The Growing Threat of Deepfakes to Brand and Executive Reputation

Alex Romero, COO, Constella Intelligence

Fake viral videos, images, and audio clips that appear indiscernibly real are catching public attention in a myriad of ways, as bad actors look to intentionally damage reputations or impersonate key individuals to obtain sensitive data or influence public opinion. Known as a “deepfake,” a portmanteau of “deep learning” and “fake,” this synthetic content is created using human image or audio synthesis based on machine learning (ML) or artificial intelligence (AI). Today, the concept and content are tangible, emerging as a legitimate threat to businesses and executives alike.

Moody’s research highlights two alarming facts about today’s digital world: AI is making it easier to damage companies’ credit, reputation, and financial health via deepfakes; and this risk will become harder to manage and mitigate as the AI that enables synthetic media continues to evolve. This means malign influence and disinformation campaigns powered by AI will take more time and resources to definitively disprove. Visit the site www.thispersondoesnotexist.com to see just how far deepfake production technologies have come. The extended time exposure to these types of campaigns is a real threat that compounds the risk associated with deepfakes.

A 2020 Brookings Institution report succinctly outlined the political and social risk presented by deepfakes: “distorting democratic discourse; manipulating elections; eroding trust in institutions; weakening journalism; exacerbating social divisions; undermining public safety; and inflicting hard-torepair damage on the reputation of prominent individuals, including elected officials and candidates for office.” The risk has been exacerbated by COVID-19. Work environments have transitioned to virtual operations, increasing the risk surface of most organizations and presenting new challenges they may not be equipped to manage – specifically regarding security and crisis response. This new paradigm means more frequent use of digital mediums of communication without the validation and checks offered by in-person communication.

At an enterprise level, deepfakes pose two specific threats: social engineering attacks and public opinion manipulation. Social engineering is characterized by manipulating individuals to perform malicious actions, such as sharing confidential information or providing access to sensitive data. Threat actors can influence public opinion through fabricated videos of executives and other high profile individuals sharing disinformation or making inappropriate statements. Depending on the nature of the weaponized information, this can have far-reaching effects including influencing brand reputation and consequently consumer behavior, in addition to affecting a company’s stock price.

Best practices to prevent such threats include real-time monitoring and analysis of digital media and websites for disinformation. Further, a swift and comprehensive approach to brand reputation management and coordinated crisis response are essential to mitigate the potential damage caused in worst-case scenarios: multiple teams within an organization, from communications to cybersecurity to compliance must now closely work together to anticipate and mitigate such emerging digital risks. Experts across a wide range of fields agree that the battle against the malign use of deepfakes will necessitate the development of advanced security solutions, including threat monitoring and mitigation technologies, along with robust education of employees, stakeholders, and everyday individuals of the risks posed by manipulated audio, image, and video content.

Encouragingly, stakeholders across the digital ecosystem are making strides toward developing methodologies for tracking, analyzing, and delivering solutions that can mitigate the risks associated with deepfakes in the public and private sectors. Some notable initiatives include the Content Authenticity Initiative (CAI) – whose objective is to establish industry-wide standards for digital authenticity verification, assisting in limiting the pernicious effects of deepfakes before they can cause personal or corporate harm. There’s also the Deepfake Detection Challenge, launched by a coalition of partners including Facebook, Microsoft, and Google, among others, in partnership with key academic institutions including Cornell, Oxford, MIT, and UC Berkeley to incentivize and catalyze the development of open-source deepfake detection tools across the broader academic and tech communities.

Is your organization equipped to safeguard against deepfakes? Ask yourself these three questions:

1. Organization: Is your organization, namely executives and key team leaders, familiar with the most likely crisis scenarios posed by deepfakes – including phishing attacks, potential stock market manipulation, and blackmail or extortion? Are your employees trained and prepared to spot and report deepfakes, suspicious synthetic or manipulated content?
2. Technology: Can you accurately identify where, when, and how your brand is mentioned both in the public digital sphere and deep & dark web to establish a basis for mitigation of incidents in real-time? Has your organization evaluated the feasibility of deepfake detection technology for mitigation of threats against your brand, executives, and key individuals?

3. Processes: Do you have a cross-organizational incident response plan that clearly details steps for security and communications remediation once an incident occurs?

Companies large and small must safeguard their reputations and assets amid a digital ecosystem witnessing the proliferation of deepfake technology. Use your peacetime wisely and address the threat of deepfakes today.

About the Author: *Alex Romero is COO and co-founder of Constella Intelligence, a company that uses advanced data analytics, artificial intelligence, and proprietary technology to analyze the digital sphere and help protect organizations against malign threats emerging from the digital media ecosystem, including synthetic media such as deepfakes.*