kyndryl

A New Approach to Proactive Deepfake Detection & Remediation

Created by Dean Pratt and Dr. Elena Huff

Contents

78 Billion Reasons for a Proactive Strategy

"In one of the largest known cases of deepfake fraud this year, a Hong Kong finance worker was duped into transferring more than $25 million to fraudsters using deepfake technology who disguised themselves as colleagues on a video call, authorities told local media in February."[1]

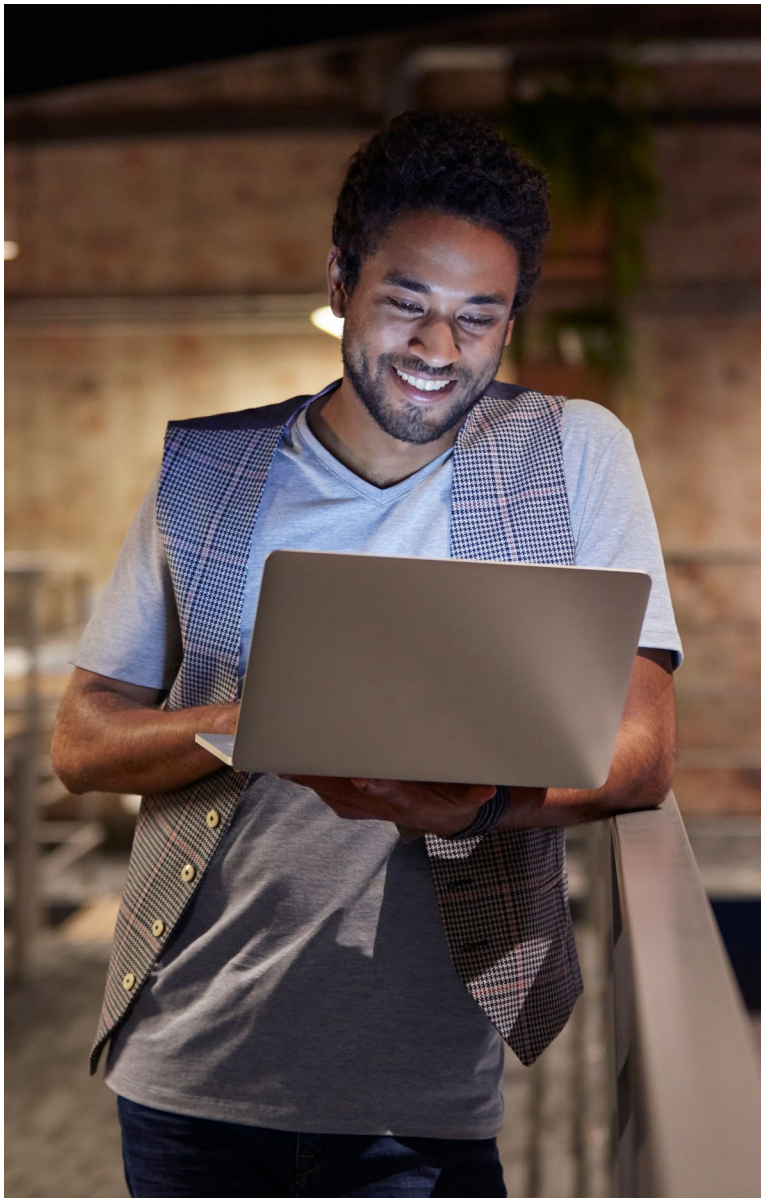
"An economic study by Tel Aviv, Israel-based cybersecurity firm CHEQ and the University of Baltimore have revealed that fake news is costing the global economy $78 billion each year."[2] *Though we often think of deepfaking a personality as something that takes a team, in the very near future, it will only take a careless thought. Opensource models with advanced audio/visual generative capabilities have become easy to access and remain unbound by the responsible use policies most foundational model creators put in place. We recently met with the CISO of one of the largest rental car companies, and their internal team created a deepfake video of their CEO to demonstrate the urgency of this issue. These malicious uses will soon morph into bad actors puppeteering a deepfake avatar on a call and even intelligent AI versions that can respond in real- time with highly nefarious content. This makes it very hard to know who you are really talking to or, what the malicious AI's intentions are. On that precipice, the floodgates will open to a whole new way of looking at the world and learning to distinguish fact from fiction. Soon, we may eventually feel the visual information we used to be able to trust without a second "thought is now, unknowable...so what does that mean for us as a society globally?*

"Apart from the obvious implications for misinformation and propaganda during times of conflict, national security challenges associated with deepfakes manifest in threats to the U.S. Government, NSS, the DIB, critical infrastructure organizations, and others."[3]

From wars to stock manipulation, what won't happen or at least be attempted with deepfakes? The only real inhibitor is the lack of access to the technology or, the creative thinking needed for ways it could successfully be used, both of which are intrinsically ephemeral. Intelligence agencies from around the world constantly work to keep threats like this from making a real impact, and for good reason. Simply put, the president of a country or the leader of a military force, giving orders for something on a video call that they never would approve, could unwittingly lead to severe and damaging outcomes if we are not prepared.

Well, How Should I Know?

Currently, Generative video or image AI models commonly utilize one or a combination of two generalized routes, emulation or simulation. Emulation refers to the model reproducing and predicting based off a text prompt and other video/image data it has been trained on. Diffusion Models, Conditional Generative Adversarial Networks (cGANs), Autoencoders, Landmarks, and 3-D Morphable Models are a few examples of this. Simulation refers to the iterations of neural rendering models such as Physics Informed Neural Networks or PINNs[4] and General World Models or GWMs.[5] These models are trained on environmental physics to simulate the real world instead of only attempting to predictively copy it. There are many different methods for doing this. We have quickly entered advancements in spatiotemporal consistency by incorporating hybrid architectures that utilize both Diffusion[6], GANs[7] and PINNs. GWMs like SORA from OpenAI are also making improvements[8] with variational autoencoders (VAEs)[9] and Recurrent Neural Networks (RNNs)[10] in tandem, creating a vast future for hybrid architectures. Combined approaches improve physics learning and generation, greatly reducing morphing and artifacts commonly found in generated content previously.



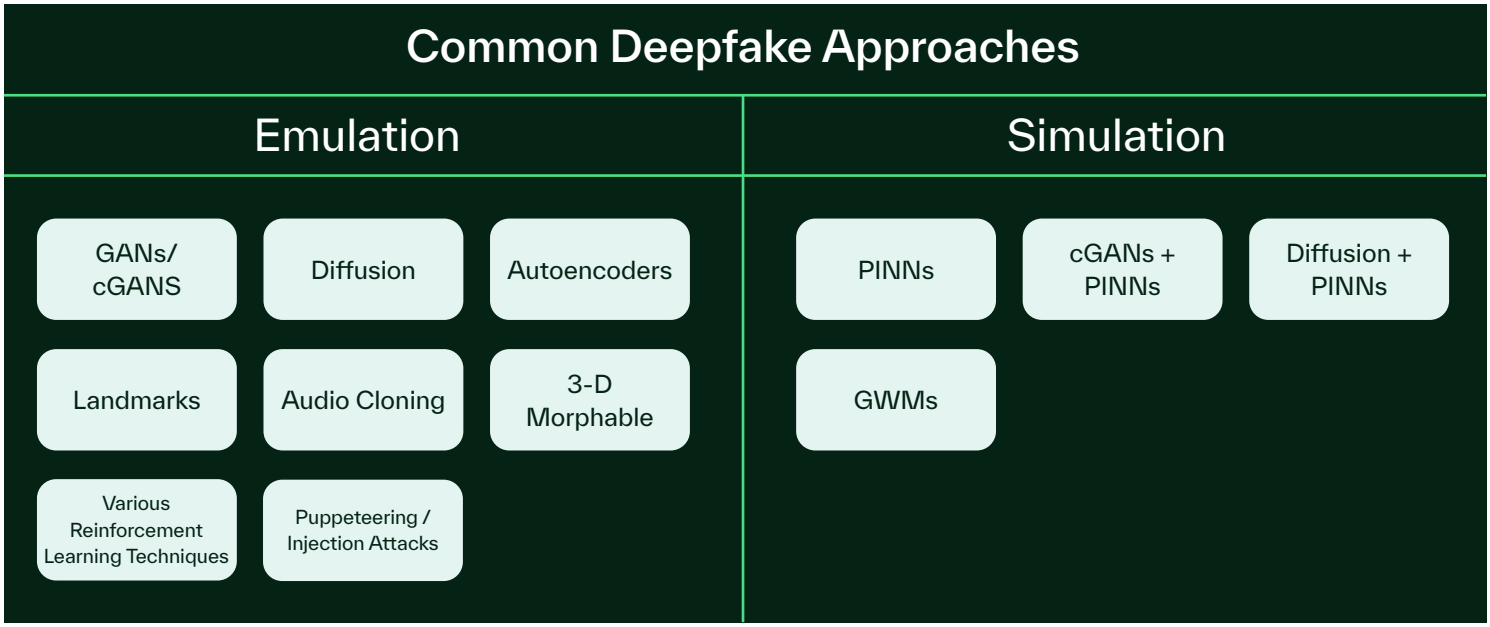


Fig. 1 - Diagram of Emulation and Simulation methods

Companies like NVIDIA AI are working hard to create AI that accurately models our physical world in various applications, from weather predictions to designing factories.[11] This approach also revolutionizes the means and magnitude of efficiencies for creating viable digital twin models. In light of self-rewarding or self-training algorithmic advancements, predicting the future soon becomes a very blurry unknown.

This topic gets topically tricky around the consistent back and forth of preventive threat measures and the opposing side's response. Often referred to as the Cat and Mouse Game[12], we are locked and engaged in a constant battle, an eternal race to see who can beat the other first. As methods for recognizing deepfakes become more intuitive, so will the generative models that produce the deepfake content. Satisfyingly accurate Facial Alignment Networks (FANs)[13] and other overlay models will arise that will take easily fake realistic content and, add image or metadata to convince even the most advanced scanners. Other methods outside of the software domain must be considered, too. We must remain creative when thinking about physical retinal scanners and close-proximity hardware injection attacks. Being an advocate of personal rights and data privacy, this line is a tough one to walk. When we think about Multi Factor Authentication (MFA)[14] and the future, what does that look like and what private data would need to be utilized in order to actually validate what is real?

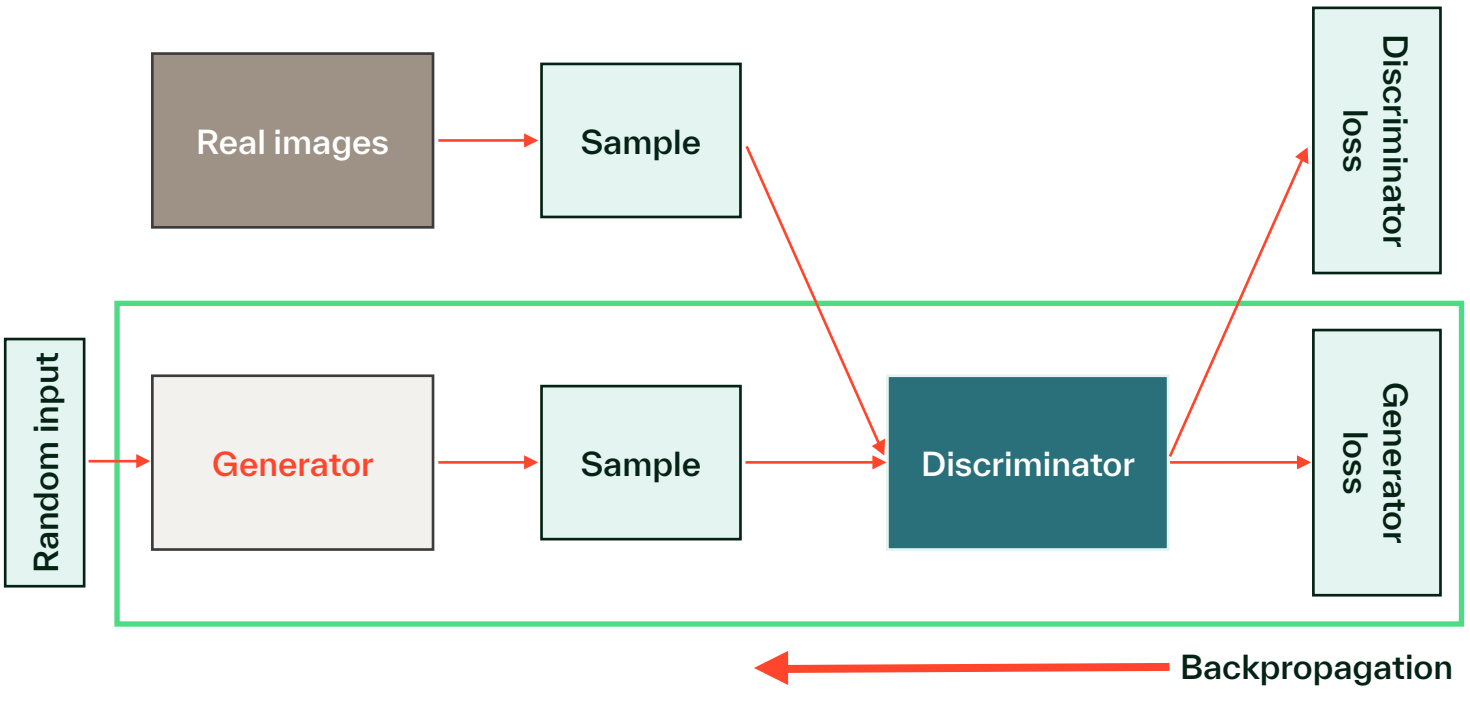


Fig. 2 - Generative Adversarial Network Backpropragation architecture diagram

Taking a Closer Look

Below are some of the most popular known technologies presently used to detect deepfake content:

1. Physiological scanning for unique, organically human functions is not yet replicated accurately by deepfake models, such as skin color augmentations due to blood flow or specific voice characteristics. This technique may be fooled when AI models replicate organic behaviors better, if video resolution is low or, if the red spectrum of color is altered as a filter.
2. Watermarking original content is a popular approach; however, even the latest techniques from leaders in the industry can be altered and made obsolete. For example, University of Maryland engineers pointed out vulnerabilities in Google's new watermarking tool called SynthID shortly after its release, which employs two neural networks and image embedding to help maintain consistency when the video content is altered.
3. Scanning for javascript injection, virtual camera and hardware video capturer events, which seek to provide a point of entry for deepfake attacks. This method only addresses bad actors, infiltration and breaches, not cases where the content is shared by authenticated users.
4. Pattern matching in algorithmic output or "Fingerprinting", as certain AI models may tend to generate more predictable, symmetrical or replicated pixel formations.
5. 3-D Convolutional Neural Networks (3-D CNNs), Temporal Facial Patterns (TFPs) and Spatiotemporal Inconsistency Learning (STIL) models are designed to detect slight incongruences in object location, physics and object behavior is recognized as well, since video and image models can lose spatio-temporal consistency from frame to frame.

These leading approaches are promising however, there are a few new solutions that can be added to the list. Introducing a cognitive layer consisting of two different AI models, that analyze the message of the content and the impact it could have, adds viability to any approach. The architecture diagram found further below, will help to clarify the proposed models purpose and overall solution workflow.

- The first model is called an Intuitive Interest Check (IIC). It is an AI model architected to check the content in question, using existing company resources and data, as well as domain specific, prefilled questionnaires about personal motivations and interests for each person of interest and the company as a whole.
- The second AI model is a Cognitive Impact Analysis (CIA) model. This model is built to accurately project the potential impact the deepfake content will have on people who are exposed to it, as well as the business' reputation and operations. It provides a summary of what was said, the potential negative and positive effects and is grounded with a predefined set of social models aligned to specific industry or business domains.

Putting it All Together

"As deepfake attacks grow it is critical for organizations to take proactive action in safeguarding their environment."[15]

"They should consider implementing a number of technologies to detect deepfakes and determine media provenance, including real-time verification capabilities, passive detection techniques, and protection of high priority officers and their communications."[16]

So, we have discussed some tools and approaches but, what does the architecture look like? How do you put together a framework that addresses the points made by the NSA, FBI and CISA concerning deepfake defense? Well to start, by using web scrapers, RSS feeds from websites, scripted functions and other alternatives, we can proactively identify malicious deepfake content before it has a chance to make a widespread impact on a company or, reputational quality. In the coming year, the capabilities of General World Models and Generative AI video models as a whole, will bring new threat vectors we must be well prepared for, along with all the benefits they may bring.

Currently, due to legal terms, this architecture will only involve proactively scanning internal systems. We have also included a list of popular external sites at the end with links to their terms and conditions which do not allow scanning of their platform data. It will be interesting to see what the future unravels in regard to the proliferation of deepfake content and individuals or companies being able to protect themselves effectively.
Let us first review the places your company can identify as a threat vector and examples of a scanning strategy. Though not detailed in the solution architecture diagram (Fig. 3), scanning for javascript and virtual hardware injection attacks should be implemented at this first step as well.

Deep Fake Remediation

High-Level Architecture

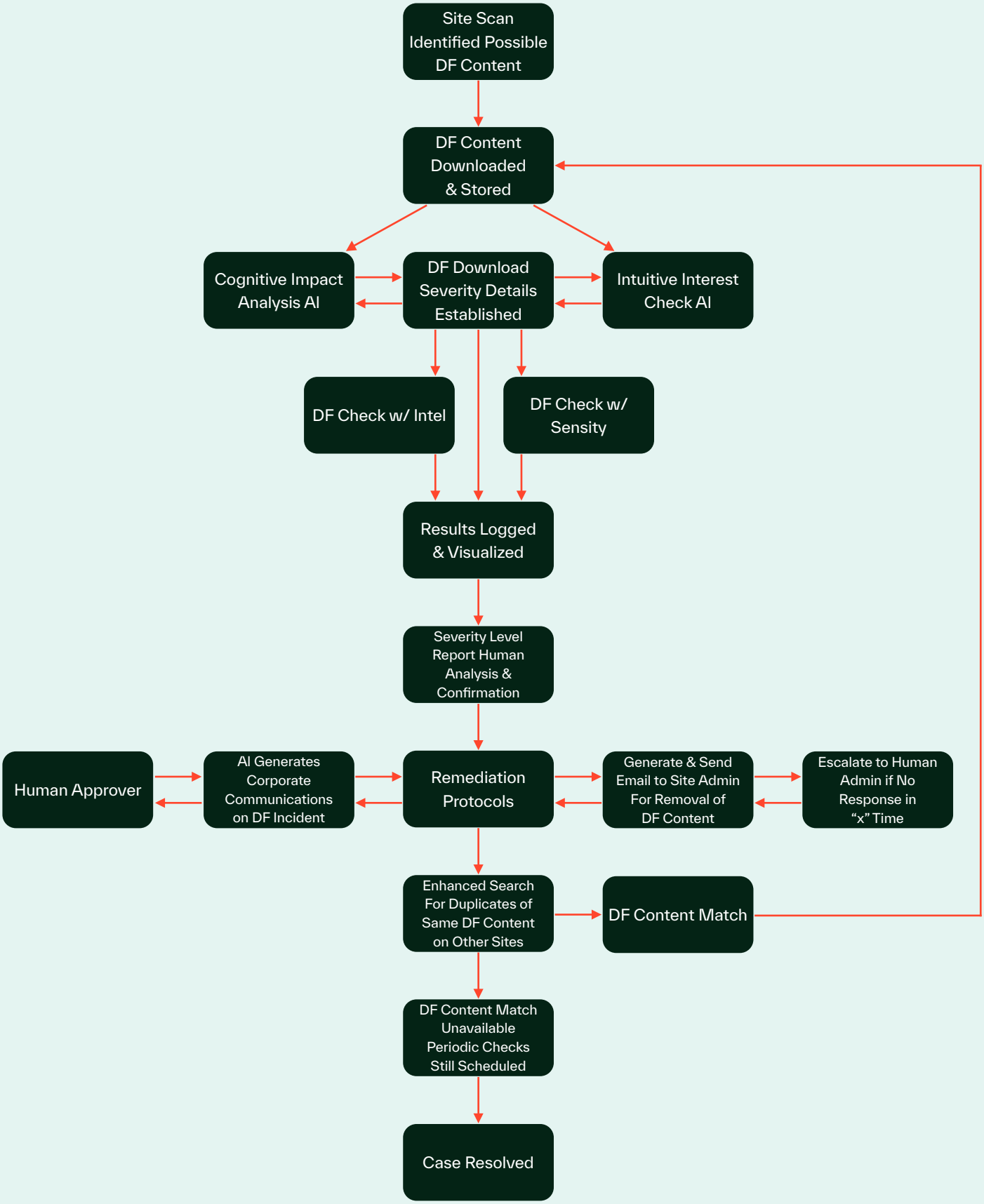


Fig. 3- Advanced deepfake detection and remediation architecture diagram

Internal Sources to Scan:

- Internally accessible data sources (SharePoint, Google Drive, etc.)
- Company messaging and chat applications (Microsoft Teams, Viva Engage (Yammer), Google Chat, Slack, Blue Jeans, etc.)
- Company Video and Phone Applications Messaging, Chat, Recordings, Transcripts and Live Streams (Teams, Webex, Google Meet, Zoom, etc.)
- Company Devices and Local Storage Devices
- Email Server
- Enterprise Application Databases

Internal Entity Scan Scheduling:

- Every 5 Minutes: CEO, CIO, COO, etc.
- Daily: VPs, high-ranking leadership
- Monthly: Management / General Org
- Ex. "CEO, [Person's Name], [Company Name]" or "[Person's Name], [Person's Name], [Person's Name], [Company Name], News, Announcements, Updates, Interviews"

Internal Topic of Interest Scheduling:

- Hourly: Highly Sensitive Information
- Daily: Current Events, Mature or Explicit Content, Political Content
- Ex. "[Country], War, [Company Name]" or "Stock, Wall Street, [Company Name]"

As seen in Fig. 3, after a scan identifies and downloads the media content that contains a person or topic of interest, it sends it through a multi-layered path of different models. The first two models that receive the data in question are the Intuitive Interest Check (IIC) and Cognitive Impact Analysis (CIA) models. They will then analyze the content for its message, meaning and impact potential. A severity level will then be established by each model, if the content is deemed irrelevant (deepfake of Donald Duck making a joke about a named entity or topic), the source is then sent to a quarantine for a human approver to review at a later date. If the content IS deemed a severity level, it is then sent to the other solutions in the market, hosted in your tenancy.

In the architecture diagram, you will see Intel and Sensity's offerings as an example but, more than just two is recommended, as market availability varies. Pragmatically, Intel has not yet made their offering publicly available at this time so, you may start with alternative options. The north star here, is having an architecture that easily can add new components in the workflow must remain a prime priority.

After the other third-party or internally developed models have analyzed the data and determined a severity level, the results from all of the models are sent to a dashboard for a human to review. A real-time alert system should be in place for certain severity levels, autonomous actions may be acceptable for lower severity levels.

Advanced Remediation

Now that the content has a severity level, predetermined remediation functions can be triggered. Two more AI models are involved in this framework at this part of the process. One AI model will take the service incident description details and generate multiple versions of a corporate communications email. These will be sent to the internal Communications and Public Relations parties responsible, to review and send to the company after any modifications.

The other AI model will generate an email to a site admin or responsible hosting party, asking them to immediately remove the content and other necessary legal details. The email must be reviewed and sent by a human in near real-time. This part of the process can be automated and include a disclaimer that the message was generated by AI protocols for deepfake detection and remediation, in case of error. For smaller companies, this may be a more viable path.

If no response is received within "X" time, additional actions must be autonomously engaged, according to the incident severity level and projected content impact. The scanning framework must also run an expanded search on the identified deepfake content and, repeat the process if necessary. Like many things in life, if there is one version, there are most likely more. If no more content is detected, the scan template is saved and run monthly until it is deemed not a threat.

Collaboration Across Industries

The fight against deepfakes is a complex, multifaceted problem that spans across multiple sectors. To effectively combat deepfakes, collaboration between media, finance, cybersecurity, law enforcement, and government agencies is essential. Each industry brings unique expertise and capabilities that, when combined, can lead to more comprehensive solutions for detecting, mitigating, and remediating the growing deepfake threat.

1. **Telecommunications, Media & Entertainment, and Technology (TMT): Protecting Credibility and Trust**
   In the TMT industry, deepfakes pose a direct threat to public trust and credibility. News organizations are particularly vulnerable to deepfakes, as altered video or audio of public figures can easily spread misinformation and cause widespread confusion. Media companies can play a critical role by working with AI developers and tech companies to create algorithms that detect deepfakes before they are disseminated on social media and news platforms.

   By collaborating with tech companies, the media industry can ensure that news and information shared with the public remains accurate and free of manipulation.

2. **Finance Industry: Preventing Fraud and Economic Disruption**
   In finance, the rise of deepfakes presents a growing risk of fraud. Deepfake audio or video of CEOs, high-level executives, or government officials can manipulate stock prices, trick employees into transferring funds, or create false market rumors. The finance industry must work closely with cybersecurity firms to develop systems that verify the authenticity of communications, particularly in situations where significant financial transactions or market-sensitive information is involved.

   In the above-mentioned example, in 2020, a bank in Hong Kong was targeted by deepfake fraud that led to over $25 million loss. As these incidents increase, financial institutions are partnering with AI experts to create authentication systems that verify facial recognition, voice patterns, and even behavioral biometrics to prevent fraud.

3. **Cybersecurity Industry: Building the Infrastructure for Detection**
   The cybersecurity industry plays a foundational role in the battle against deepfakes. It develops the tools and systems necessary for real-time detection and remediation. This industry's focus on threat intelligence and mitigation has expanded to include deepfake detection, with companies like **DeepMind** and **Sensity AI** leading the charge in developing deepfake detection algorithms.

   By collaborating with law enforcement and regulatory agencies, cybersecurity companies can help develop standards and protocols for managing deepfakes on a global scale. The challenge is not only detecting deepfakes but also ensuring secure data transmission and storage, which can be exploited by deepfake creators to distribute manipulated content.

4. **Law Enforcement: Investigating and Prosecuting Deepfake Crimes**
   Law enforcement agencies need to work closely with technology firms and cybersecurity experts to investigate deepfake-related crimes. Whether it's the creation of fake pornographic content, identity theft, or financial fraud, law enforcement is often at the frontline of addressing the consequences of deepfake technology.

One challenge is that legal frameworks have not yet caught up with the rapid development of deepfake technologies. As a result, law enforcement agencies need to work with legislators and tech companies to develop laws that specifically address deepfake creation and use. International collaboration is particularly important, given the borderless nature of digital crimes. The U.S. Department of Homeland Security has already begun collaborating with private sector partners to develop a deepfake detection challenge, where companies are encouraged to create innovative solutions for detecting and combating deepfakes.[17] This type of cross-sector collaboration is critical to staying ahead of deepfake creators, who continue to refine their techniques.

5. **Government and Regulation: Setting Global Standards**
Governments must collaborate across industries to set global standards for deepfake detection and prosecution. This includes crafting legislation that defines the use and misuse of deepfakes, imposing penalties for their malicious use, and encouraging transparency in content creation. For example, the European Union's Artificial Intelligence Act includes provisions aimed at regulating AI technologies that can create deepfakes, ensuring ethical use, and preventing widespread harm.[18]

   In the U.S., state-level laws are emerging, such as California's law prohibiting the use of deepfakes in election interference.[19] However, international agreements and frameworks are needed to regulate how these laws are enforced across borders. Regulatory bodies must also work with industries like media and finance to create trusted third-party verification systems for deepfake detection.

6. **Unified Deepfake Detection Standards**
The ultimate goal of cross-industry collaboration is to develop unified standards and frameworks for deepfake detection and response. This includes the creation of industry-wide protocols for sharing intelligence on deepfake threats, real-time detection tools, and response strategies that ensure quick remediation. The creation of shared datasets that can be used to train deepfake detection algorithms will be critical in staying ahead of evolving threats.

   Multi-layered detection systems combine multiple approaches to identify and mitigate deepfake threats, increasing the reliability and accuracy of detection. By using different layers of technology, such as AI-driven pattern recognition, blockchain-based content verification, biometric analysis, and watermarking, these systems create a comprehensive defense. Each layer targets different aspects of the deepfake, from visual inconsistencies to manipulation of audio or metadata. For example, AI might detect subtle abnormalities in facial movements, while blockchain ensures content authenticity by verifying its source and integrity. This multi-layered approach strengthens the detection process, making it harder for malicious actors to bypass all defenses and ensuring that deepfakes are identified before they cause significant harm. By integrating multiple layers of detection, organizations can significantly reduce the risk of false positives and false negatives, providing a more robust response to evolving deepfake technologies.

   Cross-industry collaboration plays a crucial role in enhancing multi-layered detection systems by allowing different sectors to contribute their expertise and resources. By working together, these industries can share insights and develop unified standards that strengthen each layer of detection, ensuring that deepfake threats are identified and mitigated more effectively. For instance, cybersecurity firms can enhance AI algorithms, while media companies focus on content verification, and law enforcement assists with legal frameworks, resulting in a more comprehensive and resilient multi-layered detection system.

A Day in the Life: Real-World Deepfake Scenarios

Cross-industry collaboration is vital to staying ahead of deepfake threats, as each sector plays a unique role in detection and response but, what does this look like in practice? To illustrate the stakes and the need for such collaboration, here are a few scenarios where deepfakes could create serious challenges and demonstrate why proactive measures are crucial:

1. **Media Integrity:** A breaking news story airs, showing what appears to be a high-profile politician making controversial remarks. The video spreads rapidly, causing public outcry. Hours later, experts reveal it was a deepfake. While the truth surfaces, the damage to public trust is already done. This scenario highlights why news organizations must work closely with tech firms to verify content before it reaches audiences.

2. **Corporate Miscommunication:** During an important board meeting, executives receive a video message from their CEO announcing a strategic change. Decisions are immediately made based on this new information. However, the CEO was not involved; it was a deepfake designed to cause confusion. This example underscores the need to authenticate internal communications, particularly at the executive level.

3. **Workplace Phishing:** An employee gets a video message from their manager asking them to click a link to access an urgent report. Believing it is legitimate, they click, unknowingly triggering malware that spreads through the company's network. This scenario shows the evolving sophistication of deepfake social engineering attacks and the need for advanced detection systems and internal training programs to protect against such threats.

These scenarios demonstrate why cross-industry collaboration and advanced detection technologies are essential in managing the risks posed by deepfakes. If you are curious about your own ability to identify deepfakes, try this interactive quiz by the New York Times: Can You Tell What's Real?

In the Future

Here are a few additional possibilities that may show up in the future, of how we can authenticate content now and looking ahead:

- **Proactive Anti-Infiltrative Development:** A team of developers will use the leading open-source models for deepfake generation and, develop on top of them, verbosely documenting the code. The team will solve for challenges of spatiotemporal consistency, in a hybrid combination of physic generators or GWMs and, diffusion/GAN methods. This will provide a better list or scripting of what to look for when identifying fakes, while simultaneously utilizing other GA offerings for checking as well.

- **Telematic Verification:** Mobile device data can be cross referenced at any time, with last known biometric login or extensible passcode evidence, and location/timestamp of the video content in question.

- **Bio-signature Near-Field Authentication & Digital Token Embedding:** When a device begins recording, it first verifies user biometric inputs (thumbprint, voice, etc.) and communicates with the other digital devices in the area that the owner is ok with a recording occurring, all digital content without that token is considered fictitious and is tagged visibly as such.

- **Retinal-based unique ID:** Must be verified through connected device when person being recorded is recognized by facial matching software, i.e., Apple Watch notification when being recorded or new video content is uploaded, can be driven by content/social group relationships

- **Personal Profile Authentication:** Tonality, timbre, word choice, eye/nose/lips/facial movements, specific clothing details, body/hand movement profiles can all be created for an individual as a reference, can easily be augmented with location and biometric authentication data.

Conclusion and Recommendations

The battle against deepfakes is too large for any single sector to tackle on its own. Cross-industry collaboration between media companies, financial institutions, cybersecurity firms, law enforcement, and government agencies is essential to create a comprehensive and unified defense. By working together, these industries can not only create more effective deepfake detection systems but also establish global standards for the ethical use of AI and ensure the integrity of information in the digital age.

As the technology we depend on climbs in intelligence, so must we. It's not about what the day-to-day changes are now, it's about the six months to three-year impact and beyond. As humans, we cannot afford to ignore that the unprecedented rate of change does not just mean innovation, it also means equivocal orders of magnitude for organic life's safety and health. Truly, the amorphous superalignment issues we have with AI now do not only apply to training the model, but its uncalculated potentiated impact and unintended consequences as well.

As different techniques and multimodal AI capabilities unfalteringly advance, all anyone will need is a picture and a prompt, to create almost anything they can imagine, regardless of technical skills.[20] Deepfake technology is one of stellar reasons that with every inhale and exhale, our reality has shifted, with or without us approving. Likewise, with our every next step wading through the deep waters of change in unfaltering, vigilant empathy, we become our causality, our chance, and our change for the better. Cheers and happy innovating to a safer future!

Dean Pratt and Dr. Elena Huff

Helpful Links and Info

1. *"'Everyone looked real': multinational firm's Hong Kong office loses HK$200 million after scammers stage deepfake video meeting",* - South China Media Post, Feb 2024
2. *"Online fake news is costing us $78 billion globally each year", ZD Net, Dec 2018*
3. *"Contexualizing Deepfake Threats to Organizations", NSA, FBI, CSI, Sept 2023*
4. *"Physics-Informed Neural Networks", Science Direct, Feb 2019*
5. *"Introducing General World Models", RunwayML, Dec 2023*
6. *"Physics-Informed Diffusion Models" arxiv.org,* Mar 2024
7. *"A physics-informed GAN framework based on model-free data-driven computational mechanics",* Science Direct, May 2024
8. *"Video generation models as world simulators",* OpenAI, Feb 2024
9. *"Variational Autoencoders",* Wikipedia, Retrieved Sep 2024
10. *"Introducing Dreamer: Scalable Reinforcement Learning Using World Models",* Google Research, Mar 2020
11. *"Physics-Informed Machine Learning Platform NVIDIA Modulus Is Now Open Source",* NVIDIA Technical Blog, Mar 2023
12. *"The AI cat and mouse game has begun",* CIO, April 2024
13. *"FT-GAN: Face Transformation with Key Points Alignment for Pose-Invariant Face Recognition",* ResearchGate, July 2019
14. *"Will AI and Deepfakes Weaken Biometric MFA",* KnowBe4, Feb 2024
15. *"Why deepfakes are set to be one of 2024's biggest cyber security dangers",* TechRadar, July 2024
16. *"Contexualizing Deepfake Threats to Organizations",* NSA, FBI, CSI, Sept 2023
17. *"Increasing Threat of Deepfake Identities",* US Homeland Security, June 2019
18. *"EU AI Act: first regulation on artificial intelligence",* European Parliament, June 2024
19. *"Defending Democracy from Deepfake Deception Act of 2024",* US Senate Judiciary Committee, June 2024
20. *"Generative models: VAEs, GANs, diffusion, transformers, NeRFs",* TechTarget, July 2024

kyndryl

This document is current as of the initial date of publication and may be changed by Kyndryl at any time without notice. Not all offerings are available in every country in which Kyndryl operates. Kyndryl products and services are warranted according to the terms and conditions of the agreements under which they are provided.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. Kyndryl products and services are warranted according to the terms and conditions of the agreements under which they are provided.