WINTER 2024 E-NEWSLETTER

At Digital Mountain, we assist our clients with their electronic discovery, digital forensics, cybersecurity, and data analytics needs. For this E-Newsletter, we discuss the development of Generative AI as it approaches two years old and how courts and states are responding to this new age of technology.

Generative AI's Terrible Twos?

If early LLMs, like ChatGPT, were Generative AI's infancy, then it is probably fair to analogize that we have now entered Generative AI's terrible twos. This is a period of a child's life beginning around eighteen months of age when rapid growth often outpaces emotional maturation, during which temper tantrums, mood swings, unpredictable behavior, and an inability to verbalize their frustrations make parenting a challenge. Developmentally, Generative AI (Gen AI) is growing in scope, skill, and utilization. We've evolved from static corpus chatbots that could "read" a prompt and "create" a response, to multimodal AI systems that combine text, video, audio, and numeric data sets in tandem, as well as to perform "live" internet searches. This explosion of near daily growth has led to the speculation that Generative AI is broaching Artificial General Intelligence (AGI) (2303.12712.pdf (arxiv.org)). While AGI is still admittedly a reach technologically, that's not all we must be concerned with when it comes to Gen AI advancements.

Threat actors are creating and accessing Malla, (an acronym for **Ma**licious **l**arge **l**anguage **a**pplications), to uplevel traditional scams with Gen AI offerings such as FraudGPT and WormGPT. There are also jailbreaking LLMs with misinformation/disinformation spreading LLM prompts such as, DarkBARD, and some 200 others studied as part of a paper released early this year (Malla: Demystifying Real-world Large Language Model Integrated Malicious Services (arxiv.org)). According to the study, the top three uses of Malla are malware creation, propagation of phishing scams, and fake/scam websites (ibid). Despite safeguards designed to prevent threat actors from taking advantage of legitimate LLM hosting sites, Malla projects were found to evade detection and takedowns for months (ibid), indicating that LLM-hosting may be growing faster than hosting platforms can monitor for malicious activity.

Despite the staggering growth of the Malla market, that may be dwarfed by the leaps and bounds made in the area of account and identity compromise. Recently, the Hong Kong arm of a multinational corporation based in the UK was at the center of a complex deepfake scam enabled by advancements in Gen AI technology (Company Lost $25 Million After Employee Tricked by Deepfakes on Call: HK Police (businessinsider.com)). The elaborate theft required the acquisition of multiple company account profiles, video and voice prints of company employees, including the CFO, a staged video-conference, and a potential Business Email Compromise to transmit instructions to effect fraudulent wire transfers in excess of $25 million. While investigations are still ongoing, the Hong Kong police confirmed that the employee who transferred the funds did not take part in the deepfake video conference, further evidencing the complexity and level of detail that was devoted to the scam (ibid). At Digital Mountain, we've been working on both Wire Transfer Fraud and Account and Identity Compromise cases with increasing frequency, a trend we hope to see the entire technology sector work to reverse.

Multiple recent deepfake events have legislators moving faster than babysitting grandparents after President Biden and billionaire superstar Taylor Swift were both victims of deepfake creations. In January 2024, New Hampshire primary voters received robocalls purportedly of President Biden encouraging voters not to vote in the New Hampshire Presidential Primary but to save their votes for the November general election and giving the impression of voting restrictions that were not accurate (NH primary AI deepfake Biden robocall sourced IDed – NBC Boston). While the FCC and the New Hampshire Attorney General's office believe they have identified the creators/disseminators of the deepfake calls, the process to create such a call takes very little time, skill, or resources thanks to Gen AI advancements. A convincing deepfake voiceprint can be created with less than a minute of voice sampling – which, for anyone who has had their voice uploaded to the internet, the small sample size required is a sobering thought. It's not illegal, per se, to create cloned voice prints, and many legitimate businesses are offering services which will create a voice print from text for a fee (https://speechify.com for example), but ask anyone who has received a deepfake call supposedly of a loved one claiming to have been kidnapped, and they'll probably tell you that they're willing to dial back on the AI enhancements.

Whether you're a Swiftie or not, the deepfake pornographic imagery of Taylor Swift that was viewed 45 million times before successfully being taken down from social media is nothing to sing about. Reportedly the deepfakes of Taylor Swift were the result of the 4chan online community and a response to a challenge whereby Gen AI image creation tools were utilized. Shortly after, the images were spread via Telegram and X (formerly Twitter), the later halting all searches for "Taylor Swift" until the images could be scrubbed (4chan daily challenge sparked deluge of explicit AI Taylor Swift images | Ars Technica). As with voice deepfakes, creating deepfake images, including deepfake video, is surprisingly easy with legitimate tools and a few jailbreaking prompts to bypass safeguards. While two pieces of deepfake legislation await votes in the US Congress, individual states are taking swifter action to criminalize damaging deepfakes (Taylor Swift deepfake pornography sparks renewed calls for US legislation | Taylor Swift | The Guardian; States turn their attention to regulating AI and deepfakes as 2024 kicks off (nbcnews.com)).

As we prepare ourselves for the next Gen AI growth spurt, we may have to adopt a "don't throw the baby out with the bath water" philosophy and instead concentrate on how we can use technology in responsible ways, including using the same technology that creates the trouble to detect, remediate, and prevent deepfakes. At Digital Mountain, we're committed to helping do our part to raise this technology child right.

Please direct questions and inquiries about electronic discovery, digital forensics, cybersecurity, and data analytics to info@digitalmountain.com.

UPCOMING INDUSTRY EVENTS

ASU-ARKFELD E-DISCOVERY, LAW & TECHNOLOGY CONFERENCE
Phoenix, AZ: March 5-6, 2024

PLANET CYBER SEC CISO-CIO FORUM
Redondo Beach, CA: March 13, 2024

MASTERS CONFERENCE MARCH 2024
Dallas, TX: March 21, 2024

IAPP GLOBAL PRIVACY SUMMIT 2024
Washington, DC: April 3-4, 2024

MAGNET USER SUMMIT 2024
Nashville, TN: April 15-17, 2024

Click here to see more upcoming events and links.

Digital Mountain, Inc. Founder and CEO, Julie Lewis, will be presenting at various upcoming industry events. Please send requests for speaker or panel participation for her to marketing@digitalmountain.com.

DIGITAL MOUNTAIN, INC.
4633 Old Ironsides Drive, Suite 401
Santa Clara, CA 95054
866.DIG.DOCS

www.digitalmountain.com

Contact us today!

FOLLOW US AT:

in f 🐦