Synthetic Reality & Deep Fakes: Considerations for Employers and Implications of the Rise of Deep Fakes in the Workplace

Authors:

Chase Perkins, Founder & CEO, Thoughtly

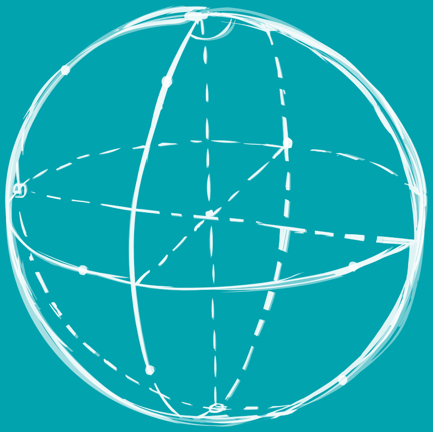
Natalie Pierce, Co-Chair, Robotics, AI & Automation Practice Group, Littler

Aaron Crews, Chief Data Analytics Officer, Littler

Jill Weimer, Shareholder, Littler

Littler®

Labor & Employment Law Solutions



In an age where computer generated imagery (CGI) and digital effects enable entire film genres to exist, like Marvel's superhero series the *Avengers* or *Guardians of the Galaxy*, audiences have no expectation that movies they consume depict actual events or reflect reality. It is therefore reasonable to assume that the context and forum of how digital media and information is communicated, observed and consumed informs our default expectations of it. Digital media, augmented reality social media filters, video games and salacious news stories are wildly successful and nearly omnipresent forms of entertainment. If we were to unwittingly rely on inauthentic digital media or false information in the course of our daily lives or during the course of business operations, however, we would find it both highly problematic and devoid of entertainment value.

The authors propose a tangible perspective on AI, one where the proliferation of *Deep Fakes* and the ability to leverage this technology may challenge standards for information dissemination, communication and our most basic assumptions of reality—where once seeing was believing. These inauthentic intrusions not only impact our society generally, and our political system and growing divisions more specifically, but also spill into our workplaces in a way that forces employers to grapple with the often inevitable effects. Employers will need to adjust to this new reality and understand the means of minimizing the potentially negative impact, including the utilization of data analytics to protect companies and their workforces from exploitative uses of false information.

From Humble Beginnings

Developments in various fields within artificial intelligence have allowed for remarkable advancements in different types of automation tools. At the core of artificial intelligence is the training of machines to learn through observation (*i.e.*, pattern recognition).[1]

What began as the narrow application of applied statistics to identify characteristics and traits of significance while analyzing data has evolved into a vast array of complex automation systems, including everything from machine vision systems used by self-driving cars, to voice-to-text systems utilized by Apple's Siri and natural language processing techniques leveraged by social media platforms.

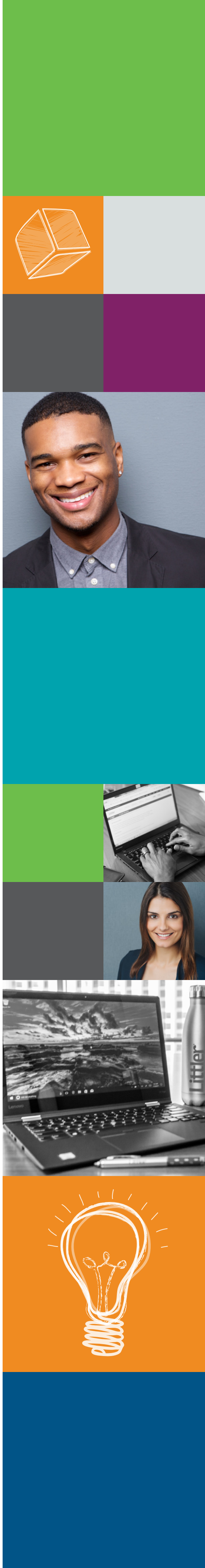
The approaches to building artificial intelligence systems can vary greatly, with a wide variety of potential learning models—from supervised learning (*i.e.*, human-curated training data), to unsupervised learning (*i.e.*, machine-collected and structured training data), to hybrid approaches like semi-supervised training and goal-oriented training like reinforcement learning.[2]

Emergence of Deep Fakes

Computer-generated and manipulated images and video have existed in various forms of sophistication for several decades. With the recent advent of generative deep-learning models, like variational autoencoders and generative adversarial networks (GANs), in which one neural network generates content (*e.g.*, realistic images of people), as a second adversarial neural network attempts to identify the computer-generated content as fabricated.[3] The results of this computational gamesmanship both enhances the GANs' training data and refines its ability to generate realistic content.

Moreover, computer-generated content is not limited to video, audio or graphical renderings, but can also apply to any domain for which learning systems have adequate training data, including natural language, art and musical composition.

OpenAI, a leading AI research institution, recently released both the source code and research findings for state-of-the-art machine-reading comprehension, real-time machine translation and coherent question-and-answering ability—all of which the machine can perform without predetermined or task-specific guidance.[4] State-of-the-art open-source code—which is accessible to anyone—such as provided by OpenAI, has demonstrated AI's ability to engage in unguided and realistic extemporaneous written dialogue and natural



[1] Wolfram Burgard, Bernhard Nebel, and Martin Riedmiller, *Foundations of Artificial Intelligence - 8. Machine Learning from Observations* (2011), http://ais.informatik.uni-freiburg.de/teaching/ss11/ki/slides/ai08_machine_learning_handout_4up.pdf.

[2] Yann LeCun, Yoshua Bengio and Geoffrey Hinton (2015), *Deep Learning*, Nature Review Insight (Vol. 521 May 28, 2015). *See* https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf.

[3] David Foster, *Generative Deep Learning* (2019). *See* https://www.oreilly.com/library/view/generative-deep-learning/9781492041931.

[4] *Better Language Models and Their Implications*, OpenAI (February 19, 2019), *See* https://blog.openai.com/better-language-models.

language content generation. Moreover, Google Duplex in 2018 unveiled the ability to engage in unscripted, realistic, real-time voice dialogue between machine and human, without the humans realizing they were speaking to a machine.[5] Imagine the impact this could have on phone interviews or allegations of misconduct with video or audio "evidence."

What's Next and What it Means for the Future of Work

As AI-based Deep Fake content-generation tools continue to mature in sophistication, fewer aspects of their operation or deployment require specialized technical knowledge and custom development. While it is not uncommon for a party to produce a fake or doctored e-mail for the purposes of retaliation, exploitation or reputational harm, fraudulent e-mails rarely withstand even cursory scrutiny, as e-mail servers, databases, mail clients and digital devices can be examined for authentication purposes. Moreover, the nature of digital communications provides numerous opportunities for forensic analysis. However, with Deep Fake tools that can easily map a person's face and digitally graft their likeness onto another person readily available on social media platforms like Snapchat and Instagram, even non-sophisticated parties have the potential to be incredibly destructive to unprepared employers.

For example, the momentary recording of a single phone call, video conference or webinar may provide sufficient training material for a nefarious party to impersonate management's likeness using nothing more than free software or features found on social media platforms. Imagine a hypothetical where an employer with a substantial number of warehouse-based employees is in the process of automating certain historically human-intensive job functions with robotic assistance. A short audio recording allegedly captures a member of management saying, "they'll all be replaced by this time next year," or mocking the concerns of employees. This video is then circulated and immediately inflames a highly delicate situation, while also generating a public relations crisis—which is subsequently used as political fodder by political operatives on social media, vilifying the company for trivializing its human workforce—all which occur before management has had adequate time to investigate or appropriately respond.

With resource constraints no longer serving as a material barrier to access state-of-the-art Deep Fake tools, combined with their increased ease-of-use, the frequency and scope of Deep Fake content generation is poised to rapidly accelerate. In other words, Deep Fake usage is no longer limited to the very technically sophisticated. The availability and potential creative use of these tools by any motivated party is increasingly relevant in the context of the workplace. Emerging startups currently designing systems to detect Deep Fakes are taking various forensic approaches to analyzing content, including scrutinizing video for imperceptible frame drops, developing algorithms that can predict accurate light reflection in an image, and identifying device-specific metadata. The use of Deep Fakes is expanding rapidly and employers need to be aware of the risk posed by this technology before the company becomes an unwittingly victim of its potentially catastrophic and nefarious applications.

Regardless of the learning model implemented for a particular function, these AI-based systems are increasingly ubiquitous, easier to replicate and less resource-intensive to leverage. AI-powered Deep Fakes can lead to risk and liability in hiring scenarios,

[5] Yaniv Leviathan and Yossi Matias, *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone* (May 8, 2018). *See* https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

investigations related to employee misconduct, public relations issues and workplace whistleblower situations, to name a few. For example, in a hostile workplace, sexual harassment or labor-relations case, the alleged appearance of impropriety or inappropriate conduct could be utilized to support claims of a specific pattern of workplace behavior and apply additional pressure on an organization during litigation.

Vulnerable Targets

With fully functioning Deep Fake tools readily available, the primary limitation to potential impersonation of an individual or fabrication of an event is the training data for which the content-generation tools use as a reference to draw inferences, extract data points and classify target- or situation-specific characteristics. Therefore, the higher the quality and volume of the training data, the better the results.

Public figures and individuals with large quantities and variations of recordings of their likeness in the public domain are obvious possible targets for Deep Fake impersonations and exploitation. But a manager who frequently leaves voice messages or other audio files may also be susceptible to fakes. Moreover, while the precise manifestations and contexts of future Deep Fakes are unknown, the following are several plausible circumstances where Deep Fakes could foreseeably arise.

- *Computer Generated Impersonations* – Re-creation of an individual's likeness, this can include fabricated video content, images, audio, writing style and simulated dialogue in the form of fabricated interpersonal exchanges, alleged recordings and even real-time impersonation and misrepresentation via audio, video or text.

One disturbing consideration is that there is no limitation as to the quantity of real-time, computer-generated impersonations that can be occurring in parallel; thus, many different people could be under the impression they are having or had entirely different conversations and interactions with the same person at the same point in time. For example, imagine if *all* your friends, family, colleagues, neighbors, social media contacts, places of business and an arbitrary number of strangers were under the impression that they were speaking to you live, right now (without your knowledge), each having a uniquely preposterous and negative personal interaction with a computer-generated impersonation of *you*. In a workplace context, the potential damage an organization could incur to strategic partnerships, public relations, workplace communications and essential operations from computer-generated, real-time impersonations of individuals from within management is extensive and profound.

- *Original Computer-Generated Personas* – Creation of a person that does not and has never existed.[6]

- *Mixed Interactions Between Machine and Humans* – Either a simulated interaction between impersonated and/or computer-generated personas, this can exist in the form of an alleged recording of an event that never occurred or in real-time between humans and machines impersonating a real or computer-generated persona.[7]

[6] *See* https://thispersondoesnotexist.com.

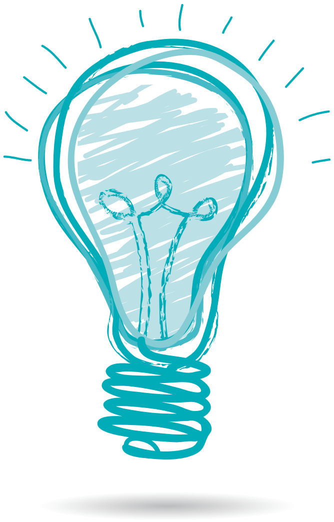
[7] *See* Google Duplex.

Preparing Employers for Scenarios Involving Deep Fakes

The authenticity of information published by or about employers is critical to internal operations and reputation management, and can directly influence market confidence of an organization.

The form in which content or communications allegedly from or concerning an organization may differ (*e.g.*, audio, video, real-time dialogue or printed text) and each form of content faces unique authentication validation challenges. Cryptographic private key signing of official communications is a basic first step employers can take to minimize manipulation or fabrication of employer-endorsed content. Cryptographic key signing could be embedded into video communication in the form of a visible or imperceptible watermark on official video, written communication and even audio. Other approaches include *fingerprinting* of content at the frame level from time of capture through dissemination. Ultimately, strategizing with counsel knowledgeable about technological vulnerabilities and legal issues that may arise will enable employers to better prepare for potential risks from Deep Fakes and incorporate safeguards unique to their circumstances. Critically, organizations must stay closely informed of the evolving threat landscape in order to prepare for and appropriately respond to highly dynamic scenarios.

While the mechanisms for detecting, authenticating and mitigating inauthentic content will continue to mature, incorporating organization-wide training and awareness for plausible hypothetical scenarios concerning Deep Fakes will allow employers to begin to develop protocols that can be implemented and initiated in a coordinated and timely manner.



Preliminary Recommendations

Until a method of authenticity verification is developed, everyone has exposure when it comes to Deep Fakes. For now, employers can employ additional methods of verification by analyzing other data such as GPS, time stamps on communications, cryptographic key signing, call logs, and other means of verification regarding allegations that may be centered around Deep Fake technology.

Employers can heighten their investigations and evaluations of claims of workplace misconduct or other areas where Deep Fakes may be used. What seems obvious may merely be deception to hurt corporations or the individuals employed by them. Employers should seek counsel for assistance related to additional methods of utilizing data analytics that may protect the company from exploitative uses of this technology.

Preparing for Deep Fakes – SPECTRE:

- Seeing is no longer believing – educate management on the potential issues and evolving vulnerabilities concerning Deep Fakes;
- Protocol – Have a response protocol in place to address harmful and questionable content relevant to management or the company;
- Engage Counsel – Strategize with counsel who is knowledgeable about technological driven vulnerabilities like Deep Fakes, familiar with legal issues that may arise from these, and familiar with computer forensic processes that can be used to test the validity of questionable content;
- Communicate – Initiate investigations by first speaking to relevant parties, as it may take time to conduct a thorough forensic analysis of pertinent digital content;
- Target Assessment – Identify parties within an organization that may be most susceptible to Deep Fakes;
- Reserve Judgment – Even seemingly innocuous digital content and communications can be used for exploitation and nefarious purposes; and
- Educate – Provide company-wide training to raise awareness of the evolving threat landscape and identify measures that can be taken to mitigate potential harm.

We welcome the opportunity to continue the discussion.

Chase Perkins, perkins@thoughtly.co
Natalie Pierce, npierce@littler.com
Aaron Crews, acrews@littler.com
Jill Weimer, jweimer@littler.com

To learn more about Littler's Data Analytics practice and Robotics, AI & Automation Practice Group, please visit:

littler.com/service-solutions/littler-big-data-initiative
littler.com/practice-areas/robotics-artificial-intelligence-ai-and-automation

Littler®

6.2019