology to place unsuspecting celebrities into adult films. The important difference to digital manipulation is that such tricks can be accomplished through autonomous editing or content generation (text, audio, video, imagery) by leveraging neural networks to merge,

combine, replace, and superimpose elements from various media (Kietzmann et al., 2020; Maras & Alexandrou, 2019). Put simply, once the algorithm understands the key characteristics that two individuals have in common (e.g., that their faces share), swapping an adult film actress and celebrity can be automatically done, with authentic results.[1] The same principle can also be used to control the words spoken by a single individual within a video – effectively making people say things they never did, or would, say. A fake but real-looking recording of a president swearing in an interview is as easily created as one of a CEO in a compromising situation (or vice versa).

Beyond the mere replacement and merging of physical features, neural networks can also be used to automatically generate altogether new content. *Generative adversarial networks* (GANs) learn how to generate new data to an ultra-realistic standard from the training data input into the model (Whittaker et al., 2020). Websites such as www.whichfaceisreal.com are modern variations of the imitation game – they put an onlooker's perceptual skills to the test to determine which human portrait image is authentically human and which has been generated by a GAN. Non-humanistic data can also be generated. GANs can be trained on the artworks of Monet and Van Gogh to generate original artwork in their artistic styles (Zhu et al., 2017). GANs can even generate entire symphonies or photorealistic images of rooms (down to the couches, chairs, and lamps), and whole urban environments (Gadde et al., 2021).

Other important synthetic media manipulation examples (and tools) include:

- The use of neural networks in tools like DALL·E that can generate photorealistic images from scratch using text descriptions (e.g., "create a car designed by Steve Jobs") (OpenAI, 2021).[2]
- Text-to-speech applications like Amazon Polly or Murf use an individual's voice as training data and, through employing an algorithm, vocalizes typed text using the voice of this individual.
- GANs can also permit 'voice conversion', where one individual's speech is synthesized from another individual's speech without altering the linguistic or phonetic content. (Saito et al., 2017). In other words, the words spoken by one person can sound exactly like they were uttered by someone else.
- Neural networks also permit the manipulation of text. The autoregressive language model GPT-3 uses deep learning to produce human-like text. It is a language prediction model facilitated by a neural network that has been trained on billions of words, including almost all data available on the internet (Romero, 2021). GPT-3 can generate news articles practically indistinguishable from those written by humans. It is this capability that powers automatic copywriting assistance, auto-generated emails, grammar correction, and most impressively, many of the chatbots used by organizations today.

[1] For more details on this process, please read "Deepfakes: Trick or Treat?" (Kietzmann et al., 2020).

[2] The reader might want to try sites like wombo.art by typing in a description of the artwork they wish to generate, selecting an art style, and watching a synthetic piece of art form before their eyes.

- Tools such as Synthesia provide business clients with automatically created videos of humanistic AI avatars (see Figure 1) which verbalize a script provided in the form of text by the client. Likewise, GANs such as FutureGAN can be trained to predict future frames in a video, creating video frames based on past frames to generate plausible, yet synthesized video (Vondrick et al., 2016).

Figure 1: An autogenerated AI avatar from Synthesia

We are Wired to Believe Synthetic Media

The impact resulting from the evolution of manipulation is clear – more and more types of fake content, generated with fewer resource requirements, skill, and time, are increasing in quality to the point where they appear undeniably genuine. Media manipulation, evolving from analog, to digital, and now to synthetic, has resulted in managers having decreased control and incurring increased risks to their brands. For illustration, Figure 2 shows stills from a Dior video advertisement that originally featured Charlize Theron[3]. Numerous deepfake iterations of the ad have been produced, with Theron being replaced with Margot Robbie, Angelina Jolie, and even Rowan Atkinson. While these examples might appear as fodder for harmless memes, none of these versions were sanctioned by Dior. Similarly, Robbie, Jolie, and Atkinson did not consent to having their likeness used in this fashion. While these examples might appear risk-free, clearly the technology could potentially be used to inflict harm on all associated brands (Dior's brand and the celebrity's personal brand).

[3] View a sample of the Dior videos at https://bit.ly/jadoredfs

Figure 2: Deepfakes of Charlize Theron

While it is impossible to tell which of the three women portrayed in the image was the original model in the ad, a question emerges regarding why we would believe them in the first place.

Traditionally, when confronted with different media types, consumers place trust in the media source when assessing the credibility of their content. With too much information around us to question everything we see and hear; we trust content from people or institutions we trust. Unsurprisingly, when consumers believe this *source credibility* to be high, they rely on and use the information more often than people who evaluated the source to be less credible (Wanta & Hu, 1994; Beaudoin & Thorson, 2005).

Today though, when more than two-thirds of Americans get news from social media, a platform where thoroughly vetted and fact-checked posts from publishers get the same packaging as eccentric conspiracy theories, we pay much less attention to the sources of online content. Especially in our increasingly connected media landscape, consumers place more and more trust in user-generated content, citizen journalism, and socially curated news. As a result, when the credibility of the source is viewed as less important, the question of how people evaluate the 'truthiness' of the content itself becomes more important and interesting. How do we assess if what is presented online is real? How do we evaluate the believability of media messages?

Our willingness and ability to assign credibility to content, as it turns out, is also tied to the type of media consumed. When we think about *media credibility* (as opposed to source credibility), we rely on the concept of *veridicality*. When the perception of phenomena or

objects presented through media is factually correct, it is considered *veridical*. A veridical perception accurately reflects how things truly are. In contrast, perceptions of phenomena can be *nonveridical.* Nonveridical perceptions are formed from media content that mischaracterizes the facts and does not accurately portray reality. This nonveridical perception exists even if one believes that their perception of the phenomena reflects reality, even though it does not (Schwartz, 2016). This notion of veridicality "has been understudied as a major factor in the credibility of a particular message" (Lee et al., 2010, p.312). Here is where our synthetic media problem starts – perceived veridicality is tied to media modality.

Humans believe themselves to be rational, but we are guided by emotional and irrational thinking more than we would care to admit or are even aware of. Extensive research on this phenomenon has been undertaken to explore the implications of these cognitive shortcuts, better known as *heuristics* (Tversky & Kahneman, 1974). Relevant to media appraisal, one such heuristic we employ is our inclination to trust visuals – information-rich media – because the richness of audiovisual material requires the allocation of more cognitive resources which can lead to cognitive overload (Lang, 2000). This overload can negatively influence one's ability to systematically process information. Instead, information-rich content is superficially processed via heuristic processing, leading individuals to believe and share it more readily (Sundar et al., 2021). Information richness also influences our appraisal of credibility. If media looks and talks like the real thing, information credibility has been found to increase regardless of source credibility (Lee et al., 2010). Accordingly, we evaluate the sources of richer audiovisual messages less systematically than leaner information presented via text, assigning more credibility to modalities such as video and audio than we do to text and images.

The realistic quality of synthetic media, especially audio and video footage, challenges our innate ability to separate what is real and what is not. Making matters worse, our reliance on cognitive shortcuts is particularly put to the test when we see people or things we recognize, as we have always relied on a *familiarity heuristic*, as it is "our tendency to assume that if something is familiar, it must be good and safe" (Steinmetz, 2018). Similarly, our perception of 'what reality is' influences our evaluation of new information. Humans are influenced by a *confirmation bias*, which describes the tendency to seek or interpret evidence in a manner consistent with one's existing beliefs or expectations (Nickerson, 1998). Put differently, the degree to which synthetic media accurately represents a familiar reality or aligns with our prior beliefs can influence our perception of its veridicality.

When we bring it all together, an interesting story emerges. Synthetic manipulation tools do not require significant time, skills, or resources to use. As such, synthetic media can be generated by anyone. This is particularly alarming as consumers of synthetic media fall victim to the "veridicality effect" – autogenerated content is perceived as being veridical due to its high level of realism. However, nonveridical perceptions are unknowingly created instead as the content is inauthentic, a fact potentially unnoticed by the consumer. This veridicality effect is enhanced by the heuristics we utilize during media appraisal. Multimodal, information-rich synthetic media are more likely to be believed due to the high degree of realism present (unless it is intentional satire like the Rowan Atkinson version of Charlize Theron above). Other heuristics, such as the familiarity heuristic and confirmation bias reinforce the misplaced belief that what is being seen and heard is veridical, as synthetic content may appear familiar to us or align with what we already believe to be true. Therefore, although seeing and hearing may be believing due to the veridical illusion assumed by synthetic media, belief may also determine what one truly perceives.

One worry is that the nonveridical nature of synthetic media will devalue authentic information by removing the 'epistemic backstop' of video and audio recordings which societies rely on for credible testimony and transmission of knowledge (Rini, 2020). After all, individuals could only discern between human and deepfake faces about 50% of the time, akin to guessing (Rössler et al., 2018). These difficulties extend to text where the human identification rate of GPT-2 generated fake reviews was roughly 55% (Salminen et al., 2022). These new synthetic realities generate 'reality apathy' by causing people to give up trying to discern between what is authentic and synthetic, ceasing their efforts to become informed citizens (Dowdeswell & Goltz, 2020; Vaccari & Chadwick, 2020) and thereby potentially eroding the perceived credibility of fundamental civic media, politics, academia institutions – and organizations.

Weaponized Synthetic Media Change the World: Five Risks for Organizations

Although some appear less worried and argue that "widespread existence is not the same as widespread impact on the public" (Duke Today, 2020), others are much more concerned for brand managers and their ability to protect themselves from the weaponization of synthetic media. These concerns are increasingly warranted as a response to five types of synthetic media risks that are surfacing, each with the ability to influence social decision-making and perceptions via deception (Bonfanti, 2020).

1. Reputational Risks

In response to the hands-off stance that Meta (then Facebook) took towards fake news (particularly the fake video of speaker Nancy Pelosi), two artists and an advertising company generated and shared a deepfake video of Mark Zuckerberg saying: "Imagine this for a second: One man, with total control of billions of people's stolen data, all their secrets, their lives, their futures." (Cole, 2019). In a similar project, Kim Kardashian was portrayed in a video saying: "When there's so many haters, I really don't care because their data has made me rich beyond my wildest dreams" (Cole, 2019) (see Figure 3). Both videos went viral on Instagram, of course leading viewers to believe them to be real even though they were identified as manipulated, synthetic media. Of course, the resulting damage to a firm's reputation can harm customer relationships and impact the financial and social capital and/or the market share.

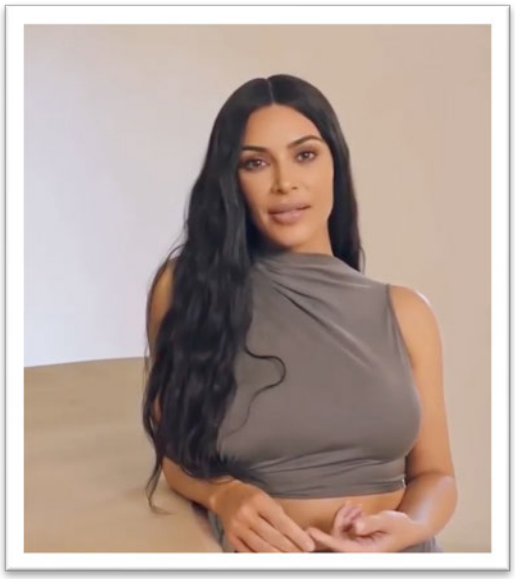

Figure 3: Still image of Kim Kardashian in a fake synthetic video[4].

Celebrities like Zuckerberg and Kardashian are popular targets of synthetic media for two reasons. First, much video footage exists of them which can be used to train the algorithm used to create deepfakes. Second, because of their global recognition and importance, the synthetic media has the highest chances of going viral. But this is not to suggest that only global celebrities can be targets. Jordan Peterson, a Canadian psychology professor, was recently featured, without his permission, on a website that allowed anyone to speak in his voice (Novak, 2019). In 2019, the Malaysian Minister of Economic Affairs Mohamed Azmin Ali found his reputation tarnished by a highly compromising video. In fact, Malaysian journalists and politicians were invited to view this video on WhatsApp to inflict the most damage to Azmin Ali's social standing.

Not even private citizens are safe anymore if some training data of them exists. An 18-year-old law student found hundreds of explicit deepfakes of her face on the bodies of adult film actresses when reverse Google image searching a photo of herself one night, despite never having taken or shared explicit photos of herself (Melville, 2019). With such risks becoming increasingly common, anyone whose reputation is of particular corporate value (e.g., every CEO or board member that has been featured on earnings calls, YouTube videos, TED talks, and podcasts) must brace themselves for the risks of synthetic media impersonation.

2. Economic and Corporate Secrecy Risks

The case of the 'Fake Johannes' is likely the most well-known instance of deepfake fraudsters tricking a company to date. The thieves cloned the voice of the CEO of a German parent company and then called a subsidiary to this parent company, a U.K-based energy firm. The call was so convincing, down to the tonality, punctuation, and German accent, that the

[4] View the Kim Kardashian video at: https://bit.ly/3l84eWa and the Mark Zuckerberg video at https://bit.ly/3yCqOOE

executive who responded to the call in the U.K. wired €220,000 (US$243,000) into an untraceable account (Stupp, 2019). Researchers at Symantec said they have found at least three cases of such cyber-impersonation campaigns (which the FBI calls 'business identity compromises') of executives' voices being mimicked to swindle companies (Harwell, 2019).

When combined, synthetic media that threaten reputational capital can have important impacts on corporate secrecy too. When apps can create deepfakes on demand (e.g., Nudify, which automatically undresses people in photos) such synthetic media can easily be used to blackmail those depicted for money through cryptoviral extortion attacks to gain leverage for access to classified information or to influence corporate decisions. Soon, organizations will increasingly need to brace themselves for such synthetic ransomware and phishing attacks.

3. Security Risks

With the ability to impersonate others, risks extend beyond merely financial. More insidiously, perhaps, malfeasants can induce employees to forward them innocuous business documentation such as transaction details or customer orders which can then be leveraged for criminal activity or fraud on an ongoing basis (Gralla, 2019).

The various ways which synthetic media can impact corporate security are relatively unsurprising. Some focus on exploiting our weakened internal alarm systems. Symantec, the enterprise security software company, warns that employees exposed to fake videos and audio might be deceived into sharing login credentials which allow attackers to gain access to an enterprise's network (Gralla, 2019). Beyond these social engineering practices, technical mechanisms that demand video identification or simple facial recognition methods are no longer safe either. When it is not clear that the person on the screen is really the one sitting in front of the camera, anyone relying on biometrics needs to brace themselves for new, synthetic security risks (Tissler, 2021).

4. Intellectual Property Risks

The original source video that was used in the deepfake of Zuckerberg belonged to CBS, the American commercial broadcast television and radio network. CBS demanded that Facebook removed the video as it displayed the unauthorized use of the CBSN trademark. With the predicted increase in synthetic media, this means that firms need to brace themselves for increasing legal costs associated with protecting and enforcing intellectual property rights to limit the loss of value of intellectual property assets (e.g., diminished licensing or product revenues).

5. Governance and Operational Risks

There are significant risks related to synthetic media and its potential impact on systems and policies that govern how organizations are controlled and operate. These risks also extend to the mechanisms by which they and their people are held to account. For instance, China's State Taxation Administration fell victim to criminals who used synthetic images to create identities to set up a shell company that issued fake tax invoices worth as much as 500 million yuan (approximately US$76.2 million) (Borak, 2021). Such malicious technology use can even undermine national security. A very alarming example recently took the form of a deepfake video of Ukraine's President Volodymyr Zelensky, in which he told his citizens to surrender, lay down their arms, and return to their families (Miller, 2022). Such deceitful, fake propaganda

can have damning consequences from the perspective of governance and operations in Ukraine. An indirect impact of this deceptive deepfake might be that Ukrainians may not believe future, real footage of their leader, including videos in which he shares well-intended guidance on how to behave in this unprecedented time.

Adopting a Synthetic Media Incident Response Playbook

These examples of early synthetic media risks are indicative of a worrying trend. The trend is particularly concerning because such attacks will continue to increase in quantity, quality, and perceived veridicality. In addition, convincing synthetic media are becoming increasingly easy to create given the development of consumer-grade apps (Kietzmann et al., 2020). These conditions create a perfect storm for synthetic media attacks against enterprises, which experts believe are to become more frequent (Davis, 2021; Gralla, 2019). However, enterprises are unprepared to weather malicious media attacks (Gow, 2021). Perhaps this is a product of managers being cognitively biased via a miscalibration of subjective probabilities. Unknowingly, we are guided by comparative-optimism effects – it is human to believe that bad or rare events are more likely to happen to others than oneself (Ucbasaran et al., 2010). Unfortunately, this means that we make fewer good choices to protect ourselves, whether that impacts us as private citizens (e.g., too few go for routine cancer screening), or as corporate entities.

In the face of new adversaries and their new weapons engaged in false information warfare, a passive attitude towards synthetic media threats is inadvisable. In order to safeguard carefully built organizational assets (e.g., reputation, financial resources, and intellectual property), managers are well advised to develop and implement appropriate sets of procedures to identify, remediate, and recover from vulnerabilities and synthetic media incidents affecting their brands.

For this purpose, we use the 2021 Cybersecurity Incident & Vulnerability Response Playbooks issued by the U.S. Government's Cybersecurity & Infrastructure Security Agency (CISA, 2021) as a foundation for a 'Synthetic Media Incident Response Playbook'. It is important to keep in mind that this is the first version of such a playbook. Naturally, as synthetic media threats and response techniques evolve, so too does the need to update the Playbook.

The Playbook consists of six phases, including the preparation and assessment phases (which happen before an incident), the detection and containment phases (which take place during an incident), and the post-incident activities and coordination phases (which take place after an incident). With the mindset that "after one incident is before the next incident", these phases are arranged in a circular fashion (see Figure 4).

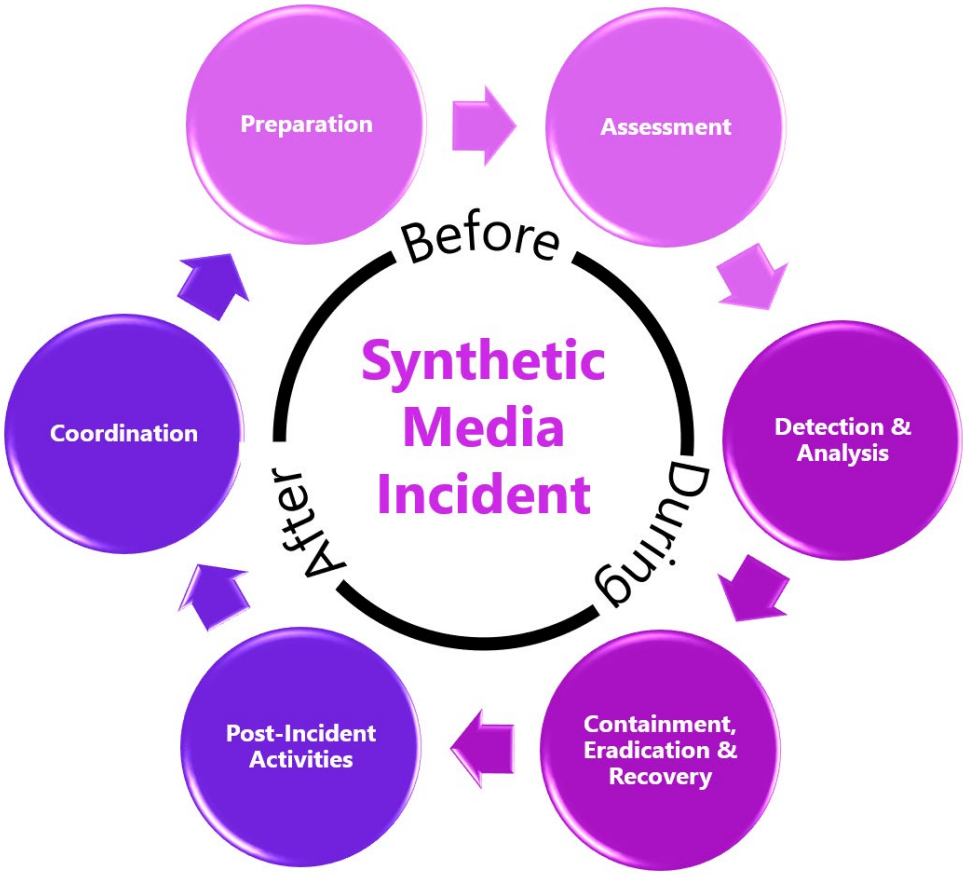

Figure 4: Six phases of the Synthetic Media Incident Response Playbook

1. Preparation Phase

The preparation phase focuses on activities that need to happen *before fake synthetic media incidents occur.* The goal is to reduce the impact such events can have on the organization. Since many organizations are still ill-prepared even for common cybersecurity events, brand managers are more likely completely unprepared for fake media incidents. We suggest that managers:

Educate Everyone on Threats and Notification Procedures: Teach members of the organization what synthetic media incidents are, why they are so hard to tell apart from genuine media, and why we are primed to believe what we see. Make sure this includes high-value, high-risk individuals such as board members, C-level executives, middle management with significant discretion, and the finance department. Also, ensure this becomes part of the onboarding procedures for all new hires. A worthwhile activity might also be to include external stakeholders in this exercise (e.g., supply chain partners, investors, key clients), so that they are also prepared.

Establish a Fake Media Posture: Ensure that fake media incidents are treated as serious threats, not as memes, by everyone in the organization. Communicate the potential financial losses,

reputational losses, and eroded customer trust that can result from malicious fake media attacks. Establish a "you see something, you say something" culture, where individuals know that it is the responsibility of everyone to raise an alert on anomalous media. Establish a "Don't Trust, Verify" attitude, where activities above a certain risk threshold are confirmed before they are executed (not unlike two-factor authentication). Even simple passphrases might help protect from instances like the 'Fake Johannes'.[5]

Establish Incident Response Teams: Designate specific individuals and ensure they are ready to respond when an incident materializes. Who should assess any reports of suspicious activities? Who should lead mitigation activities when a fake media event happens? Someone from the branding team, cybersecurity, or from a third-party organization? Without such clarity, team members will look towards others to take charge. Nobody will see such an event as their responsibility, and valuable time will be lost.

Train: How will the designated individual or individuals be trained on how to respond to a fake media event? In the cybersecurity field, there are appropriate training and education tools. As these do not yet exist widely in the fake media context, organizations need to prepare by directing resources toward those who will deal with incidents.

Devise Escalation Policies: Are synthetic media threats on the organization's risk register? When, how, by whom, and to whom will fake media incidents be escalated and reported? The marketing team? The cybersecurity team? Legal counsel? The board? Develop policies and plans regarding the potential notification, interaction, and evidence sharing with authorities (e.g., the Securities and Exchange Commission) and law enforcement.

Test: To thwart cybersecurity incidents, organizations often (although not often enough) engage in so-called 'penetration testing' or 'white hat hacking' activities. Penetration testing is a "method for gaining assurance in the security of an IT system by attempting to breach some or all of that system's security, using the same tools and techniques as an adversary might." (NCSC, 2017). Much like running a fire-drill to ensure everyone knows what to do, imitate the actions of a hostile fake media attack to see if the organization is well prepared for it.

Protect: In times of synthetic media, intellectual property laws (e.g., copyrights and trademarks) are ineffective, reactive, and time-consuming means for protecting authentic media (as can be seen from the Zuckerberg example). Increasingly, organizations are turning to protecting image provenance, often through blockchain-based digital signatures which verify that media have not been unknowingly altered or tampered with. At the same time, The Coalition for Content Provenance and Authenticity (C2PA), with members including Adobe, Microsoft, Arm, Intel TruePic, and the BBC, is working on developing a secure media provenance standard to allow content creators and editors to create media that cannot be secretly tampered with (Woollacott, 2022).

[5] Interestingly, when the thieves made a second attempt to defraud more money, the director grew suspicious and called the real CEO directly. When the thieves called back, the director talked to the fake Johannes while being on the phone with the real Johannes at the same time.

2. Assessment Phase

While all organizations could fall victim to synthetic media attacks, a methodological security and risk assessment of the threat environment can help organizations understand their overall attractiveness for malicious attackers. We suggest that managers undertake such an assessment to facilitate better-informed management decisions, which should include the following considerations:

Know your Crown Jewels: In cybersecurity, an organization's crown jewels are its mission-critical information assets which become targets of attacks. In a synthetic media context, managers need to identify possible reasons why others might want to attack them. Reasons could include a provocative product launch, activities that might be seen as controversial, political engagement on topics that are divisive, but also leaders who are in the public eye or have been featured a lot in the media (thereby offering plenty of training data for the synthetic media generation process). Nonetheless, this is not to instill a false sense of security – everyone can become a target, and even the most unsuspecting companies have fallen victim to such attacks.

Know your Enemy: Once an organization knows about its crown jewels, it needs to assess the adversarial threats to these assets. Who might engage in the creation of fake media to hurt a brand? How do their motives, habits, methods, and levels of commitment vary? Are these non-malicious, or malicious? Among others, some profiles include technologists who want to showcase their technical prowess (but have nothing against the brand), brand fans (who want to promote the brand but might inadvertently endanger it through their efforts), hacktivists (who support particular political viewpoints or causes), subversive or extremist groups, criminals (with goals of personal gain), and even nation-state sponsored attackers (during political warfare). Are they individual actors, or communities? How big, powerful, and committed are they? Prioritizing this list and knowing enemy #1 is the goal of this activity.

Understand Potential Attack Types: Are actors more likely to attack a single target or stage a full spectrum attack with multiple targets? Are their targets internal (e.g., the Fake Johannes), external (e.g., reputation damage), or both? What types of media are the most likely delivery mechanisms? Are they likely more of a nuisance (e.g., a phishing attack) or a significant threat (e.g., a multistage exploit)?

3. Detection & Analysis Phase

Even when an organization has prepared for potential fake media events, when they know what their most likely attack surface is (i.e., their crown jewels) and understand their enemies and their most likely attack types, this is no assurance that events will not happen. The most challenging aspect of a synthetic media attack is accurately detecting the media as fake. This is where most of the victims to date have failed. This is not hard to believe, as compared with technical cybersecurity incidents, fake media attacks are even more difficult to detect, either by people or technical means. In addition to promoting a strong fake media posture within the organization, brand managers might want to engage in the following to identify media activities as anomalous:

Engage in Social Listening: While this term can be confusing (the difference between social listening, buzz analysis, brand monitoring, social media intelligence, and social media

monitoring can be unclear), the underlying idea is straightforward. Monitor mentions of the brand, key people, organization, and campaign hashtags across platforms such as social media platforms, news websites, blogs, and forums to "get advance warning of any effort to spread disinformation or manipulated media about them" (Segal, 2021). Brand managers likely already do this but would benefit from focusing such activity on unexpected spikes in mentions potentially tied to synthetic media.

Outsource Media Forensics: In the cybersecurity context, most organizations know that they are out of their element when it comes to protecting their crown jewels – the most important data on their servers. This is true for much of the social listening mentioned above, too. However, the usual services do not specialize in fake media. Fortunately, start-ups around the world are starting to offer their synthetic media detection services. Firms like Sensity offer technical Forensic Deepfake Detection that "recognize[s] the latest AI-based media manipulation and synthesis techniques, including fake human faces in social media profiles, and realistic face swaps in videos" (Sensity, 2022). Other firms, like Sentinel, advertise their services in protecting companies from disinformation campaigns, synthetic media and information attacks by automatically authenticating digital media and checking if it is generated by artificial intelligence (Sentinel, 2022).

Gather Incident Indicators: Together, the assessment and detection phase should provide a technical and contextual understanding of fake media incidents. If an incident occurred, assess the conditions that led to the attack. Such an analysis might reveal the root causes of an incident and additional threat information, such as the TTPs (Tactics, Techniques, and Procedures) used by the attacker. Such analytical insights can help improve an organization's ability to detect future attacks earlier and strengthen its overall resilience to future fake media events.

4. Containment, Eradication & Recovery Phase

In the cybersecurity context, the goal of containment is to "prevent further damage and reduce the immediate impact of the incident by removing the adversary's access" (CISA, 2021, p.14). Eradication and recovery refer to a "return to normal" once the adversarial activity is contained. In the case of synthetic media attacks, these activities are difficult as there is no "intrusion" into organizational information systems and returning to normal is not a technical process with which all traces of the malicious event can be removed. Instead, the Containment, Eradication & Recovery phase demands much more public activities, which may require the manager to consider engaging in the following practices:

Remove Synthetic Media: Companies are well advised to work with their legal counsel and, working with social media platforms, have the synthetic media removed. The Kardashian deepfake was removed from YouTube through a copyright claim by the original publisher, but this process is arguably only "available for the privileged few" (Katz, 2019). However, containing a public attack is complicated. Once synthetic media becomes increasingly shared, it is exceedingly difficult to remove from circulation. For instance, the Kardashian deepfake still exists on other social media platforms. Complete eradication is likely not possible.

Rebut through Authentic Media: Work during the preparation phase should include incident response workflows and escalation procedures that help with containment. Legal counsel has suggested that such procedures should involve firms turning directly to their customers, partners, the media, and the public to let them know that they have fallen victim to a synthetic

media attack (Segal, 2021). Responding with legitimate content allows organizations to set the record straight, so to speak.

Report: In accordance with the escalation policies, if there is a risk of a material impact, the authorities (e.g., police or FBI) or regulators (e.g., the Security and Exchange Commission) might need to be alerted and involved in the containment phase. While these are generally good steps to take, they are not without risk. Depending on the type of organization (or brand) and its crown jewels, such responses might further incentivize the enemy to orchestrate further attacks. When Scarlett Johansson initially tried to fight deepfakes portraying her in adult videos, she found herself featured in more and more of such videos. After all, the attackers already have her 'personal decoder' with which they can create additional videos. Eventually, Johansson realized that "to protect yourself from the internet and its depravity is basically a lost cause, for the most part." (Harwell, 2018).

5. Post-Incident Activities Phase

The goal of this phase is to document the incident and to collect and apply lessons learned to improve the handling of future synthetic media incidents, which we suggest may include the following activities:

Log Incident Indicators: At this stage, organizations need to collect and record as much pertinent data as possible. These data can be related to the previous steps and to the actual incident. Any impact on reputation, economic well-being, corporate secrets, intellectual property, security, and the ability to govern and carry out daily operations needs to be logged in detail.

Perform Hotwash: Review the effectiveness and efficiency of the previous phases. Evaluate all the activities in each phase and capture what worked and what did not. If possible, identify root causes of problems and revise the step (e.g., the escalation policy and training manual may need to be adjusted if people did not know how to respond to the attack). Add, remove, revise, and realign policies. Update roles and responsibilities to improve the organization's overall resilience to synthetic media incidents.

6. Coordination Phase

This last phase is critical for improving everyone's ability to respond effectively to synthetic media attacks. In a cybersecurity context, it is well known that companies often suffer from repeat incidents (e.g., 80% of ransomware victims are targets for repeat attacks) (Sganga & Bidar, 2021). During this last phase, coordinating and sharing lessons with important stakeholders (e.g., supply chain partners, investors, cybersecurity firms) will increase the odds that everyone is better prepared for the next attack.

Share: Sharing synthetic media threats, lessons learned, and best practices will help improve the Synthetic Media Incident Response Playbook. Illustrating this point is a significant development in the resource sector. Goldcorp Inc. of Vancouver had been the victim of a ransomware attack in 2016. After the breach, Goldcorp invited the whole industry to talk about their experience. It turned out that every organization suffered from a miscalibration of subjective probabilities. Nobody thought that their mining company would be targeted. The breach was a wake-up call, and today, various mining companies participate in a threat-sharing centre they created collectively (Solomon, 2017).

Conclusion

Although humans have long employed manipulation to create a sense of falsity – from analog to digital techniques – falsity has arguably never been as hard to discern as it has become with the recent emergence of synthetic media manipulation. With the increasing quantity and quality of synthetic media, we find ourselves exposed to a new type of imitation game. We are particularly susceptive to believing synthetic media due to its ability to portray phenomena in a seemingly veridical manner through a high degree of realism. Even though such perceptions are in fact nonveridical, we are primed to believe particularly through our reliance on heuristics. Synthetic content which is information-rich (i.e., uses video and audio), familiar to us, or aligns with our existing beliefs may make us prone to processing the content in a non-systematic fashion, and subsequently believe it, as we have little reason to question such content.

Criminals increasingly exploit such cognitive biases, and by weaponizing synthetic media they can leverage our own vulnerabilities to create deception. Malicious synthetic media attacks create unique risks for organizations, such as risks to their reputation, economic and corporate secrecy, security, intellectual property, and governance and operations.

To brace themselves for these risks evolving into actual incidents, we recommend that managers learn from the cybersecurity context and adopt a Synthetic Media Incident Response Playbook. This Playbook outlines six phases an organization should consider before, during, and after a synthetic media incident. Firstly, before an incident occurs, organizations should *prepare* by conducting organizational education, establishing a fake media posture, appointing and training incident response personnel, devising escalation policies, engaging in penetration testing, and taking proactive steps to protect organizational assets. Organizations should then *assess* their threat environment by understanding what might be attacked (e.g., the reputation of a CEO or other brand representative), why and how an attack may occur, and who may initiate it. When a synthetic media incident occurs, organizations should engage in *detection & analysis* by conducting social listening, outsourcing media forensics, and gathering incident indicators to reveal the root causes of the event. Next, *containment, eradication, & recovery* can begin, which involves the organization working with legal counsel and social media platforms to remove the synthetic media content, actively rebutting the synthetic content with authentic and legitimate media, and, when necessary, reporting the incident to authorities or regulators. Following a synthetic media incident, organizations should engage in *post-incident activities* which include logging incident indicators (collecting all data on the impact of the incident) and performing a hotwash by reviewing the effectiveness of the previous phases and updating policies which align with lessons learned from the incident. Lastly, an organization should *coordinate* with other likely organizational targets by sharing their experiences with dealing with the synthetic media incident. This alerts other organizations to the current threat environment, key lessons, and general practices which they may be yet to consider due to underestimating the likelihood of an incident. Crucially, this also improves the threat resilience of their industry in general.

We hope that this Playbook, in addition to identifying key cognitive mechanisms which influence our appraisal of synthetic media, provides valuable insights for managers to aid them in bracing their organizations for this new age of synthetic falsity.

References

Beaudoin, C. E., & Thorson, E. (2005). Credibility perceptions of news coverage of ethnic groups: The predictive roles of race and news use. *The Howard Journal of Communications, 16*(1), 33-48.

Bonfanti, M.E. (2020, July 14). The weaponisation of synthetic media: what threat does this pose to national security? https://www.realinstitutoelcano.org/en/analyses/the-weaponisation-of-synthetic-media-what-threat-does-this-pose-to-national-security/

Borak, M. (2021, March 31). Chinese government-run facial recognition system hacked by tax fraudsters: report. *South China Morning Post*. https://www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax

Campbell, C., Plangger, K., Sands, S., & Kietzmann, J. (2021). Preparing for an era of deepfakes and AI-generated ads: A framework for understanding responses to manipulated advertising. *Journal of Advertising, 51*(1), 22-38.

Cho, C. H., Phillips, J. R., Hageman, A. M., & Patten, D. M. (2009). Media richness, user trust, and perceptions of corporate social responsibility: An experimental investigation of visual web site disclosures. *Accounting, Auditing & Accountability Journal, 22*(6), 933-952.

CISA. (2021). CISA releases incident and vulnerability response playbooks to strengthen cybersecurity for federal civilian agencies. https://www.cisa.gov/news/2021/11/16/cisa-releases-incident-and-vulnerability-response-playbooks-strengthen

Cole, S. (2019, June 12). This deepfake of Mark Zuckerberg tests Facebook's fake video policies. *Vice*. https://www.vice.com/en/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy

Daft, R. L., & Lengel, R. H. (1984). Information richness: A new approach to managerial behavior and organizational design. *Research in Organizational Behavior, 6,* 191–233.

Davis, V. (2021, December 06). Deepfakes to become a growing trend in 2022 says IntSights. *Cyber Magazine*. https://cybermagazine.com/cyber-security/deepfakes-become-growing-trend-2022-says-intsights

Dowdeswell, T. L., & Goltz, N. (2020). The clash of empires: regulating technological threats to civil society. *Information & Communications Technology Law, 29*(2), 194-217.

Duke Today. (2020, August 5). Navigating fake news: How Americans should deal with misinformation online. https://polisci.duke.edu/news/navigating-fake-news-how-americans-should-deal-misinformation-online

Fabe, M. (2014). Expressionism and Realism in Film Form: F.W. Murnau's The Last Laugh and Charles Chaplin's The Adventurer. In *Closely Watched Films: An Introduction to the Art of Narrative Film Technique* (1st ed., pp. 37–58). University of California Press.

French, R. M. (2000). The Turing Test: the first 50 years. *Trends in Cognitive Sciences, 4*(3), 115-122.

Gadde, R., Feng, Q., & Martinez, A. M. (2021). Detail Me More: Improving GAN's Photo-Realism of Complex Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13950-13959).

Gow, G. (2021, May 02). The scary truth behind the FBI warning: deepfake fraud is here and it's serious—we are not prepared for an attack. *Forbes*. https://www.forbes.com/sites/glenngow/2021/05/02/the-scary-truth-behind-the-fbi-warning-deepfake-fraud-is-here-and-its-serious-we-are-not-prepared/?sh=2d8abebb3179

Gralla, P. (2019, September 24). Here's how deepfakes can harm your enterprise — and what to do about them. *Symantec Enterprise Blogs*. https://symantec-enterprise-blogs.security.com/blogs/feature-stories/heres-how-deepfakes-can-harm-your-enterprise-and-what-do-about-them

Harwell, D. (2019, September 04). An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft. *The Washington Post.* https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/

Harwell, D. (2018, December 31). Scarlett Johansson on fake AI-generated sex videos: 'Nothing can stop someone from cutting and pasting my image'. *The Washington Post*. https://www.washingtonpost.com/technology/2018/12/31/scarlett-johansson-fake-ai-generated-sex-videos-nothing-can-stop-someone-cutting-pasting-my-image/

Holzer, H. (1979). How the printmakers saw Lincoln: not-so-honest portraits of "Honest Abe". *Winterthur Portfolio*, *14*(2), 143-170.

Katz, M. (2019, June 17). Kim Kardashian can get a deepfake taken off YouTube. It's much harder for you. *Digital Trends*. https://www.digitaltrends.com/social-media/kim-kardashian-deepfake-removed-from-youtube/

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons, 63*(2), 135-146.

Landay, L. (2018). The Moviola and Other Analog Film Editing Machines. In *The Routledge Companion to Media Technology and Obsolescence* (pp. 136-147). Routledge.

Lang, A. (1995). Defining audio/video redundancy from a limited-capacity information processing perspective. *Communication Research, 22*(1), 86-115.

Lee, H., Park, S. A., Lee, Y., & Cameron, G. T. (2010). Assessment of motion media on believability and credibility: An exploratory study. *Public Relations Review, 36*(3), 310-312.

Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof, 23*(3), 255-262.

Melville, K. (2019, August 30). The insidious rise of deepfake porn videos — and one woman who won't be silenced. *ABC News*. https://www.abc.net.au/news/2019-08-30/deepfake-revenge-porn-noelle-martin-story-of-image-based-abuse/11437774

Miller, J.R. (2022, March 17). Deepfake video of Zelensky telling Ukrainians to surrender removed from social platforms. *The New York Post*. https://nypost.com/2022/03/17/deepfake-video-shows-volodymyr-zelensky-telling-ukrainians-to-surrender/

Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press, USA.

NCSC. (2017). Penetration testing. https://www.ncsc.gov.uk/guidance/penetration-testing

Nickerson, R.S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Novak, M. (2019, August 08). make Jordan Peterson say anything you want with this spooky audio generator. *Gizmodo*. https://gizmodo.com/make-jordan-peterson-say-anything-you-want-with-this-sp-1837306431

OpenAI. (2021, January 5). DALL·E: Creating images from text. https://openai.com/blog/dall-e/

Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers' Imprint,* 20(4), 1-16.

Romero, A. (2021, May 24). A complete overview of GPT-3 — The largest neural network ever created. https://towardsdatascience.com/gpt-3-a-complete-overview-190232eb25fd

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179.

Saito, Y., Takamichi, S., & Saruwatari, H. (2017). Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 84-96.

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services, 64*, 102771.

Schwartz, R. (2016). Perceptual veridicality. *Philosophical Topics, 44*(2), 381-403.

Segal, E. (2021, January 11). Deepfakes: 7 ways to guard against this new form of disinformation. *Forbes*. https://www.forbes.com/sites/edwardsegal/2021/01/11/deepfakes-7-ways-to-guard-against-this-new-form-of-disinformation/?sh=693855933c80

Sensity. (2022). Forensic deepfake detection. https://sensity.ai/deepfakes-detection/

Sentinel. (2022). Defending against deepfakes and information warfare. https://thesentinel.ai/

Sganga, N. & Bidar, M. (2021, June 17). 80% of ransomware victims suffer repeat attacks, according to new report. *CBS News*. https://www.cbsnews.com/news/ransomware-victims-suffer-repeat-attacks-new-report/

Solomon, H. (2017). Canadian cyber attack led to new mining industry threat sharing centre. https://www.itworldcanada.com/article/canadian-cyber-attack-led-to-new-mining-industry-threat-sharing-centre/393850

Steinmetz, K, (2018, August 9). How your brain tricks you into believing fake news. *Time*. https://time.com/5362183/the-real-fake-news-crisis/

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

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps?. *Journal of Computer-Mediated Communication*, *26*(6), 301-319.

Swerzenski, J. D. (2021). Fact, fiction or Photoshop: Building awareness of visual manipulation through image editing software. *Journal of Visual Literacy*, *40*(2), 104-124.

Tissler, J. (2021, April 28). Deepfakes as a security risk: What's behind it?. *DSwiss.* https://www.dswiss.com/en/news/deepfakes-as-a-security-risk-whats-behind-it

Turing, A. M., (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Ucbasaran, D., Westhead, P., Wright, M., & Flores, M. (2010). The nature of entrepreneurial experience, business failure and comparative optimism. *Journal of Business Venturing, 25*(6), 541-555.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1), 1-13.

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 613-621.

Wanta, W., & Hu, Y. W. (1994). The effects of credibility, reliance, and exposure on media agenda-setting: A path analysis model. *Journalism Quarterly*, *71*(1), 90-98.

Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). "All around me are synthetic faces": the mad world of AI-generated media. *IT Professional, 22*(5), 90-99.

Woollacott, E. (2022, January 27). New standard aims to protect against deepfakes. *Forbes.* https://www.forbes.com/sites/emmawoollacott/2022/01/27/new-standard-aims-to-protect-against-deepfakes/?sh=425da4f7265a

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2223-2232).