Responding to Malicious Corporate Deepfakes

January 25, 2023

Content generated by artificial intelligence ("AI") continues to improve and become more convincing. These realistic images, audio, and videos, where used for purposes of a misrepresentation or to falsely spread information, are commonly dubbed "deepfakes." Governments around the world are taking notice of deepfakes and beginning to respond. As reported by the Wall Street Journal, China's internet regulator announced rules last month to restrict the creation of deepfakes by, for example, prohibiting their use to spread "fake news" or other information disruptive to the economy or national security. In the European Union, the recently updated Code of Practice on Disinformation now covers deepfakes and will require platforms to conduct periodic reviews of disinformation tactics used by malicious actors, also mandating the implementation of policies to cover those practices.

In the United States, however, the federal government has not yet offered legislation or regulation to address the general problem posed by deepfakes, and efforts to restrict deepfake creation would likely face First Amendment challenges. And while deepfakes have been a thorn in the side of celebrities and politicians for several years, most U.S. companies have not considered deepfakes to be a serious threat. With the widespread availability of cheap software that can generate quality deepfakes, that may be about to change. The recent proliferation of fake twitter accounts impersonating companies like Eli Lilly and Lockheed Martin, and the resulting losses for each company, illustrates the risk that deepfakes might pose and the damage they may cause.

For example, imagine a deepfake of Company X's CEO promising to do something popular that would nonetheless hurt the company's stock price—like the fake Eli Lilly tweet promising to release insulin for free. There are several reasons that a bad actor might choose to create and post such a video. They might wish to trade on the stock knowing that their deepfake would be likely to temporarily lower the company's stock price. They might intend to put pressure on a company to make a payment if said company was already the victim of a ransomware or data extortion attack. The deepfake might be an act of corporate activism or vandalism in response to some perceived wrongdoing by the company or the company's executive. Or perhaps it is a state-sponsored attack intended to harm competitors of important domestic industries.

All of these motivations have already led to attacks against companies; deepfakes will create an entirely new vector to which companies will have to respond. To prepare, companies should consider updating their cybersecurity programs and incident response plans to address corporate deepfakes, as well as consider running a tabletop to test those new protocols.

Detecting Corporate Deepfakes

As deepfakes become easier to make and deploy, the risk to companies increases. Whereas in the early days of the technology, creating a deepfake required bespoke coding and took substantial time, it is now possible to create a convincing deepfake using inexpensive off-the-shelf software. Tools like Open AI's Dall-E 2 and Stability AI's open-source Stable Diffusion expand access to this creative power, and may further increase the ubiquity of deepfakes, while making it more and more difficult to distinguish between legitimate and fraudulent AI-generated content.

These new tools give bad actors new opportunities to target businesses. For the modern-day company, vigilance is key. Companies should be diligent in monitoring social media for deepfake threats so that malicious videos and soundclips can be caught and addressed before they go viral. Even though there is no one-size-fits-all approach for an organization faced with a deepfake, companies should have a rapid reaction plan in place in the event that a deepfake does surface in order to ensure a quick and effective response.

Such plans should be tailored to the specific characteristics and circumstances of each company. If your organization does not already have an AI Incident Response Plan ("AI IRP") in place, it may be time to consider one. Though deepfake response is of a different nature than the incidents typically contemplated in an AI IRP, this rapid reaction plan has a natural home in that document. In adding deepfake response to an AI IRP, organizations should consider including:

- A means to quickly prove that the video, picture or audio is fake; draft press statements ready to go declaring the video to be fake, with distribution plans appended; and draft takedown notices ready to go for the sites that are hosting the deepfakes.

- The appropriate personnel that should be brought in to manage the response to the deepfake, and the clear delineation of lines of communication, escalations, and approvals to handle the response.

- Identification of key tasks and decisions that need to be made. This includes an assessment of the appropriate avenue to get the deepfake removed, options for which we discuss in the next section. In addition, processes should be in place to consider whether separate legal action is warranted should a takedown notice or request to remove the content be unsuccessful.

Companies might also consider investing in deepfake detection tools. The market for deepfake detection is growing, and employing such a tool can allow for both quick detection and documentation to support a takedown or removal request. For example, Intel recently announced that its deepfake video detection tool boasts a 96% accuracy rate and identifies such AI-generated content in real-time.

Responding to Corporate Deepfakes

Once a company has become aware of a targeted deepfake making the rounds on social media, what legal mechanisms are available? The answer is context-dependent, and the exact nature of the deepfake needs to be taken into account. There are a number of means of redress, each covering different circumstances, and we describe the landscape below.

Takedown Regime

If a deepfake makes use of copyrighted material to which a company owns the rights, the company can file what is known as a "takedown notice" under the Digital Millennium Copyright Act ("DMCA"). Such a notice must be sent to the website on which the allegedly infringing deepfake is hosted. The DMCA takedown regime offers copyright owners a way to seek quick removal of allegedly infringing material without resorting to formal legal action. At the same time, the DMCA regime offers the hosting site or platform immunity from suit for the alleged infringement, so long as it acts expeditiously to remove the content—a paramount consideration for platforms in light of the fact that the immunity provided by Section 230 of the Communications Decency Act does not apply to federal copyright claims and arguably does not apply to state law copyright claims.

Of course, for a takedown notice to be an effective means of removing a deepfake, there must actually be copyrighted material at issue. The DMCA takedown regime does not extend to trademarks or defamatory content.

Consider our earlier example involving a deepfake of a hypothetical company CEO making a statement harmful to the company's stock price. A takedown notice may not be the appropriate avenue to remove an audio deepfake—trained on publicly available

voice clips of the CEO's past speeches—where no copyrighted material is at issue. For a real-world analogue, consider the failed takedown request for a deepfake of Jay-Z reciting the "To Be, or Not to Be" soliloquy from *Hamlet*. YouTube, in reinstating the content, stated that the takedown request was "incomplete," perhaps because it did not identify any copyrighted material in the use of Jay-Z's voice or mannerisms.

At the same time, if the deepfake uses video of the CEO—say from a corporate promotional video—for which the company holds a copyright, this takedown request could fare much better. For example, a 2019 video deepfake featuring Kim Kardashian was successfully removed from YouTube. That takedown request came from Condé Nast, which had a copyright over the original video published by *Vogue*, which was used to create the deepfake.

Though the Condé Nast example demonstrates one circumstance in which a DMCA takedown notice can successfully target deepfakes, the waters get muddied by "fair use" considerations, which may lead hosting websites and platforms to stop short of removing some content. Fair use allows for the use of copyrighted materials for "transformative" purposes without the copyright owner's consent. These "transformative" purposes typically include commentary, criticism, or parody, and where a deepfake arguably straddles one of those lines, a request may not succeed. In addition, as the United States Supreme Court considers the bounds of what constitutes a "transformative" purpose for the sake of fair use, the legal landscape on this issue may shift.

Terms of Service

Another avenue for removing deepfakes—which may be particularly useful where no copyrighted material is involved—relies on hosting websites' or platforms' terms of service. These relevant terms of service cover two potentially applicable categories: those covering deepfakes and those addressing trademark infringement. Discussing these in order:

Deepfake-Specific Terms of Service

Platforms like Meta and Twitter specifically address "manipulated media" and "synthetic and manipulated media," respectively, in their policies. Though each platform's requirements for the removal of deepfakes differ, the thrust is that substantially edited media that is both AI-generated and shared in a deceptive manner can be reported and removed. Returning to our hypothetical CEO deepfake, this is where a robust AI IRP that addresses deepfake response can be of particular value. Platforms have no legal obligation to remove deepfakes where non-copyrighted material is involved. Having a team ready to collect the necessary information as soon as the deepfake is detected—including clear proof that the deepfake actually is a fake—

and put forward a strong case to the hosting platform will help put the business in the best possible position to prevail on its request.

Trademark Infringement

With no corresponding legal regime like DMCA to cover trademark infringement on hosting websites and platforms, terms of service preventing misuse of trademarks can help address deepfakes. However, as with the deepfake-specific terms of service, the hosting website or platform will have discretion in deciding whether to remove the content. Platforms are sometimes hesitant to intervene in trademark disputes, so highlighting the fraudulent and deceptively confusing uses of trademarks in deepfakes will be key when making takedown requests. Legal action for trademark infringement also remains an option for the trademark holders, but identifying the creator of the content and filing a federal lawsuit is considerably more expensive and time-consuming. The same lesson with respect to the hypothetical holds true—highlighting the strength of the company's request, demonstrating that there is a strong legal basis behind it—will best position the business to have the deepfake of the CEO removed.

State Laws

Though not essential for responding to corporate deepfakes, our discussion of deepfake takedowns would not be complete without a primer on the various state laws that have been passed to address deepfakes. Deepfakes are not categorically illegal to make, share, or host. Instead, states prohibit deepfakes in specific, harmful contexts. Though these laws are useful in highlighting that legislative bodies have started to think about deepfake regulation, they are unlikely to provide businesses with a means of redress when deepfakes are created that bear on their reputational or financial interests.

These laws fall into two main categories: laws regulating deepfakes in the context of nonconsensual intimate imagery ("NCII")—which make up 96% of all extant deepfake videos as of 2019—and laws regulating deepfakes in the context of election misinformation. Outside of these two categories, attempts at regulation are rarer, although New York has passed a law that regulates deepfakes of deceased celebrities as a component of their right of publicity. No state has, as of yet, passed a law banning deepfakes in general, nor is it likely that any will. As previously mentioned, such legislation would raise significant First Amendment concerns.

Only time will tell what impact the rise of deepfakes will have on companies. In order to protect your organization, awareness of the possibilities and threats that deepfakes represent is a vital first step. Furthermore, even though there is no one-size-fits-all approach for an organization faced with a deepfake, being prepared with a plan will put a company in a good position to successfully combat the threat.

Individual Remedies

The individual targeted by a corporate deepfake can take legal action in their personal capacity, too. Defamation, right of publicity, and false endorsement are three causes of action that the individual can bring against those responsible for perpetrating a deepfake fraud, assuming the creator can be identified and prosecuted. These actions can present a means to recoup losses stemming from the deepfake, and depending on the jurisdiction and type of action, offer deepfake targets an opportunity to seek injunctive relief.

Defamation

Targets of allegedly manipulated media have already brought defamation claims, and subjects of true deepfakes will likely follow. For example, former President Donald Trump's campaign sued a Wisconsin television station during the 2020 campaign for running an advertisement allegedly combining a number of clips of Trump's voice to make false and defamatory statements.

For a defamation claim to succeed, however, the false statement must purport to be fact. If the deepfake's creator intends to fraudulently move markets or otherwise cause individuals to act in a way that allows the creator to profit, a court may well find that the deepfake's message was meant as fact.

Though there may be a trend among the states toward permitting injunctive relief in defamation cases, typically the remedy is not available on a preliminary basis and is only an option following full adjudication of the lawsuit. Injunctions in the defamation context must also be narrowly tailored to avoid constituting a "prior restraint" running afoul of the First Amendment. While a defamation suit could be worthwhile for a deepfake target, it therefore does not necessarily offer the prompt and preventive relief of a DMCA takedown request.

Rights of Publicity

The rights of publicity, which vary in scope from state to state, typically protects individuals against the misappropriation of their name, likeness, voice, or other indicia of personality for commercial benefit. Though courts have not yet spoken on whether a deepfake fraud could constitute such a "commercial benefit," returning to our hypothetical, a prominent corporate executive may find it worthwhile to assert such a claim if the executive or his or her organization can identify the perpetrator.

This cause of action could become very appealing to high-profile deepfake targets seeking to halt a deepfake's spread, as a court can issue a preliminary injunction in response to a right of publicity claim. In some jurisdictions, the court can make use of its inherent injunctive power, while in others like New York, injunctive relief is expressly authorized by statute.

False Endorsement

Individuals may also have a claim for false endorsement under trademark law if a deepfake causes consumers to misleadingly believe that the individual sponsors or approves of a product or service. This cause of action requires a showing of actual deception, or at least a tendency to deceive a substantial portion of the intended audience, and a likelihood of injury to plaintiff. While most of these cases arise in the context of a celebrity, some courts have held that celebrity status is not a necessary prerequisite for a successful false endorsement claim. While there are hurdles to proving false endorsement, the benefit of this cause of action is entry into federal court and the availability of immediate injunctive relief.

To subscribe to our Data Blog, click here.

The authors would like to thank Debevoise Law Clerk Abigail Liles for her work on this Debevoise Data Blog post.

* * *

Please do not hesitate to contact us with any questions.

NEW YORK

Megan K. Bannigan
mkbannigan@debevoise.com

Avi Gesser
agesser@debevoise.com

Scott M. Caravello
smcaravello@debevoise.com

SAN FRANCISCO

Anna R. Gressel
argressel@debevoise.com

Christopher S. Ford
csford@debevoise.com