CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

WORKING PAPER

JULY 2020

Cyber Policy Initiative Working Paper Series | “Cybersecurity and the Financial System” #7

Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios

Jon Bateman

Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios

Jon Bateman

+ CONTENTS

Cybersecurity and the Financial System

Carnegie's working paper series "Cybersecurity and the Financial System" is designed to be a platform for thought-provoking studies and in-depth research focusing on this increasingly important nexus. Bridging the gap between the finance policy and cyber policy communities and tracks, contributors to this paper series include government officials, industry representatives, and other relevant experts in addition to work produced by Carnegie scholars. In light of the emerging and nascent nature of this field, these working papers are not expected to offer any silver bullets but to stimulate the debate, inject fresh (occasionally controversial) ideas, and offer interesting data.

If you are interested in this topic, we also invite you to sign up for Carnegie's FinCyber newsletter providing you with a curated regular update on latest developments regarding cybersecurity and the financial system: CarnegieEndowment.org/subscribe/fincyber.

If you would like to learn more about this paper series and Carnegie's work in this area, please contact Tim Maurer, co-director of the Cyber Policy Initiative, at tmaurer@ceip.org.

Papers in this Series:

- "Cyber Mapping the Financial System," Jan-Philipp Brauchle, Matthias Göbel, Jens Seiler, and Christoph von Busekist, April 2020
- "Lessons Learned and Evolving Practices of the TIBER Framework for Resilience Testing in the Netherlands" Petra Hielkema and Raymond Kleijmeer, October 2019
- "Cyber Risk Scenarios, the Financial System, and Systemic Risk Assessment" Lincoln Kaffenberger, Emanuel Kopp, September 2019
- "Cyber Resilience and Financial Organizations: A Capacity-building Tool Box," Tim Maurer and Kathryn Taylor, July 2019
- "The Cyber Threat Landscape: Confronting Challenges to the Financial System" Adrian Nish and Saher Naumaan, March 2019
- "Protecting Financial Institutions Against Cyber Threats: A National Security Issue" Erica D. Borghard, September 2018
- "Toward a Global Norm Against Manipulating the Integrity of Financial Data" Tim Maurer, Ariel (Eli) Levite, and George Perkovich, March 2017

About the Author

Jon Bateman is a fellow in the Cyber Policy Initiative of the Technology and International Affairs Program at the Carnegie Endowment for International Peace. He previously worked as a senior intelligence analyst, policy adviser, and speechwriter in the U.S. Defense Department, most recently serving as special assistant to the Chairman of the Joint Chiefs of Staff.

Acknowledgments

The author would like to recognize Tim Maurer, Matt Noyes, and Paul Williams for providing valuable written feedback on this paper. Special gratitude goes to more than thirty international experts from the financial sector, tech industry, and regulatory community for their wealth of insight shared during a roundtable discussion on an early draft. Evan Burke, Ronit Langer, Ariel (Eli) Levite, Arthur Nelson, Mike Nelson, and Natalie Thompson also made important suggestions during an internal workshop.

Finally, the author thanks Charlotte Stanton and more than three dozen social media platform representatives, technologists, and leading experts, who privately convened in 2018 and 2019 to discuss safeguarding elections from synthetic and manipulated media. Those dialogues helped to create a starting point for this subsequent work.

Summary

Rapid advances in artificial intelligence (AI) are enabling novel forms of deception. AI algorithms can produce realistic "deepfake" videos, as well as authentic-looking fake photos and writing. Collectively called synthetic media, these tools have triggered widespread concern about their potential in spreading political disinformation. Yet the same technology can also facilitate financial harm. Recent months have seen the first publicly documented cases of deepfakes used for fraud and extortion.

Today the financial threat from synthetic media is low, so the key policy question is how much this threat will grow over time. Leading industry experts diverge widely in their assessments. Some believe firms and regulators should act now to head off serious risks. Others believe the threat will likely remain minor and the financial system should focus on more pressing technology challenges. A lack of data has stymied the discussion.

In the absence of hard data, a close analysis of potential scenarios can help to better gauge the problem. In this paper, ten scenarios illustrate how criminals and other bad actors could abuse synthetic media technology to inflict financial harm on a broad swath of targets. Based on today's synthetic media technology and the realities of financial crime, the scenarios explore whether and how synthetic media could alter the threat landscape.

The analysis yields multiple lessons for policymakers in the financial sector and beyond:

Deepfakes and synthetic media do not pose a serious threat to the stability of the global financial system or national markets in mature, healthy economies. But they could cause varying degrees of harm to individually targeted people, businesses, and government regulators; emerging markets; and developed countries experiencing financial crises.

Technically savvy bad actors who favor tailored schemes are more likely to incorporate synthetic media, but many others will continue relying on older, simpler techniques. Synthetic media are highly realistic, scalable, and customizable. Yet they are also less proven and sometimes more complicated to produce than "cheapfakes"—traditional forms of deceptive media that do not use AI. A bad actor's choice between deepfakes and cheapfakes will depend on the actor's strategy and capabilities.

Financial threats from synthetic media appear more diverse than political threats but may in some ways be easier to combat. Some financial harm scenarios resemble classic political disinformation scenarios that seek to sway mass opinion. Other financial scenarios involve the direct targeting of private entities through point-to-point communication. On the other hand, more legal tools exist to fight financial crime, and societies are more likely to unite behind common standards of truth in the financial sphere than in the political arena.

These ten scenarios fall into two categories, each presenting different kinds of challenges and opportunities for policymakers. Six scenarios involve "broadcast" synthetic media, designed for mass consumption and disseminated widely via public channels. Four scenarios involve "narrowcast" synthetic media, tailored for small, specific audiences and delivered directly via private channels. The financial sector should help lead a much-needed public conversation about narrowcast threats.

Organizations facing public relations crises are especially vulnerable to synthetic media. Broadcast synthetic media will tend to be most powerful when they amplify pre-existing negative narratives or events. As part of planning for and managing crises of all kinds, organizations should consider the possibility of synthetic media attacks emerging to amplify the crises. Steps taken in advance could help mitigate the damage.

Three malicious techniques appear in multiple scenarios and should be prioritized in any response. *Deepfake voice phishing (vishing)* uses cloned voices to impersonate trusted individuals over the phone, exploiting victims' professional or personal relationships. *Fabricated private remarks* are deepfake clips that falsely depict public figures making damaging comments behind the scenes, challenging victims to refute them. *Synthetic social botnets* are fake social media accounts made from AI-generated photographs and text, improving upon the stealth and effectiveness of today's social bots.

Effective policy responses will require a range of actions and actors. As in the political arena, no single stakeholder or solution can fully address synthetic media in the financial system. Successful efforts will involve changes in technology, organizational practices, and society at large. The financial sector should consider its role in the broader policymaking process around synthetic media.

Financial institutions and regulators should divide their policy efforts into three complementary tracks: internal action, such as organizational controls and training; industry-wide action, such as information sharing; and multistakeholder action with key outside entities, including tech platforms, AI researchers, journalists, civil society, and government bodies. Many notional responses could draw on existing measures for countering financial harm and disinformation.

Introduction

The advent of deepfakes and other synthetic, AI-generated media has triggered widespread concern about their use in spreading disinformation (see box 1). Most attention so far has focused on how deepfakes could threaten political discourse. Carnegie, for example, has extensively researched how to protect elections against malicious deepfakes.[1] In contrast, there has been relatively little analysis of how deepfakes might impact the financial system.

Disinformation is hardly new to the financial world. Crimes of deceit, such as fraud, forgery, and market manipulation, are endemic challenges in every economy. Moreover, bad actors often incorporate new technologies into their schemes. It is therefore worth considering how novel deception tools like deepfakes could enable financial crimes or other forms of financial harm.

This paper merges two of Carnegie's research areas. The FinCyber project works to better protect the financial system against cyber threats and to strengthen its resilience. The Deepfakes project has sought to develop safeguards against malicious deepfakes and other AI-generated disinformation. Through both projects, Carnegie has engaged extensively with leading stakeholders from industry, government, and academia.

In February 2020, Carnegie convened a private roundtable to discuss deepfakes in the financial sector. More than thirty international experts from the financial sector, tech industry, and regulatory community participated. This paper is informed by their collective insights, though it does not attempt to reflect any consensus.

Experts disagree sharply about the magnitude of financial threats posed by deepfakes. There have been only a handful of documented cases to date, making future trends difficult to judge. Some in the financial industry rank deepfakes as a top-tier technology challenge, predicting that they will corrode trust across the financial system and require significant policy changes. Others believe that deepfakes have been overhyped and that existing systems of trust and authentication can readily adapt to this new technology.

To advance the debate, this paper (1) identifies specific ways that deepfakes and other synthetic media could facilitate financial harm, (2) assesses their likely impact, and (3) offers lessons for policymakers. Based on today's synthetic media technology and the realities of financial crime, plausible threat scenarios are explored for four likely target groups: individuals, companies, markets, and regulatory structures. Although no set of scenarios can be comprehensive, this paper attempts to broadly outline the challenge.

BOX 1

What Are Deepfakes and Synthetic Media?

Deepfakes are AI-generated media that depict made-up events, sometimes quite realistically. As a slang term, "deepfake" has no agreed-upon technical definition.[2] It most commonly refers to fabricated video or audio of a person saying or doing something they never said or did.

Deepfakes are created using an AI method called deep learning. It relies on a complex computing system called a deep neural network, loosely modeled on biological brains. The network first ingests training data (samples) of a targeted person's face or voice and then applies an algorithm to extract mathematical patterns from the data. Based on these patterns, the network can generate new, synthetic representations of the target's face or voice.[3]

The best-known type of deepfake is a **face-swap video**, which transposes one person's facial movements onto someone else's features. (The term "deepfake" was coined in 2017 when internet users began transposing female celebrities' faces onto other faces in pornographic videos.[4]) Another type of deepfake is **voice cloning**, which copies a person's unique vocal patterns in order to digitally recreate and alter their speech.

Face swaps and voice cloning aim to mimic real individuals, but deep learning can also be used to "imagine" entirely fictitious people or objects.[5] Algorithms can generate synthetic photographs of nonexistent people, animals, and landscapes or create synthetic voices that belong to no one. AI can also produce synthetic text meant to emulate human-authored text. All of these fabrications are sometimes loosely labeled as deepfakes, but for clarity, this paper uses that term only for video or audio representations. Deepfakes are thus a subset of **synthetic media**, a broad category including all AI-generated video, images, sound, and text.[6]

Technologies used to create synthetic media have rapidly developed and proliferated in the last few years. Breakthroughs in machine learning science and in large-scale data collection, processing, storage, and transmission have made synthetic media appear much more realistic. And user-friendly software and cheap cloud computing now enable any technically savvy person to generate synthetic media—in some cases, through a free web interface or mobile app.

Despite these advances, however, synthetic media varies in quality. Face-swap videos can be strikingly lifelike, yet close observation usually reveals a lack of fine detail, fuzzy edges, or odd artifacts.[7] Algorithms for synthetic photographs and text can produce hyperrealistic results but also occasional absurdities.

Synthetic media quality generally depends on three factors: employing a strong and well-tailored algorithm; ingesting a large, diverse set of training data; and applying sufficient computing power and processing time. Innovation is occurring in all three areas as developers create more powerful algorithms that require less training data and computing resources.

Methodology

Deceptive synthetic media could be used to inflict financial harm on a wide range of potential targets. Obvious targets include financial institutions, such as banks, stock exchanges, clearinghouses, and brokerages—all of which rely on truthful information to conduct transactions—as well as financial regulators and central banks, which oversee general market conditions and combat harmful misinformation. But companies and individuals outside the financial sector also could become targets.

This paper therefore assesses ten threat scenarios facing four groups of potential victims in order of increasing systemic importance: (1) individuals, (2) companies (financial and nonfinancial), (3) markets (national and global), and (4) financial regulatory structures. Although each scenario is assigned to a group for the sake of simplicity, some scenarios could eventually impact multiple groups. For example, Scenario 1 depicts how synthetic media could enable identity theft. Identity theft harms individuals (people whose identities are stolen) but could also harm companies (for example, banks that issue fraudulent credit cards to the perpetrator and retailers that unwittingly process sales charged to those cards). In other words, small-scale harms, if aggregated, could theoretically have higher-order effects; and large-scale harms would inevitably trickle down to individuals.

Bad actors already have many proven, cost-effective techniques at their disposal, including digital technologies. Thus, this paper compares synthetic media against today's commonly used technologies for financial crime and harm. Each scenario outlines an established illicit scheme, imagines how synthetic media could enhance the scheme, and assesses the threat level and potential harm.

Directly comparing synthetic media against more commonly used tools helps illuminate the most concerning scenarios—those that could greatly empower bad actors and require new policy responses. Likewise, it helps identify the least concerning scenarios—those that are no more dangerous than today's threats and may not require additional responses.

While largely unrealized at this point, the ten scenarios are all feasible because they involve widely available synthetic media technologies. In other scenarios, the relevant technologies exist but remain proprietary or commercially controlled, suggesting a window of opportunity to control their proliferation and restrict future abuse.

Scenario Overview

Narrowcast and Broadcast Synthetic Media

All ten scenarios fall under one of two categories of synthetic media. Scenarios 1–4 involve what might be called **narrowcast synthetic media**, which are tailored for small, individual targets (such as a business's payroll officer) and delivered directly via private channels (such as a phone call). Scenarios 5–10 involve **broadcast synthetic media**, which are designed for mass targets (such as the investment community) and disseminated widely via public channels (such as social media).

Each category presents different challenges and opportunities for policymakers. For example, broadcast synthetic media tend to spread through content-monitored channels like social media platforms or news reports, which provide central opportunities for detection, moderation, and fact-checking. Narrowcast synthetic media, in contrast, may transit through phone lines, SMS, or email, which are relatively unmonitored spaces. Countermeasures are still possible but would take different forms, such as antispam and antispoofing technologies that help to authenticate identities and flag or filter suspicious actors.[8]

Broadcast synthetic media have already received significant attention from policymakers worried about political subterfuge and election interference. But narrowcast synthetic media have so far attracted little notice. The financial sector should seek to remedy this imbalance and foster greater public awareness of narrowcast threats, given their potential for financial harm in multiple scenarios.

Three Key Malicious Techniques

Across all ten scenarios, three malicious techniques appear repeatedly: deepfake voice phishing (vishing), fabricated private remarks, and synthetic social botnets. **Deepfake vishing**, a type of narrowcast threat, uses cloned voices for social engineering phone calls. This technique has innumerable applications, including identity theft, imposter scams, and fraudulent payment schemes. A well-crafted deepfake vishing operation can exploit the call recipient's trust in the impersonated party. Many victims will attribute any flaws in the cloned voice to a faulty line or succumb to emotional manipulation during a high-pressure phone call.

Fabricated private remarks, a type of broadcast threat, are deepfake video or audio clips that falsely depict a public figure making damaging comments behind the scenes. Again, there are many applications, including stock manipulation, malicious bank runs and flash market crashes, and fabricated government actions. It is difficult to prove a negative—that purported private remarks never occurred—so victims may need to rely on their reputations to refute false charges, which some will be unable to do.

Synthetic social botnets, another broadcast threat, are also of primary concern. Fake social media accounts could be constructed from synthetic photographs and text and operated by AI, potentially facilitating a range of financial harm against companies, markets, and regulators. Synthetic botnets would be more effective and harder to expose than the bots existing today. It seems likely that social bots will incorporate more AI over time, intensifying the technology competition between social media platforms and bad actors.

Table 1 provides an overview of the ten scenarios, the distinction between narrowcast and broadcast synthetic media, and role of three key techniques used. Following the overview are detailed assessments of each scenario, organized by the four target groups.

TABLE 1

Ten Synthetic Media Scenarios for Financial Harm

Target	Scenario	Role of Synthetic Media	Key Malicious Technique
Individuals	1. Identity theft	Voice cloning or face-swap video is used to impersonate a wealthy individual and initiate fraudulent transactions. Alternatively, it is used to impersonate a corporate officer and gain access to databases of personal information, which can enable larger-scale identity theft.	🔊
	2. Imposter scam	Voice cloning or face-swap video is used to impersonate a trusted government official or family member of the victim and coerce a fraudulent payment.	🔊
	3. Cyber extortion	Synthetic pornography of the victim is used for blackmail.	
Companies	4. Payment fraud	Voice cloning or face-swap video is used to impersonate a corporate officer and initiate fraudulent transactions.	🔊
	5. Stock manipulation via fabricated events	Voice cloning or face-swap video is used to defame a corporate leader or falsify a product endorsement, which can alter investor sentiment.	◌
	6. Stock manipulation via bots	Synthetic photos and text are used to construct human-like social media bots that attack or promote a brand, which can alter investor perception of consumer sentiment.	👤
	7. Malicious bank run	Synthetic photos and text are used to construct human-like social media bots that spread false rumors of bank weakness, which can fuel runs on cash.	👤
Markets	8. Malicious flash crash	Voice cloning or face-swap video is used to fabricate a market-moving event.	◌
Regulatory Structures	9. Fabricated government action	Voice cloning or face-swap video is used to fabricate an imminent interest rate change, policy shift, or enforcement action.	◌
	10. Regulatory astroturfing	Synthetic text is used to fabricate comments from the public on proposed financial regulations, which can manipulate the rulemaking process.	

🔊 **Deepfake voice phishing**

◌ **Fabricated private remarks**

👤 **Synthetic social botnet**

Narrowcast

Broadcast

Scenarios Targeting Individuals

Scenario 1: Identity Theft

Pre-existing threat. Identity theft is the most common type of consumer complaint received by the U.S. Federal Trade Commission (FTC).[9] In the typical case, a criminal opens a new credit card account in the victim's name.[10] This requires the victim's personal information, which hackers can acquire in bulk by breaching commercial databases.[11] Large-scale cyber breaches have fueled the growth of sophisticated criminal ecosystems to facilitate identity theft.[12] Stolen data caches are bought and sold as virtual commodities in online black markets.[13]

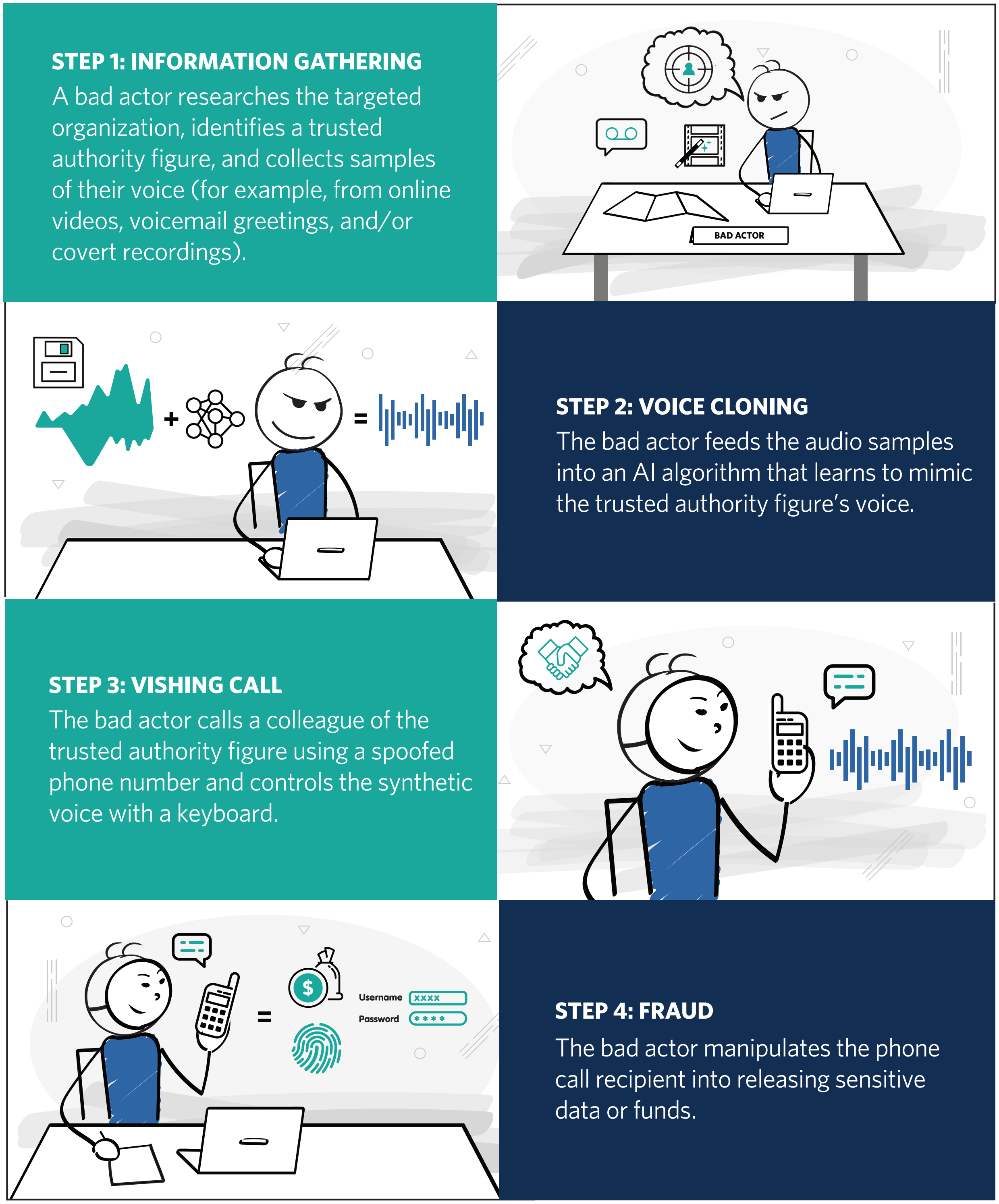
Synthetic media scenario. There are at least two ways that deepfakes might enable identity theft. First, deepfakes could be used in a targeted fashion to steal the identities of individuals. For instance, a phone call made in a victim's synthesized voice could trick her executive assistant or financial adviser into initiating a fraudulent wire transfer.[14] Another form of targeted identity theft would be deepfake audio or video used to create bank accounts under false identities, facilitating money laundering.[15]

Second, deepfakes could facilitate identity theft conducted at scale. Criminals could use deepfakes in social engineering campaigns to gain unauthorized access to large databases of personal information. For example, an e-commerce company official might receive a deepfake phone call—synthesizing the voice of a supervisor or IT administrator—asking for his username and password. In this scenario, deepfakes simply take the place of the initial phishing for these credentials, with the actual identity theft occurring later.

Assessment. These forms of deepfake vishing (voice phishing) are technically feasible today (see figure 1). Current technology enables realistic voice cloning, which can be controlled in real time with keyboard inputs.[16] A leading seller of commercial voice synthesis technology claims that its technology can convincingly clone a person's voice based on just five minutes of original recorded speech and has shown passable results with just one minute of sample audio.[17] Other algorithms can generate very crude cloned voices with as little as three seconds of sample audio.[18]

FIGURE 1

How Deepfake Vishing Works



STEP 1: INFORMATION GATHERING

A bad actor researches the targeted organization, identifies a trusted authority figure, and collects samples of their voice (for example, from online videos, voicemail greetings, and/or covert recordings).

STEP 2: VOICE CLONING

The bad actor feeds the audio samples into an AI algorithm that learns to mimic the trusted authority figure's voice.

STEP 3: VISHING CALL

The bad actor calls a colleague of the trusted authority figure using a spoofed phone number and controls the synthetic voice with a keyboard.

STEP 4: FRAUD

The bad actor manipulates the phone call recipient into releasing sensitive data or funds.

The requirement for only small amounts of sample audio means that the voices of many people—not just prominent, frequently recorded individuals—could theoretically be cloned and used for identity theft or other malicious purposes. Furthermore, as technology continues to advance, the amount of audio needed for accurate voice cloning is likely to shrink further.

Identity thieves might clone someone's voice based on a voicemail greeting or a social media clip of the victim.[19] Or they might call the victim under some pretense and surreptitiously record her voice to synthesize and use in phone calls with others. A malicious insider embedded within a target organization would be particularly well-positioned to capture someone's voice and to know exactly what to say (in that person's voice) to exploit a relationship of trust.

On the other hand, these scenarios seem more complex and labor-intensive than existing criminal techniques. Skilled human impersonators can already mimic voices without the use of AI technologies and attempt to mask any flaws by creating a false sense of urgency.[20]

In addition, today's large-scale identity theft employs cyber techniques that scale more easily than deepfakes. For example, in search of sensitive personal data caches, broad-based email phishing attacks can simultaneously target many employees of a company or even multiple companies. A deepfake phone call, in comparison, must be carefully planned and personally controlled by the perpetrator using keyboard inputs. In theory, deepfake phone calls could be automated using so-called interactive voice response software, but integrating these two technologies would be a complex and novel undertaking.[21]

With identity theft already occurring at such a high volume and efficiency, deepfakes are unlikely to fully displace current tools. However, deepfakes could eventually become powerful, supplementary tools for the most sophisticated financial crimes, particularly those committed by insiders.

Scenario 2: Imposter Scam

Pre-existing threat. Imposter scams are the second most common complaint received by the FTC. Criminals impersonate "the government, a relative in distress, a well-known business, or a technical support expert" to pressure the victim into paying money.[22] Scammers typically make contact by phone, using spoofed phone numbers or voice-over-IP (digital calling) services like Skype to disguise themselves as local callers or even specific people.[23] In other cases, hackers take control of someone's email and then contact friends and family to perpetrate the scam.[24]

After making contact, scammers often manipulate victims by threatening imminent harm unless money is paid—for example, claiming the victim or a loved one faces criminal charges or suspension of government benefits. U.S. residents reported $667 million in losses from imposter scams during 2019.[25]

Synthetic media scenario. Deepfakes could enhance the realism of imposter scams. Scammers might clone the voice of a specific individual, such as a victim's relative or a prominent government official known to many victims. A sophisticated scam operation might involve cloning the voices of all fifty U.S. state governors, robo-dialing many victims, and automatically simulating the appropriate governor's voice based on a victim's area code.

Alternatively, scammers could use AI technology to synthesize a generic voice rather than clone a specific person.[26] Far-away scammers could generate a voice that mimics local accents or that sounds vaguely similar to a well-known actor or celebrity, thereby creating an air of trustworthiness with some victims.

Assessment. Deepfakes could be a boon to skilled scammers that conduct extensive online research to map family relationships and develop convincing vocal impersonations.[27] There are already unverified reports from scam victims who believe their family member's voice was cloned using artificial intelligence.[28] However, these types of scams are labor-intensive and will likely remain less common than more indiscriminate scams.

Deepfake scams that impersonate government leaders would be easier to produce, given the availability of audio recordings to use as training data. However, this would still mean shifting away from better-proven scam strategies. Many of today's scammers claim to be lower-level officials such as law enforcement officers, whose contact with the victim is plausible and whose identities are not readily verified. Scam calls that appear to come from a governor, senator, or president would seem strange and suspicious to some victims. Still, imposter scams do not need to be perfectly cogent. Manipulating the emotions of victims and creating a false sense of urgency help to paper over gaps and inconsistencies, and many scams intentionally target more vulnerable victims, including the elderly and military families.[29]

Scenario 3: Cyber Extortion

Pre-existing threat. In a cyber extortion scheme, criminals claim to have embarrassing information about the victim and threaten to release it unless they are paid or given further sensitive material.[30] The information is often sexual in nature (for example, purported nude images or videos of the victim or alleged evidence of the victim's online pornographic viewing habits).[31]

In some cases, the blackmail material is real, acquired through hacking. But more often, the scheme is a bluff and no compromising information exists.[32] To make the scheme more personalized and convincing, cyber extortionists sometimes reference a password or phone number of the victim in their communications, typically taken from a publicly available data dump.[33] In 2019, U.S. residents reported $107 million in losses from cyber extortion (not including losses from ransomware), according to the Federal Bureau of Investigation.[34]

Synthetic media scenario. Cyber extortionists could use deepfakes to generate more convincing fake blackmail material. For example, blackmailers might send a victim images or videos of her own face synthetically stitched into pornography as purported proof of their access to sensitive material.

Just as today's cyber extortionists harvest data dumps to conduct large-scale email targeting, deepfake extortionists could scrape social media platforms to collect personal images and then automate the production of synthetic blackmail material. The extortionists could use those same social media platforms to contact victims and/or release the damaging images.

Assessment. Aspects of this scenario already happen today, though not yet commonly for the purpose of extortion. Since the term "deepfakes" originated in 2017, nonconsensual synthetic pornography has been the most common type of deepfake.[35] A 2019 study found that 96 percent of online deepfakes are pornographic (and presumed nonconsensual).[36] Pornographic deepfakes have already been used for targeted harassment and character assassination, particularly of women, for personal and political ends.[37]

Using existing technology, deepfake makers have transformed normal photos (with clothed people) into realistic nude simulations.[38] They have also transposed someone's face, as captured in still images, onto another person's face in pornographic videos. Initially, these deepfakes mainly targeted celebrities, because mapping facial patterns required numerous video samples of the victim. Today, technological advances enable the production of full-motion deepfakes based on a single photo of the victim, whose facial movements are made to mimic those of the person in the original video.[39] This means that synthetic pornography could be automated and produced on a mass scale using only one publicly available social media image per victim.

It is easy to imagine how these kinds of deepfakes could facilitate large-scale cyber extortion for profit. While some victims would recognize the blackmail material as fictitious and refuse to pay, other victims may believe the images or feel enough uncertainty to comply with the criminal's demands. Victims might also choose to pay if they fear that others—family members, friends, or coworkers—could believe the images are real.[40] In April 2020, local authorities in India said they had received reports of deepfake cyber extortion.[41]

Nevertheless, conventional methods of cyber extortion appear to be profitable and require little technical skill, so a major shift to using deepfakes is unlikely.[42] A deepfake would be much more technically challenging—requiring software that efficiently integrates photo scraping, deepfake production, and victim contact. But it would likely be more effective on a per-victim basis.

Scenarios Targeting Companies

Scenario 4: Payment Fraud

Pre-existing threat. "Business email compromise" is an umbrella term for various schemes to trick firms into initiating fraudulent payments. Criminals often hack or spoof an email account of a chief executive officer (CEO) and then contact a financial officer to request an urgent wire transfer or gift card purchase.[43] Criminals may also masquerade as trusted suppliers (using false invoices) or employees (diverting direct deposits).[44] Complex cases involve weeks or even months of priming victims through phone calls and emails.[45]

In 2019, U.S. businesses reported more than $1.7 billion in losses from this type of fraud—nearly half of the reported total loss from all cyber crimes, according to the Federal Bureau of Investigation.[46]

Synthetic media scenario. Deepfakes could make phone calls used in business email compromise schemes sound more authentic. In fact, a convincing deepfake vishing call could eliminate the need for email hacking or spoofing in some cases. Unwitting recipients might find deepfake video calls even more compelling than audio calls.[47]

Assessment. The use of deepfakes to commit fraud has already been documented on a small scale. Last year, criminals apparently used voice cloning technology to impersonate a German CEO and successfully trick his British subordinate into sending a $243,000 wire transfer.[48] The subordinate "recognized his boss' slight German accent and the melody of his voice on the phone." Commercial software was likely used in the crime, according to the victim's insurance company.

A more ambitious scheme could incorporate deepfakes into live video calls, adding another layer of persuasion. Current technology could enable a criminal to swap one face with another on the fly during a video call.[49] And because video calls often have poor image quality, flaws in the deepfake might go unnoticed or overlooked.

Scenario 5: Stock Manipulation via Fabricated Events

Pre-existing threat. The internet provides multiple ways for disinformation campaigns to manipulate stock prices.[50] Anonymous bad actors frequently spread false or misleading claims about a targeted stock via blogs, forums, social media, bot networks, or spam.[51] These campaigns seek to either artificially increase the stock's price (a "pump and dump" scheme) or lower it (a "short and distort" or "poop and scoop" scheme) for quick profit.[52] Small companies have been the most common targets as small-cap stocks can be more easily manipulated. However, large companies have sometimes been victims of sophisticated disinformation campaigns, which may have political as well as financial motives.[53]

Synthetic media scenario. Deepfakes could lower a company's stock price by generating seemingly credible false narratives, perhaps by fabricating the private remarks of a corporate leader. For example, a bad actor could release a deepfake video that portrays a targeted CEO supposedly acknowledging her company's insolvency, committing or confessing to misconduct, or making highly offensive comments (see figure 2).[54]

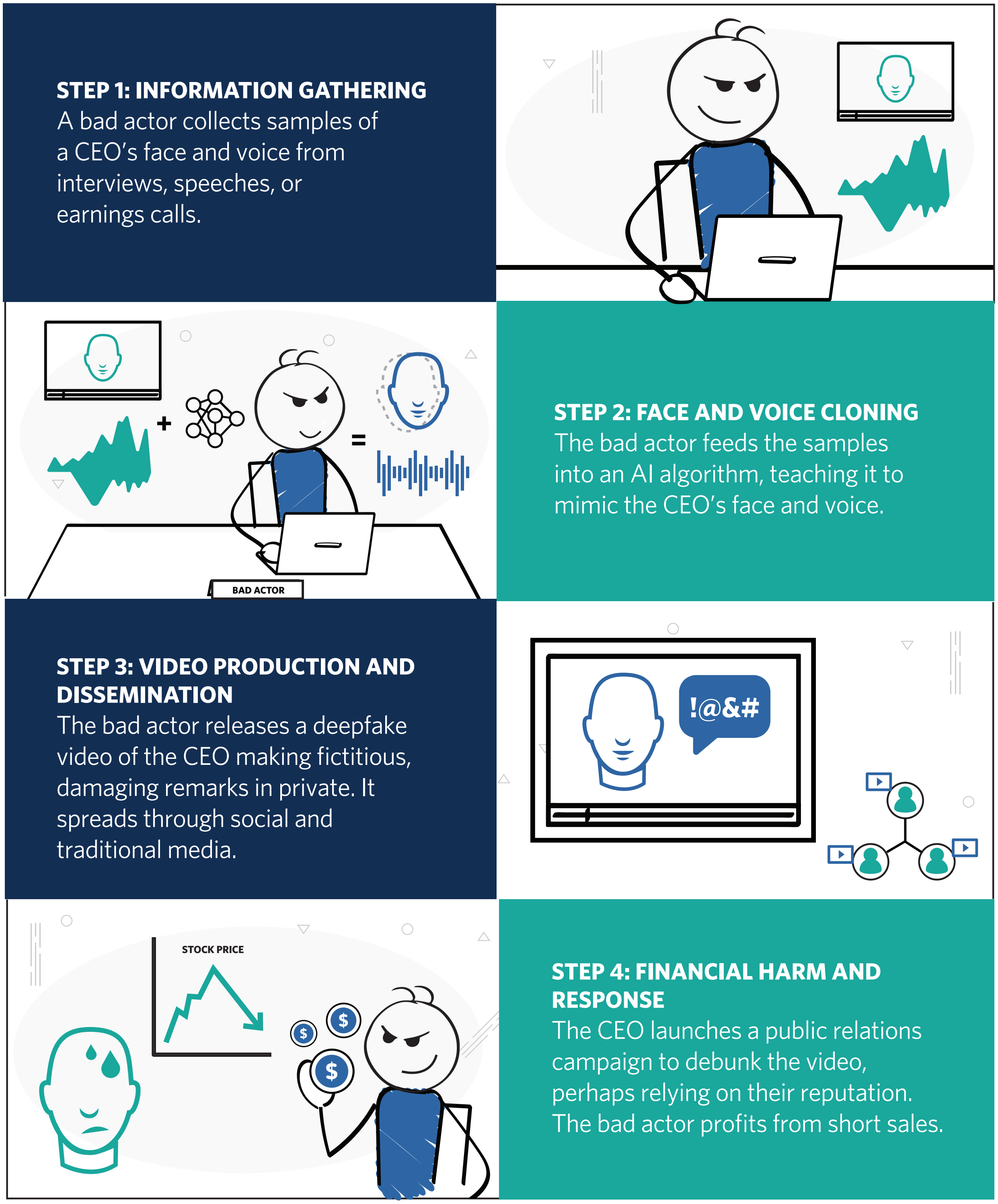
Alternatively, deepfakes could be designed to raise a company's stock price by fabricating positive events. For example, a bad actor could use deepfake videos to falsely portray celebrities or politicians endorsing or using a product.

Assessment. A well-crafted deepfake shared through social media or spam networks could be effective in manipulating small-cap stocks. Smaller companies often lack sufficient resources and goodwill to mount a rapid, persuasive self-defense against short and distort schemes.[55] Even if a deepfake is quickly debunked, perpetrators could still profit from short-term trades.

Deepfakes might also represent a new vulnerability for large companies, whose stock prices are traditionally more resistant to manipulation. Highly visible corporate leaders generate large volumes of media interviews, earnings calls, and other publicly available recordings. These would enable bad actors to produce relatively accurate deepfakes.

FIGURE 2

How Deepfake Fabrication of Private Remarks Works



STEP 1: INFORMATION GATHERING
A bad actor collects samples of a CEO's face and voice from interviews, speeches, or earnings calls.

STEP 2: FACE AND VOICE CLONING
The bad actor feeds the samples into an AI algorithm, teaching it to mimic the CEO's face and voice.

STEP 3: VIDEO PRODUCTION AND DISSEMINATION
The bad actor releases a deepfake video of the CEO making fictitious, damaging remarks in private. It spreads through social and traditional media.

STEP 4: FINANCIAL HARM AND RESPONSE
The CEO launches a public relations campaign to debunk the video, perhaps relying on their reputation. The bad actor profits from short sales.

An especially damaging scenario would involve the fabrication of private remarks—for example, a synthesized recording of a CEO purportedly using sexist language.[56] Definitively proving that a private conversation never took place might be impossible; instead, the CEO may need to rely on his reputation to manage the fallout. Highly trusted CEOs would be well-positioned to rebuff a false recording. But a CEO with prior credibility problems would face a much more challenging situation.[57] Truth could also begin mixing with fiction, further influencing the market—for example, a deepfake recording of sexist remarks could inspire real women to come forward with sincere cases of discrimination.

Even if such a deepfake could be authoritatively disproven, it would likely still have long-term consequences for a company's reputation. As with other forms of misinformation, deepfakes can leave lasting psychological impressions on some viewers even after being debunked.[58] Experiments have shown a substantial minority of people will believe a deepfake is real despite explicit warnings that it is fake.[59] Long-term loss of goodwill could reduce revenue and stock price over time, especially for consumer-facing companies.

Scenario 6: Stock Manipulation via Bots

Pre-existing threat. In the previous scenario, stock prices are manipulated using portrayals of fictitious remarks or events that influence investor or consumer attitudes. But stock prices can also be manipulated through generating false pictures of mass sentiment—for example, by manufacturing evidence of a grassroots backlash against a brand on social media.

Social media bots are already used for this purpose. Bad actors craft large numbers of fake personas on a platform and then coordinate mass postings that promote or denigrate specific companies.[60] Spikes in such bot activity have been tied to small, temporary changes in targeted stock prices.[61]

Of course, much of this activity violates social media platforms' policies against spam or false personas. And platforms can look for characteristic hallmarks to identify and remove illicit bots. Possible red flags include the recent creation of numerous similarly focused accounts, stolen or stock profile photos, inexplicably frequent posting, and inconsistent personality traits or biographical data.[62] Platforms also use machine learning to identify subtler behavioral patterns and detect spam bots at scale.[63]

Still, illicit social media bots remain endemic for several reasons. First, not all automated social media accounts actually violate platform policies. Distinguishing harmful bots from others is as much art as it is science.[64] Second, malicious bot behavior occurs on a vast scale and is constantly evolving. Third, algorithmic techniques for identifying bots remain imperfect. They may misidentify bots as human (allowing them to escape removal) or misidentify humans as bots (inhibiting platforms from using such algorithms aggressively).[65] Fourth, large-scale account purges can frighten investors (who want to see strong user growth) or alienate prominent users (who want to retain a high follower count).[66] In short, identifying and removing illicit bots is already a difficult and fraught task, even when platforms have access to sophisticated AI algorithms and bad actors do not.

Synthetic media scenario. Although no cases have yet been publicly documented, deep learning could theoretically be used to create AI-driven synthetic social botnets that better evade detection and improve persuasion. The end goal would remain the same: depicting false trends in sentiment about a company, thereby driving stock prices higher or lower. A bad actor could seek to concoct a wholly false shift in sentiment, or more likely, amplify an emergent trend.

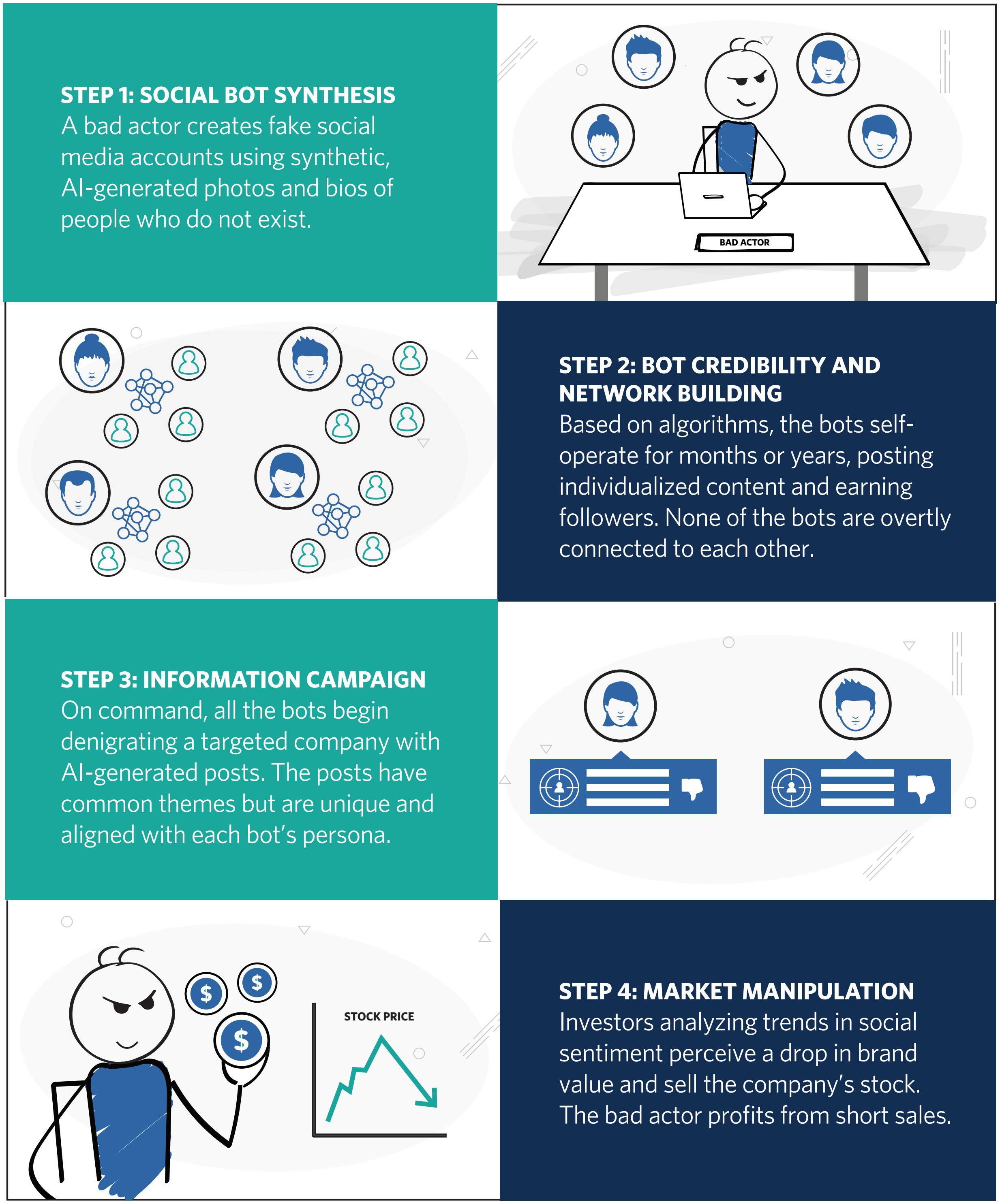
Bad actors have already begun using AI-generated profile photos that depict people who do not exist, thwarting efforts to spot picture reuse.[67] A few sophisticated influence campaigns, carried out by unscrupulous media companies and suspected intelligence operatives, have employed this technique.[68] The next step would be for algorithms to author synthetic posts (see figure 3).

Whereas traditional bots make duplicative or crudely randomized posts, synthetic social bots could publish novel, individualized content.[69] Conscious of their prior postings, they could maintain consistent personalities, writing styles, subject matter interests, and biographies over time. The most convincing synthetic bots would earn organic human followings, increasing the impact of their messages and making them harder to detect. Traditional bots often follow each other to create the appearance of influence, which leads to telltale clustering patterns.[70]

Synthetic social bots could operate with minimal supervision for months or even years to build credibility and clout. When called upon by the bad actor that created them, each bot in the hidden network could then start posting about a targeted company. Each bot would use unique language and storytelling, consistent with its persona. The overall campaign could appear to represent grassroots consumer sentiment and thus influence stock price. For example, the bots might all claim to have contracted food-borne illnesses at the same fast food chain, or they might feign shared outrage at a company's recent advertising campaign.[71]

The social media profile and post shown in figure 4 are just mock-ups created for this paper, but their individual components were generated with artificial intelligence to demonstrate what is already possible. The profile photo,[72] banner photo,[73] bio text, and post text[74] were all synthesized using publicly available, web-based AI tools.[75] The profile name was chosen by a random generator.[76]

FIGURE 3
How Synthetic Social Botnets Could Work



STEP 1: SOCIAL BOT SYNTHESIS
A bad actor creates fake social media accounts using synthetic, AI-generated photos and bios of people who do not exist.

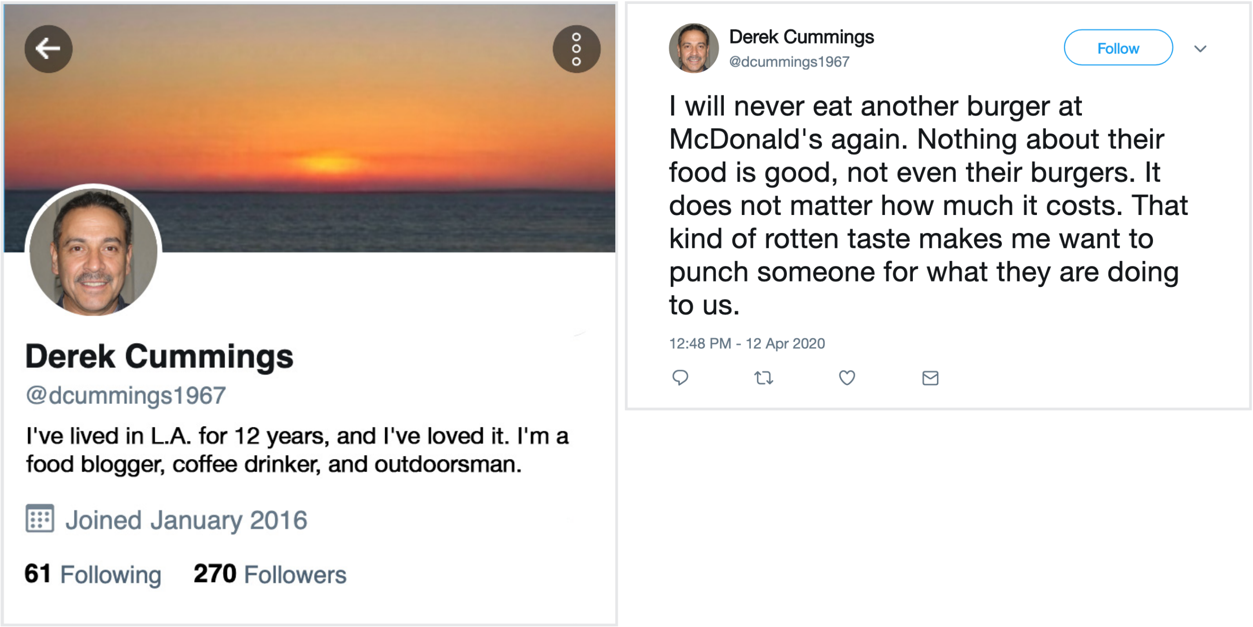
STEP 2: BOT CREDIBILITY AND NETWORK BUILDING
Based on algorithms, the bots self-operate for months or years, posting individualized content and earning followers. None of the bots are overtly connected to each other.

STEP 3: INFORMATION CAMPAIGN
On command, all the bots begin denigrating a targeted company with AI-generated posts. The posts have common themes but are unique and aligned with each bot’s persona.

STEP 4: MARKET MANIPULATION
Investors analyzing trends in social sentiment perceive a drop in brand value and sell the company’s stock. The bad actor profits from short sales.

FIGURE 4

Demonstration of a Synthetic Social Bot



Derek Cummings

@dcummings1967

I've lived in L.A. for 12 years, and I've loved it. I'm a food blogger, coffee drinker, and outdoorsman.

Joined January 2016

61 Following 270 Followers

Derek Cummings

@dcummings1967

I will never eat another burger at McDonald's again. Nothing about their food is good, not even their burgers. It does not matter how much it costs. That kind of rotten taste makes me want to punch someone for what they are doing to us.

12:48 PM - 12 Apr 2020

Assessment. Tools currently exist to help detect AI-generated photographs and text, the basic building blocks of synthetic social bots.[77] Moreover, social media platforms can often detect bots based on their behavior and interrelationships, not just their individually posted content.[78] However, bot detection remains imperfect, even with today's relatively crude bots. No platform has successfully eliminated illicit bots, and platforms must continuously innovate to keep up with evolutions in bot technology.[79] Synthetic social bots would represent another leap forward by bad actors, challenging platforms to advance the science and art of bot detection. Tools for abuse and detection will likely continue to co-evolve, and who has the overall advantage could shift repeatedly over time.

For stock manipulators, synthetic social bots could exploit investors' desire to analyze social media activity for trends in consumer sentiment. An increasing number of fintech companies market "social sentiment" tools that analyze what social media users say about companies.[80] In some cases, social sentiment data have been integrated with automated trading algorithms—empowering computers to execute trades, without human intervention, based on apparent trends in social media activity.[81] More widespread use of social sentiment analysis and its greater integration with automated trading algorithms could enhance the power of synthetic social bots to manipulate stock prices.[82]

Scenario 7: Malicious Bank Run

Pre-existing threat. Individual banks in several countries have experienced runs that were partly driven by social media rumors of financial weakness.[83] Social media itself is not a principal cause of bank runs; online rumors about bank weakness are usually responses to genuine problems in the banking sector, though the rumors sometimes outpace reality.[84] Social media, alongside traditional media and word of mouth, provides a venue for such rumors to circulate. When public doubts about a bank's health start to become widespread, social media can rapidly amplify them.

The sources of bank rumors, and their ultimate basis in fact, can be difficult to determine. In 2014, the Bulgarian government blamed unspecified parties for coordinating a "criminal attack" on the reputation of several banks that had suffered runs. The episode involved text messages, internet posts, and media leaks—some factual, others contested.[85] Incidents like this, whether malicious or not, suggest a template that bad actors might follow in the future.

Synthetic media scenario. A synthetic social botnet, described in the previous scenario, could be used to foment or intensify rumors that drive bank runs. Alternatively, a deepfake video released on social media could depict a bank executive or government official describing severe liquidity problems. Any successful deepfake-driven bank run would likely occur during times of trouble in a country's financial system.

Assessment. While a synthetic social botnet might be more effective in fomenting bank fears than a traditional social botnet, bad actors might avoid botnets and synthetic media altogether. One of the most enduring and effective forms of online deception is also one of the simplest: pairing real images and videos with false or misleading captions.[86] For example, bad actors might spread years-old photos of long lines at another bank in a different country and claim that they represent current local activity.

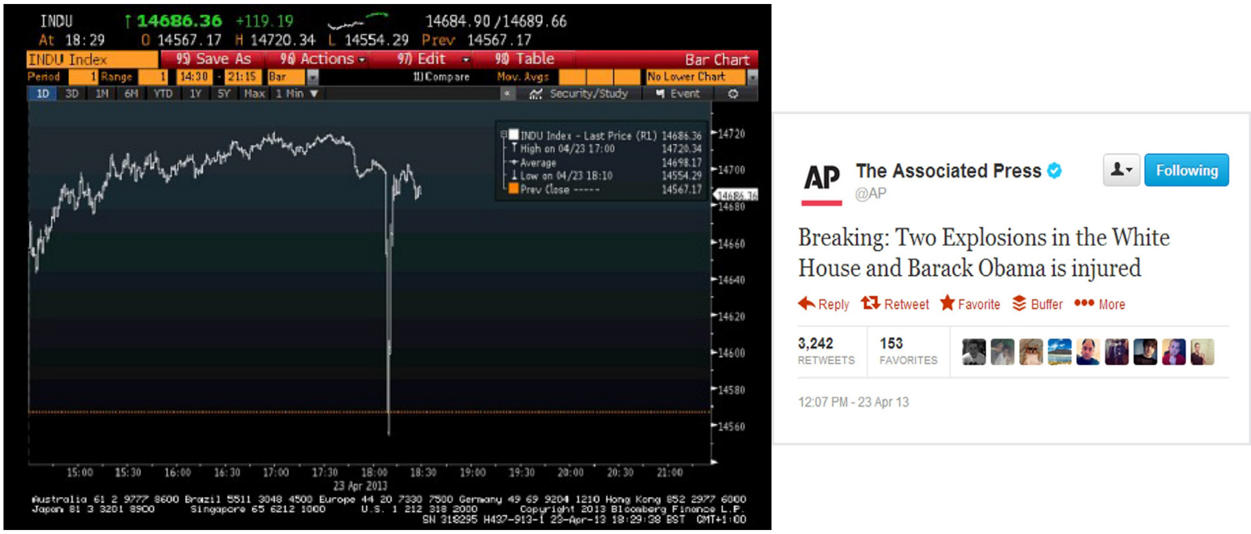
Scenarios Targeting Financial Markets

Scenario 8: Malicious Flash Crash

Pre-existing threat. On April 23, 2013, a state-sponsored hacking group called the Syrian Electronic Army hijacked the Twitter account of the Associated Press and then tweeted, "Breaking: Two Explosions in the White House and Barack Obama is injured." This false claim triggered an instant deluge

FIGURE 5

The Syrian Electronic Army's Malicious Flash Crash


INDU ↑14686.36 +119.19 14684.90/14689.66
At 18:29 O 14567.17 H 14720.34 L 14554.29 Prev 14567.17
INDU Index
95) Save As
96) Actions
97) Edit
98) Table
Bar Chart
Period 1 Range 1 14:30 - 21:15 Bar
11) Compare
Mov. Avgs
No Lower Chart
1D 3D 1M 6M YTD 1Y 5Y Max 1 Min
Security/Study
Event
INDU Index - Last Price (R1) 14686.36
High on 04/23 17:00 14720.34
Average 14698.17
Low on 04/23 18:10 14554.29
Prev Close ----- 14567.17
14720
14700
14686.36
14680
14660
14640
14620
14600
14580
14560
15:00
15:30
16:00
16:30
17:00
17:30
18:00
18:30
19:00
19:30
20:00
20:30
21:00
23 Apr 2013
Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2013 Bloomberg Finance L.P.
SN 318295 H437-913-1 23-Apr-13 18:29:38 BST GMT+1:00
AP
The Associated Press
@AP
Following
Reply
Retweet
Favorite
Buffer
More
3,242 RETWEETS
153 FAVORITES
12:07 PM - 23 Apr 13

Breaking: Two Explosions in the White House and Barack Obama is injured

of trading, which E-Trade called "the most active two minutes in stock market history."[87] Automated trading algorithms drove much of the volume.

In just three minutes, the S&P 500 index lost $136 billion in value, and crude oil prices and Treasury bond yields also fell.[88] But the shock ended as quickly as it began. Markets fully recovered after another three minutes (see figure 5).[89]

Synthetic media scenario. Politically or financially motivated actors could attempt something similar using deepfakes. For example, a synthetic recording of Saudi and Russian oil ministers haggling over production quotas could be released online in a bid to influence oil prices and the wider stock market.

Some clips could be quickly debunked—for instance, if the depicted person is widely trusted and issued a prompt denial, ideally backed by hard evidence. But leaders who already face trust deficits could find it harder to quash suspicions caused by a deepfake, potentially allowing market impacts to linger.

For example, imagine that during the height of U.S. President Donald Trump's impeachment trial, someone had "leaked" an authentic-sounding, synthetic audio clip of the key phone call between Trump and Ukrainian President Volodymyr Zelensky. The recording would have closely mirrored the published transcript, except that several new passages would have been inserted to convey a damning quid pro quo. Such a deepfake might have spread for hours or days until being fully disproven. Meanwhile, financial markets would respond to the heightened political uncertainty.

Assessment. A persuasive deepfake could have a more lasting effect than the simple Twitter hijacking perpetrated by the Syrian Electronic Army. Deepfake videos in particular benefit from the "picture superiority effect," a psychological bias toward believing and remembering visual images more than other types of data.[90] Deepfakes can also aim to spread disinformation organically through social and traditional media, obviating the need to hack an influential news account to publicize the false claim. However, creating an effective deepfake of this kind would require great political sophistication as well as technical skill. It would not be easy to craft a fake scene capable of fooling informed observers and moving entire markets.

It is worth noting that the Syrian Electronic Army's Twitter hijack was the high watermark of modern, market-moving disinformation—and in the seven years since, the United States has suffered no other similar event. High-frequency trading, which fueled the crash, has declined in overall volume.[91] Meanwhile, savvy traders and other market watchers have become more wary of breaking news claims.[92] Still, another flash crash remains possible.[93] Even a short-lived crash could enable bad actors to profit from well-timed trades and inflict lasting psychological impacts.

Scenarios Targeting Central Banks and Financial Regulators

Scenario 9: Fabricated Government Action

Pre-existing threat. Central banks and financial regulators around the world have battled rumors—many circulating online—about imminent market-moving actions.[94] In 2019, the central banks of India and Myanmar each sought to quash social media rumors that authorities would soon close certain commercial banks.[95] In 2010, false claims spread online that China's central bank governor had defected, briefly spooking short-term lending markets.[96] In 2000, U.S. stocks slid for several hours amid false rumors that the Federal Reserve chair had suffered a car accident.[97]

Synthetic media scenario. Deepfakes could be used to concoct recordings of central bankers privately discussing future interest rate changes, liquidity problems, or enforcement actions. For example, a "leaked" audio clip of a fabricated central bank meeting could depict officials fretting over inflation and making plans to increase interest rates.

Deepfakes might also target central bankers or financial regulators as individuals, perhaps for political purposes. A regulator could be shown accepting a bribe from a business leader to drop a corruption investigation, for example.

Assessment. These deepfakes would likely have a greater impact in countries where financial oversight mechanisms are already less trusted. Trust is critical for effectively debunking a deepfake, especially when the victim must prove a negative—that a purported private conversation never happened. In times of financial crisis, deepfakes could exploit and amplify pre-existing economic fears.

Even in major, stable economies, central banks and financial authorities are often criticized for unclear, obtuse, or slow public communication.[98] A botched governmental response to a deepfake would lengthen the time window during which bad actors can sow chaos and profit from short-term trades.[99]

Scenario 10: Regulatory Astroturfing

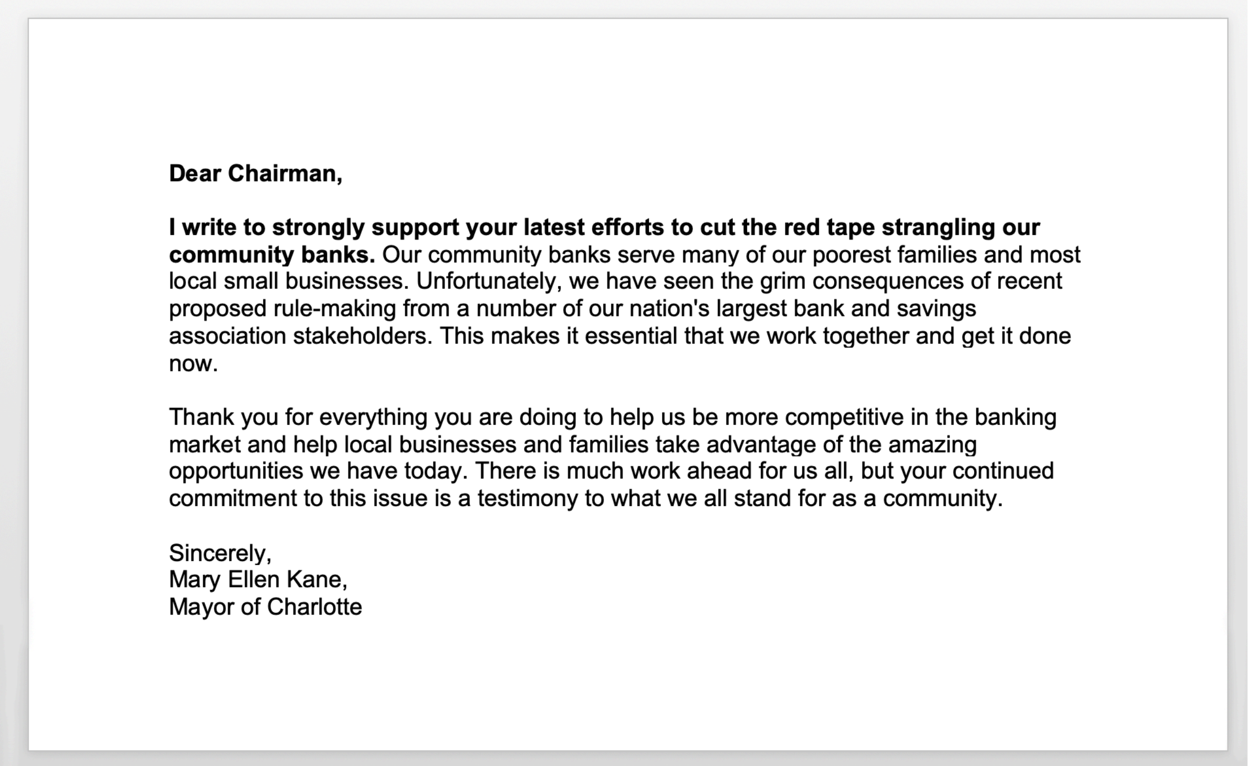
Pre-existing threat. Regulators, including those overseeing the financial industry, must increasingly deal with digital astroturfing—covert attempts to manipulate policymaking by creating the false appearance of grassroots support for certain positions. The U.S. Securities Exchange Commission and Consumer Financial Protection Bureau have both experienced large-scale abuse of online systems for submitting public comments on proposed regulations.

In several cases, lobbyists and activists have submitted huge volumes of comments in the names of unwitting, fictitious, or deceased people.[100] Many comments are sent on behalf of compromised or phony email accounts.[101] Some comments are carbon copies, while others are partially randomized using computer software in a bit to appear unique.[102]

Synthetic media scenario. AI-generated content could make digital astroturfing appear more authentic. Algorithms that generate synthetic text can produce any amount of writing on any topic.[103] Astroturfers could use this technique to create thousands or millions of fake comments that oppose or support a specific financial regulation. The comments would vary in content and language, avoiding crude copy-and-paste or dictionary-randomizing techniques that have exposed traditional false campaigns.

The letter shown in figure 6 is only an illustration for this paper, but it was generated using artificial intelligence to demonstrate what is already possible. The bolded portion of the letter was a human-written prompt. The remainder was written by a publicly available, web-based AI text generator in response to the prompt.[104] The result is highly realistic, except for one obvious flaw: the mayor of Charlotte, North Carolina, is not named Mary Ellen Kane.

FIGURE 6
Demonstration of Synthetic Astroturfing



Dear Chairman,

I write to strongly support your latest efforts to cut the red tape strangling our community banks. Our community banks serve many of our poorest families and most local small businesses. Unfortunately, we have seen the grim consequences of recent proposed rule-making from a number of our nation's largest bank and savings association stakeholders. This makes it essential that we work together and get it done now.

Thank you for everything you are doing to help us be more competitive in the banking market and help local businesses and families take advantage of the amazing opportunities we have today. There is much work ahead for us all, but your continued commitment to this issue is a testimony to what we all stand for as a community.

Sincerely,
Mary Ellen Kane,
Mayor of Charlotte

Assessment. Today's AI-generated text varies in quality and apparent authenticity.[105] It can be incoherent or uncanny. Synthesized text can be greatly enhanced by tailoring the algorithm to write in specific genres and styles. For example, an astroturfer might feed large amounts of prior regulatory comments (which are publicly available) into the deep learning algorithm, helping its output mimic the style and substance of authentic submissions.

A Harvard University student successfully tested this approach in 2019. Using previous regulatory comments as training data, he synthesized 1,001 comments on a real proposed rule.[106] The AI-generated comments were of high quality and expressed a diverse set of arguments. Humans asked to review both the synthetic and real comments were unable to distinguish between them. Notably, the research cost less than $100 and was performed by a college senior and self-described "novice coder" using an "an older, everyday model HP laptop."[107]

However, tools do exist to help distinguish AI-generated text from human-written text.[108] Human writings tend to include creative and unexpected word choices, whereas synthetic text follows more predictable statistical patterns and algorithms can be trained to spot the difference. Longer bodies of text, like regulatory submissions, might allow these detection algorithms to make more confident judgments. Even so, detection algorithms are not foolproof and a dedicated adversary would seek to counter them.[109] For example, synthetic text generators could be designed to produce more irregular, human-like output.

Moreover, synthetic text detection tools will only work if they are implemented effectively. It is therefore worrying that U.S. agencies have not yet implemented much more basic forms of comment authentication. A 2019 U.S. Senate report found that none of the fourteen agencies it surveyed used CAPTCHAs, or indeed any technology, to verify that public commenters are real people.[110] In the Harvard experiment, all 1,001 synthetic comments were successfully submitted to the agency. The AI-generated comments, before being voluntarily withdrawn, made up a majority of the 1,810 comments the agency had received.[111]

The bottom-line impact of digital astroturfing, whether AI-generated or otherwise, is debatable. In practice, agencies give much more weight to comments from well-known businesses and organizations than to those from average individuals.[112] Still, digital astroturfing is a growing regulatory problem that saps public confidence in the rulemaking process and could have legal or political consequences.[113] Synthetic media appears to be a powerful new method for digital astrourfing.

Policy Implications

Overall Threat to Financial Stability and the Macroeconomy

None of the scenarios depict a serious threat to the stability of the global financial system or national markets in mature, healthy economies. Major markets seem generally resilient to disinformation campaigns, regardless of the technique used. Prior to the invention of deepfakes, misinformation sometimes had market-wide effects—but instances have been rare in recent decades, and they resulted in only limited, short-term swings. To threaten market stability, synthetic media would need to be orders of magnitude more powerful than traditional disinformation tools. There is no reason yet to expect that.

More likely, synthetic media will inflict financial harm on targeted individuals and businesses. This harm could be significant from the victim's perspective—as is the case with existing forms of fraud, extortion, and stock manipulation. But the aggregate financial harm from all synthetic media schemes is very unlikely to have a macroeconomic impact. Financial wrongdoing is already a constant, even in advanced economies. To have macroeconomic consequences, the aggregate harm from synthetic media abuse would need to surpass most other forms of illicit activity—which almost certainly will not happen.

However, emerging markets face greater threats from synthetic media. Countries with shakier economies and less-trusted institutions already struggle more with financial disinformation; synthetic media could exacerbate this problem. Developed countries experiencing financial crises could also be more vulnerable. For example, Scenario 7 explores how synthetic media could contribute to bank runs. If a banking system is already suspect, deepfakes could play upon widespread fears by falsely "revealing" liquidity shortfalls at ailing banks. Debunking the deepfake would be harder where bankers and government regulators lack public credibility.

Synthetic Media Versus Other Malicious Tools

Deepfakes and synthetic media may be powerful, but they are also less proven and more technically complex than some other criminal techniques. The extent of synthetic media abuse will depend largely on its cost-effectiveness; bad actors will weigh the required inputs (generating and using synthetic media) with the resulting outputs (monetary gains) and then compare this equation against other illicit methods. Criminals gravitate toward the most profitable tools, and even state actors are cost-conscious. The ultimate question for many bad actors will be how deepfakes size up against "cheapfakes" or traditional forms of media manipulation.

Cheapfakes have proven highly effective, including in the financial sphere, and show no signs of obsolescence (see box 2). Each of this paper's ten synthetic media scenarios is paired with a description of how bad actors already carry out financial deception using pre-AI technologies. In some cases, pre-existing tools are very profitable—as with identity theft (Scenario 1), which occurs on an industrial scale. The continued prevalence of manipulated media and other relatively simple tricks indicates that technical wizardry is neither necessary nor sufficient for deception. Even as AI technology advances, many bad actors will find manipulated media to be more cost-effective and simpler to produce.

BOX 2

What Are Cheapfakes and Manipulated Media?

Cheapfakes (or shallowfakes) are traditional forms of deceptive media that do not rely on AI.

The term is a retronym, coined to differentiate deepfakes and synthetic media from everything that came before. It serves as an implicit reminder that bad actors have long manipulated video, doctored photographs, distorted sound, and forged documents. These activities predate the AI era and indeed the digital age. "Cheapfake," like "deepfake," is ill-defined slang. This paper uses **manipulated media** to describe the broad category of all nonsynthetic (non-AI generated) deceptive media.

Manipulated media remains far more widespread than synthetic media. Common media manipulation techniques include mislabeling a real video or image, isolating a short clip from its surrounding context, chopping and resplicing recordings into a new sequence, speeding up or slowing down a clip, and altering individual images or frames.[114] Many techniques are low-tech or no-tech, such as paper document forgery, human voice impersonation, and faux film footage of props or body doubles.[115]

Still, synthetic media has important advantages. First, it can be highly realistic—in some cases, uniquely so. Without deepfake technology, producing a face-swap video would require Hollywood-level postproduction capabilities. Second, it can scale. For example, algorithms can produce infinite numbers of synthetic photos on demand, which no human could do. Third, it can be custom-tailored. Synthetic text generators can quickly learn new genres or writing styles, whereas humans might need years of experience to do so. Synthetic media will sometimes be the stronger, more cost-effective option.

A bad actor's choice between synthetic and manipulated media will depend on both his strategy and his capabilities. Some financial schemes, by their nature, make one option more practical than another. For example, indiscriminate "grandparent scams" target thousands of senior citizens with bogus calls from a purported grandchild who needs fast cash.[116] This style of scam often rewards quantity over quality; many fraudsters improvise during calls instead of doing prior research on the victim. Deepfakes would be neither feasible nor profitable for such scattershot scams. However, more sophisticated versions of this crime also exist. Some imposter scammers perform substantial research to tailor their scheme in the hopes of coercing larger payoffs from each victim. Deepfakes would support this and other relatively complex criminal strategies.

In some cases, the bad actor's skillset will be a determining factor. For instance, a criminal with strong language and social skills might opt to use traditional vishing techniques to defraud a business over the phone. But a less charismatic, more technically oriented criminal could prefer to use a cloned voice rather than his own. In essence, this would mean outsourcing key aspects of the impersonation to a computer—gaining the benefit of a skilled, customizable impersonator without having to pay wages or divide up the illicit gains.

Amplification of Existing Narratives and Crises

Broadcast synthetic media will tend to be most powerful when it amplifies pre-existing narratives. For example, Scenario 6 envisions a synthetic social botnet that falsely simulates grassroots consumer backlash against a targeted brand, lowering its stock price. If a genuine consumer backlash against a targeted brand is already underway, synthetic social botnets could magnify its apparent extent while blending in with authentic social media activity. And the disinformation campaigns would be harder to combat because of the company's diminished credibility and reduced corporate bandwidth.

Companies, financial institutions, and government regulators facing public relations crises are especially vulnerable to deepfakes and synthetic media. As part of planning for and managing crises of all kinds, organizations should consider the possibility of synthetic media attacks emerging to amplify the crisis. Steps taken in advance—for example, building trust with key audiences and curating evidence to counter potential false narratives—could help mitigate the damage.

Comparison With Political Deepfakes

The extensive commentary on political deepfakes has helped shape perceptions of deepfakes and synthetic media more generally. First, with election integrity framed as the central issue, deepfakes are typically imagined as targeting mass opinion and spreading through social and traditional media. Second, legal responses to deepfakes are not viewed as promising—partly because there are legal protections for political speech and because legal processes move more slowly than election cycles.[117] Third, deepfakes in the political context trigger serious philosophical dilemmas. How can we distinguish legitimate satire and activism from illegitimate deception? Who should be allowed to draw and enforce these lines? Can political actors overcome distrust to agree on common standards of discourse?[118]

The analysis of financial deepfakes in this paper complicates this prevailing picture of deepfakes and synthetic media. As in the political arena, financial deepfakes will sometimes manipulate the public at large—but multiple scenarios involve the targeting of private entities through direct, point-to-point communication. Almost all these scenarios are blatantly illegal, creating opportunities for criminal, civil, and administrative action. Legal responses would still face challenges, such as identifying and pursuing shadowy, far-away perpetrators. But there might be some hope of recouping ill-gotten gains—in contrast to elections, which are rarely overturned. Finally, societies should be able to find common ground on combatting the abuse of synthetic media for financial crime and harm. For example, social media platforms are unlikely to face calls of censorship for banning deepfakes that falsify celebrity endorsements of commercial products.

Where there are similarities between political and financial threats from synthetic media, a common set of policy responses can perform double duty. Where there are differences, the financial sector should ensure that tech platforms, governments, and others involved in countering synthetic media do not overlook or neglect financial scenarios. Meanwhile, policymakers in both areas should learn from each other's successes and failures. The financial system can move faster in some areas (for example, law enforcement) while the political system leads in others (for example, public education).

A Diverse, Multistakeholder Response

The ten scenarios share a common general sequence of events. Each scenario begins with bad actors producing synthetic media. Next they disseminate the media to a target audience. Someone then views that media and responds—for example, by believing and acting upon a deepfake. Finally, an indirect victim affected by the synthetic media may also seek to respond—for example, the person depicted in a deepfake seeks to refute the narrative.

Although detailed policy prescriptions are outside the scope of this paper, this sequence of events could provide a basic framework for organizing policy responses. Each event engages different stakeholders and therefore offers distinct opportunities for policy responses (see table 2 for a simple overview). Of course, specific interventions each have costs and limitations as well as potential benefits. Many of these interventions are already being explored or could draw on existing measures for countering financial harm and disinformation.

TABLE 2

Notional Policy Interventions and Stakeholders

Events	1. Synthetic media production	2. Synthetic media dissemination	3. Viewer response	4. Victim response
Notional interventions	• AI research and development controls/ethics • AI dissemination controls/ethics	• Synthetic media detection • Content moderation • Identity authentication • Antispoofing • Antispam • Bot policy enforcement • Fact checking • Information and intelligence sharing • Legal action	• Public education • Organizational controls/training • Financial institution controls/training	• Prior trust building • Prior mitigation and recovery planning • Curation of evidence to counter false narrative • Public relations campaign
Policy stakeholders	• AI researchers • AI developers • AI technology investors	• Social media platforms • Traditional media outlets • Phone/voice-over-IP providers • Video call services • Email providers • Intelligence agencies • Law enforcement and financial regulators	• General public • Businesses • Financial institutions • Governments • Civil society	• Businesses • Financial institutions • Central banks and financial regulators • Journalists

This analysis confirms the results of Carnegie's earlier research on synthetic media and election integrity, which found that no single solution or stakeholder can fully address the challenge.[119] Successful efforts will require new technologies, organizational practices, and societal changes. The financial sector, then, should consider its place in a broader ecosystem of synthetic media policy.

Again, this paper does not offer a detailed policy road map. But generally, financial institutions and regulators should divide their efforts into three complementary tracks:

1. Internal action, such as training and controls (for example, reevaluating methods of customer authentication that rely on voice or face)

2. Industry-wide action, such as information sharing (for example, expanding cyber intelligence sharing mechanisms to encompass synthetic media schemes)

3. Multistakeholder action with key outside entities (for example, tech platforms, AI researchers, journalists, civil society, and government bodies

Whether through formal partnerships or informal dialogues, the financial sector should look for ways to voice its unique concerns to other relevant stakeholders, stay informed about outside activities, and facilitate collaboration. For example, social media platforms are still debating the best policies for moderating synthetic media content; the platforms should therefore be fully apprised of the scenarios and malicious techniques most concerning to financial institutions.[120]

Conclusion

The scenarios explored in this paper suggest several policy lessons. Synthetic media is unlikely to threaten global financial stability or the macroeconomy—but individual people, companies, and government institutions are vulnerable, as are emerging economies and financial systems under stress. Synthetic media joins a long list of malicious tools; sophisticated bad actors are more likely to incorporate synthetic media, while others may continue relying on older techniques. Financial threats from synthetic media are both similar to and different from political threats, offering lessons for stakeholders in both realms.

Policymakers concerned about deepfakes and synthetic media face significant uncertainty. On the one hand, this technology presents obvious concerns. Synthetic media can be more realistic, scalable, and custom-tailored than traditional forms of deception. Financial criminals and other bad actors have historically sought out advanced technology, and they will doubtless explore the use of synthetic media. In fact, as noted earlier, aspects of a few scenarios have already been documented in the last few months. If the financial system does not act now, it could lose valuable time in the innovation race against bad actors. It could be years before technical investments, public-private partnerships, and other policy efforts come to fruition.

On the other hand, it may be difficult to justify devoting scarce resources to a largely theoretical problem. Synthetic media use has not yet facilitated widespread financial harm; none of the scenarios in this paper have manifested at scale. From a policymaker's perspective, synthetic media is just one of many technology-based risks to the financial system. Most financial institutions and other organizations have a long list of unmet technology needs. Some of this "technical debt"—such as insecure computer networks—can be tied to ongoing, measurable financial losses, including from conventional criminal schemes, whereas most harm from deepfakes remains speculative.[121]

This is a classic risk management dilemma. In an ideal world, policy interventions are grounded in reliable risk modeling and return-on-investment estimates. But with emerging technologies like synthetic media, sufficient data do not yet exist and will likely come too late. To deal with this dilemma, financial stakeholders should take an incremental approach: making modest interventions at first, while continuing to monitor how the problem evolves over time.

The analysis in this paper can help guide decisionmaking around these initial interventions. In the absence of hard data on synthetic media abuse, this paper bounds the risk management problem through realistic scenarios—rooted in how bad actors already use technology and what would be feasible with today's synthetic media tools. Given scarce resources, financial institutions should focus on combatting three key malicious techniques: deepfake vishing, fabrication of private remarks, and synthetic social botnets. They should consider how to address both narrowcast and broadcast attacks; for the latter, the priority should be synthetic media that amplifies pre-existing narratives or crises.

While focusing on these interventions, it will be particularly important to situate them within a diverse, multistakeholder effort. Countering synthetic media in the financial system will require new technologies, institutional practices, and education in the financial sector and beyond.

Notes

1 "Synthetic Media and Potential Safeguards," Carnegie Endowment for International Peace, https://carnegieendowment.org/siliconvalley/synthetic-media.

2 James Vincent, "Why We Need a Better Definition of 'Deepfake'," The Verge, May 22, 2018, https://www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news.

3 Timothy B. Lee, "I Created My Own Deepfake—It Took Two Weeks and Cost $552," *Ars Technica*, December 16, 2019, https://arstechnica.com/science/2019/12/how-i-created-a-deepfake-of-mark-zuckerberg-and-star-treks-data/.

4 Samantha Cole, "Deepfakes Were Created As a Way to Own Women's Bodies—We Can't Forget That," *Vice*, June 18, 2018, https://www.vice.com/en_us/article/nekqmd/deepfake-porn-origins-sexism-reddit-v25n2.

5 Martin Giles, "The GANfather: The Man Who's Given Machines the Gift of Imagination," *MIT Technology Review*, February 21, 2018, https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/.

6 Charlotte Stanton, "November 2018 Convening Mapping Synthetic Media's Problem and Solution Space," Carnegie Endowment for International Peace, November 16, 2018, https://carnegieendowment.org/2018/11/16/november-2018-convening-mapping-synthetic-media-s-problem-and-solution-space-pub-79892.

7 Lee, "I Created My Own Deepfake—It Took Two Weeks and Cost $552."

8 "Combating Spoofed Robocalls With Caller ID Authentication," Federal Communications Commission, https://www.fcc.gov/call-authentication.

9 Federal Trade Commission, "Consumer Sentinel Network Data Book 2019," January 2020, https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2019/consumer_sentinel_network_data_book_2019.pdf.

10 Ibid.

11 Identity Theft Resource Center, "2018 Annual Report," 2019, https://www.idtheftcenter.org/wp-content/uploads/2019/02/ITRC_ANNUAL-IMPACT-REPORT-2018_web.pdf.

12 "Thirty-six Defendants Indicted for Alleged Roles in Transnational Criminal Organization Responsible for More Than $530 Million in Losses From Cybercrimes," press release, Department of Justice, February 7, 2018, https://www.justice.gov/opa/pr/thirty-six-defendants-indicted-alleged-roles-transnational-criminal-organization-responsible.

13 McAfee, "The Hidden Data Economy," 2017, https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-data-economy.pdf.

14 Catherine Stupp, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case," *Wall Street Journal*, updated August 30, 2019, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

15 “Deepfakes: The Threat to Financial Services,” iProov, January 29, 2020, https://www.iproov.com/newsroom/blog/deepfakes-the-threat-to-financial-services.

16 Samantha Cole, “A Site Faking Jordan Peterson’s Voice Shuts Down After Peterson Decries Deepfakes,” *Motherboard*, August 26, 2019, https://www.vice.com/en_us/article/43kwgb/not-jordan-peterson-voice-generator-shut-down-deepfakes.

17 “How Lyrebird Uses AI to Find Its (Artificial) Voice),” *Wired*, October 15, 2018, https://www.wired.com/brandlab/2018/10/lyrebird-uses-ai-find-artificial-voice/; and Natasha Lomas, “Lyrebird Is a Voice Mimic for the Fake News Era,” *TechCrunch*, April 25, 2017, https://techcrunch.com/2017/04/25/lyrebird-is-a-voice-mimic-for-the-fake-news-era/.

18 Samantha Cole, “‘Deep Voice’ Software Can Clone Anyone’s Voice With Just 3.7 Seconds of Audio,” *Motherboard*, https://www.vice.com/en_us/article/3k7mgn/baidu-deep-voice-software-can-clone-anyones-voice-with-just-37-seconds-of-audio; and Reina Qi Wan, “Clone a Voice in Five Seconds With This AI Toolbox,” *Synced*, September 9, 2019, https://syncedreview.com/2019/09/03/clone-a-voice-in-five-seconds-with-this-ai-toolbox/.

19 Ellie Rushing, “A Philly Lawyer Nearly Wired $9,000 to a Stranger Impersonating His Son’s Voice, Showing Just How Smart Scammers Are Getting,” *Philadelphia Inquirer*, updated March 9, 2020, https://www.inquirer.com/news/voice-scam-impersonation-fraud-bail-bond-artificial-intelligence-20200309.html.

20 Julie Hirschfeld Davis, “Prankster Calls the President, and the White House Puts Him Right Through,” *New York Times*, June 29, 2018, https://www.nytimes.com/2018/06/29/us/politics/prank-call-donald-trump-stuttering-john.html; and Zachary Evans, “Graham Tricked By Russian Prank Call, Called Kurds a ‘Problem’ for Turkey,” National Review, October 10, 2019, https://www.nationalreview.com/news/lindsey-graham-tricked-by-russian-prank-call-called-kurds-a-problem-for-turkey/.

21 Sean Gallagher, “The New Spam: Interactive Robo-calls From the Cloud as Cheap as E-mail,” *Ars Technica*, April 15, 2015, https://arstechnica.com/information-technology/2015/04/the-new-spam-interactive-robo-calls-from-the-cloud-as-cheap-as-e-mail/.

22 Federal Trade Commission, “Consumer Sentinel Network Data Book 2019.”

23 Federal Trade Commission, “Consumer Sentinel Network Data Book 2019; and “Watch Out for ‘Grandparent’ Scams,” Federal Communications Commission, May 22, 2018, https://www.fcc.gov/watch-out-grandparent-scams.

24 “Imposter Scams,” Office of the Minnesota Attorney General, https://www.ag.state.mn.us/Consumer/Publications/ImposterScams.asp.

25 Federal Trade Commission, “Consumer Sentinel Network Data Book 2019.”

26 “We Are VocaliD, Your Voice AI Company,” VocaliD, https://vocalid.ai/about-us/.

27 Rushing, “A Philly Lawyer Nearly Wired $9,0-00 to a Stranger Impersonating His Son’s Voice, Showing Just How Smart Scammers Are Getting.”

28 Ibid.

29 Sarah Krouse, “Robocall Scams Exist Because They Work—One Woman’s Story Shows How,” *Wall Street Journal*, November 21, 2019, https://www.wsj.com/articles/robocall-scams-exist-because-they-workone-womans-story-shows-how-11574351204; Federal Trade Commission, “Consumer Sentinel Network Data Book 2019”; and Federal Bureau of Investigation Internet Crime Complaint Center, “2019 Internet Crime Report,” https://pdf.ic3.gov/2019_IC3Report.pdf.

30 “What Is Sextortion?” Federal Bureau of Investigation, https://www.fbi.gov/video-repository/newss-what-is-sextortion/view.

31 “The Revival and Rise of Email Extortion Scams,” Symantec, July 30, 2019, https://symantec-blogs.broadcom.com/blogs/threat-intelligence/email-extortion-scams.

32 Kate Fazzini, “Email Sextortion Scams Are on the Rise and They’re Scary—Here’s What to Do If You Get One,” CNBC, June 17, 2019, https://www.cnbc.com/2019/06/17/email-sextortion-scams-on-the-rise-says-fbi.html.

33 “The Revival and Rise of Email Extortion Scams.”

34 Federal Bureau of Investigation Internet Crime Complaint Center, “2019 Internet Crime Report.”

35 Cole, “Deepfakes Were Created As a Way to Own Women’s Bodies.”

36 Joseph Cox, “Most Deepfakes Are Used for Creating Non-Consensual Porn, Not Fake News,” *Vice*, October 7, 2019, https://www.vice.com/en_us/article/7x57v9/most-deepfakes-are-porn-harassment-not-fake-news.

37 Rana Ayyub, “In India, Journalists Face Slut-Shaming and Rape Threats,” *New York Times*, May 22, 2018, https://www.nytimes.com/2018/05/22/opinion/india-journalists-slut-shaming-rape.html.

38 James Vincent, “New AI Deepfake App Creates Nude Images of Women in Seconds,” The Verge, June 27, 2019, https://www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnude-non-consensual-pornography.

39 Jon Porter, “Another Convincing Deepfake App Goes Viral Prompting Immediate Privacy Backlash,” The Verge, September 2, 2019, https://www.theverge.com/2019/9/2/20844338/zao-deepfake-app-movie-tv-show-face-replace-privacy-policy-concerns.

40 Fazzini, “Email Sextortion Scams Are on the Rise and They’re Scary.”

41 Gautam S. Mengle, “Law Enforcers Worried as Deep Nude Makes a Return,” *Hindu*, April 13, 2020, https://www.thehindu.com/news/national/law-enforcers-worried-as-deep-nude-makes-a-return/article31334415.ece.

42 Fazzini, “Email Sextortion Scams Are on the Rise and They’re Scary.”

43 Federal Bureau of Investigation Internet Crime Complaint Center, “Business E-mail Compromise,” public service announcement, January 22, 2015, https://www.ic3.gov/media/2015/150122.aspx; and Federal Bureau of Investigation Internet Crime Complaint Center, “Business Email Compromise: Gift Cards,” public service announcement, October 24, 2018, https://www.ic3.gov/media/2018/181024.aspx.

44 Federal Bureau of Investigation Internet Crime Complaint Center, “2019 Internet Crime Report.”

45 Federal Bureau of Investigation Internet Crime Complaint Center, “Business E-Mail Compromise”; and “Recognizing and Avoiding Business Email Compromise Attacks,” Proofpoint, 2019, https://www.proofpoint.com/sites/default/files/pfpt-us-ig-bec.pdf.

46 Federal Bureau of Investigation Internet Crime Complaint Center, “2019 Internet Crime Report.”

47 Stupp, “Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case.”

48 Ibid.

49 Alessandro Cauduro, “Live Deep Fakes—You Can Now Change Your Face to Someone Else’s in Real Time Video Applications,” Medium, April 4, 2018, https://medium.com/huia/live-deep-fakes-you-can-now-change-your-face-to-someone-elses-in-real-time-video-applications-a4727e06612f.

50 Lawrence Delevingne, “Short & Distort? The Ugly War Between CEOs and Activist Critics,” Reuters, March 21, 2019, https://www.reuters.com/article/us-usa-stocks-shorts-insight/short-distort-the-ugly-war-between-ceos-and-activist-critics-idUSKCN1R20AW; and “SEC Fires Warning Shot Against ‘Short and Distort’ Schemes,” DLA Piper, October 18, 2018, https://www.dlapiper.com/en/us/insights/publications/2018/10/sec-fires-warning-shot-against/.

51 Joshua Mitts, “Short and Distort,” Columbia Law and Economics Working Paper No. 592, February 20, 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3198384; and Thomas Renault, “Pump-and-dump or News? Stock Market Manipulation on Social Media,” European Financial Management Association, 2017, https://efmaefm.org/0EFMAMEETINGS/EFMA%20ANNUAL%20MEETINGS/2017-Athens/papers/EFMA2017_0387_fullpaper.pdf.

52 Will Kenton, “Poop and Scoop,” Investopedia, May 11, 2018, https://www.investopedia.com/terms/p/poopandscoop.asp.

53 Claire Atkinson, “Fake News Can Cause ‘Irreversible Damage’ to Companies—and Sink Their Stock Price,” *NBC News*, April 25, 2019, https://www.nbcnews.com/business/business-news/fake-news-can-cause-irreversible-damage-companies-sink-their-stock-n995436.

54 Leroy Terrelonge and Jim Hempstead, “Deepfakes Can Threaten Companies’ Financial Health,” *Moody’s Investors Service*, August 1, 2019, https://www.moodys.com/research/Moodys-Deepfakes-can-threaten-companies-financial-health--PBC_1188117?showPdf=true.

55 Delevingne, “Short & Distort?”

56 Darla Mercado, “Fisher Withdrawals Top $3 Billion as Texas Retirement Plan Exits,” CNBC, October 25, 2019, https://www.cnbc.com/2019/10/25/fisher-withdrawals-top-3-billion-as-texas-retirement-plan-exits.html.

57 Ramona Shelburne, “When the Donald Sterling Saga Rocked the NBA—and Changed It Forever,” August 20, 2019, https://www.espn.com/nba/story/_/id/27414482/when-donald-sterling-saga-rocked-nba-changed-forever.

58 Drew Harwell, “Doctored Images Have Become a Fact of Life for Political Campaigns. When They’re Disproved, Believers ‘Just Don’t Care.’” *Washington Post*, January 14, 2020, https://www.washingtonpost.com/technology/2020/01/14/doctored-political-images.

59 Megan Metzger, “Effectiveness of Responses to Synthetic and Manipulated Media on Social Media Platforms,” Carnegie Endowment for International Peace, November 15, 2019, https://carnegieendowment.org/2019/11/15/legal-ethical-and-efficacy-dimensions-of-managing-synthetic-and-manipulated-media-pub-80439#effectiveness.

60 Stefano Cresci, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, Maurizio Tesconi, “Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter,” *ACM Transactions on the Web* 13 (2018), https://arxiv.org/pdf/1804.04406.pdf.

61 Rui Fan, Oleksandr Talavera, and Vu Tran, "Social Media Bots and Stock Markets," Swansea University School of Management, Working Paper No. 30, 2018, https://rahwebdav.swan.ac.uk/repec/pdf/WP2018-30.pdf.

62 "Platform Manipulation and Spam Policy," Twitter, September 2019, https://help.twitter.com/en/rules-and-policies/platform-manipulation; and "How to Spot a Twitter Bot," Symantec, October 26, 2018, https://symantec-blogs.broadcom.com/blogs/election-security/spot-twitter-bot.

63 Jurgen Knauth, "Language-Agnostic Twitter Bot Detection," *Proceedings of Recent Advances in Natural Language Processing*, September 2–4, 2019, https://www.aclweb.org/anthology/R19-1065.pdf; John Gramlich, "Q&A: How Pew Research Center Identified Bots on Twitter," Pew Research Center, April 19, 2018, https://www.pewresearch.org/fact-tank/2018/04/19/qa-how-pew-research-center-identified-bots-on-twitter/; and Alyssa Newcomb, "Twitter Is Purging Millions of Fake Accounts—and Investors Are Spooked," *NBC News*, July 9, 2018, https://www.nbcnews.com/tech/tech-news/twitter-purging-millions-fake-accounts-investors-are-spooked-n889941.

64 Yoel Roth and Nick Pickles, "Bot or Not? The Facts About Platform Manipulation on Twitter," Twitter Blog, May 18, 2020, https://blog.twitter.com/en_us/topics/company/2020/bot-or-not.html.

65 Gramlich, "Q&A: How Pew Research Center Identified Bots on Twitter"; and Michael Kreil, "The Army That Never Existed: The Failure of Social Bots Research," Github, November 2, 2019, https://michaelkreil.github.io/openbots.

66 Newcomb, "Twitter Is Purging Millions of Fake Accounts"; and Luis Sanchez, "Conservatives Say They've Lost Thousands of Followers on Twitter," *The Hill*, February 21, 2018, https://thehill.com/policy/technology/374842-conservatives-say-theyve-lost-thousands-of-followers-on-twitter.

67 Paris Martineau, "Facebook Removes Accounts With AI-generated Profile Photos," *Ars Technica,* December 23, 2019, https://arstechnica.com/tech-policy/2019/12/facebook-removes-accounts-with-ai-generated-profile-photos.

68 Davey Alba, "Facebook Discovers Fakes That Show Evolution of Disinformation," *New York Times*, December 20, 2019, https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html; and Raphael Satter, "Experts: Spy Used AI-generated Face to Connect With Targets," Associated Press, June 13, 2019, https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

69 James Vincent, "OpenAI Has Published the Text-generating AI It Said Was Too Dangerous to Share," The Verge, November 7, 2019, https://www.theverge.com/2019/11/7/20953040/openai-text-generation-ai-gpt-2-full-model-release-1-5b-parameters.

70 Renault, "Pump-and-dump or News?"; and Elana Lyn Gross, "Why Peloton Stock Dropped More Than 10% After 'Sexist' Ad Backlash," *Forbes*, December 5, 2019, https://www.forbes.com/sites/elanagross/2019/12/05/peloton-stock-is-down-more-than-10-following-backlash-about-sexist-ad/.

71

72 This Person Does Not Exist, https://www.thispersondoesnotexist.com/. License: https://nvlabs.github.io/stylegan2/license.html.

73 "GauGAN Beta," NVIDIA, http://nvidia-research-mingyuliu.com/gaugan/. License: http://nvidia-research-mingyuliu.com/gaugan/term.txt.

74 Adam King, "Talk to Transformer," https://talktotransformer.com/.

75 The first sentence of the mock Tweet was a human-written prompt, which enabled the algorithm to write the remainder. The algorithm was run several times, and the most convincing output was selected for this mock-up. The mock Twitter bio was completely AI-generated on the first try, using the prompt "About me:". Although this AI algorithm requires a human prompt, prompts could be algorithmically generated in the future.

76 "20th Century English Name Generator," Fantasy Name Generators, https://www.fantasynamegenerators.com/20th-century-english-names.php.

77 Satter, "Experts: Spy Used AI-generated Face to Connect With Targets"; and Ravie Lakshmanan, "This AI Tool Is Smart Enough to Spot AI-generated Articles and Tweets," The Next Web, July 29, 2019, https://thenextweb.com/artificial-intelligence/2019/07/29/this-ai-tool-is-smart-enough-to-spot-ai-generated-articles-and-tweets/.

78 Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi, "The Paradigm-shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race," Proceedings of the 26th International Conference on World Wide Web Companion (2017), https://arxiv.org/pdf/1701.03017.pdf.

79 Cresci, et al., "The Paradigm-shift of Social Spambots."

80 "Applying Social Sentiment to Your Stock Research," Fidelity, 2017, https://www.fidelity.com/learning-center/tools-demos/research-tools/social-sentiment-research-video.

81 "The Day Social Media Schooled Wall Street," *Atlantic Re:think*, https://www.theatlantic.com/sponsored/etrade-social-stocks/the-day-social-media-schooled-wall-street/327.

82 "Investor Bulletin: Social Sentiment Investing Tools—Think Twice Before Trading Based on Social Media," Securities Exchange Commission, April 3, 2019, https://www.sec.gov/oiea/investor-alerts-and-bulletins/ib_sentimentinvesting.

83 Ben Chapman, "Metro Bank Says Customers' Money Safe After WhatsApp Rumour Sparks Panic," *Independent*, May 13, 2019, https://www.independent.co.uk/news/business/news/metro-bank-whatsapp-money-account-safe-deposit-box-a8911296.html.

84 Issaku Harada, "Online Rumors Spark Runs on Smaller Chinese Banks," *Nikkei Asian Review*, December 19, 2019, https://asia.nikkei.com/Economy/Online-rumors-spark-runs-on-smaller-Chinese-banks.

85 Matthias Williams and Tsvetelia Tsolova, "Accusations Fly in Bulgaria's Murky Bank Run," Reuters, July 4, 2014, https://www.reuters.com/article/us-bulgaria-banking-insight/accusations-fly-in-bulgarias-murky-bank-run-idUSKBN0F90SG20140704.

86 Nadine Ajaka, Elyse Samuels, and Glenn Kessler, "Seeing Isn't Believing: The Fact Checker's Guide to Manipulated Video," *Washington Post*, 2019, https://www.washingtonpost.com/graphics/2019/politics/fact-checker/manipulated-video-guide/.

87 "The Day Social Media Schooled Wall Street."

88 Ibid.

89 Shawn Langlois, "This Day in History: Hacked AP Tweet About White House Explosions Triggers Panic," *MarketWatch*, April 23, 2018, https://www.marketwatch.com/story/this-day-in-history-hacked-ap-tweet-about-white-house-explosions-triggers-panic-2018-04-23.

90 Metzger, “Effectiveness of Responses to Synthetic and Manipulated Media on Social Media Platforms.”

91 “How High-frequency Trading Hit a Speed Bump,” *Financial Times*, January 1, 2018, https://www.ft.com/content/d81f96ea-d43c-11e7-a303-9060cb1e5f44.

92 Atkinson, “Fake News Can Cause ‘Irreversible Damage’ to Companies.”

93 Jean-Philippe Serbera, “Flash Crashes: If Reforms Aren’t Ramped Up, the Next One Could Spell Global Disaster,” The Conversation, January 7, 2019, https://theconversation.com/flash-crashes-if-reforms-arent-ramped-up-the-next-one-could-spell-global-disaster-109362.

94 Ding Yi, “Chinese Central Bank Denies Digital Currency Issuance Rumors,” *CX Tech*, November 14, 2019, https://www.caixinglobal.com/2019-11-14/chinese-central-bank-denies-digital-currency-issuance-rumors-101483430.html.

95 Suvashree Ghosh, “India Central Bank Denies Rumors of Bank Closures,” *Bloomberg*, September 25, 2019, https://www.bloomberg.com/news/articles/2019-09-25/frayed-nerves-force-india-watchdog-to-douse-bank-closure-rumors; and “Myanmar Central Bank Refutes Rumors About Bank Closure,” Xinhua, August 19, 2015, http://www.globaltimes.cn/content/937860.shtml.

96 Venkatesan Vembu, “Buzz of $430 Billion Loss; Top Banker Defection Freaks Out China,” *DNA India*, September 1, 2010, https://www.dnaindia.com/business/report-buzz-of-430-billion-loss-top-banker-defection-freaks-out-china-1431780.

97 Beth Piskora, “Fed Head Is Not Dead—Rumor False; Markets Drop on Earnings, Rate Hike Fears,” June 23, 2000, https://nypost.com/2000/06/23/fed-head-is-not-dead-rumor-false-markets-drop-on-earnings-rate-hike-fears.

98 Sarah Foster, “The Fed Wants to Be Easier to Understand—and It May Be Risky to the Markets,” Bankrate, June 4, 2019, https://www.bankrate.com/banking/federal-reserve/fed-simple-communication-may-be-confusing-markets; and Alister Bull, “Fed Slammed for Poor Communication by Its Own Advisory Council,” Reuters, October 4, 2013, https://www.reuters.com/article/us-usa-fed-communication-idUSBRE99313220131004.

99 BBC, “Call for Bank of England Executive to Quit Over Security Breach,” December 19, 2019, https://www.bbc.com/news/business-50849479.

100 U.S. Senate Permanent Subcommittee on Investigations, “Abuses of the Federal Notice-and-Comment Rulemaking Process,” October 24, 2019, https://www.hsgac.senate.gov/imo/media/doc/2019-10-24%20PSI%20Staff%20Report%20-%20Abuses%20of%20the%20Federal%20Notice-and-Comment%20Rulemaking%20Process.pdf?mod=article_inline.

101 Jeremy Singer-Vine and Kevin Collier, “Political Operatives Are Faking Voter Outrage With Millions Of Made-Up Comments To Benefit The Rich And Powerful,” Buzzfeed News, October 3, 2019, https://www.buzzfeednews.com/article/jsvine/net-neutrality-fcc-fake-comments-impersonation.

102 James V. Grimaldi and Paul Overberg, “Millions of People Post Comments on Federal Regulations. Many Are Fake,” *Wall Street Journal*, December 12, 2017, https://www.wsj.com/articles/millions-of-people-post-comments-on-federal-regulations-many-are-fake-1513099188.

103 Vincent, “OpenAI Has Published the Text-generating AI It Said Was Too Dangerous to Share.”

104 King, “Talk to Transformer,” https://talktotransformer.com/. This output was produced on the first try. The only human intervention was to separate paragraphs.

105 Tom McKay, "Turns Out Elon Musk–Backed OpenAI's Text Generator Is More Funny Than Dangerous, For Now," Gizmodo, November 7, 2019, https://gizmodo.com/turns-out-elon-musk-backed-openais-text-generator-is-mo-1839705114.

106 Max Weiss, "Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished From Human Submissions," *Technology Science*, December 18, 2019, https://techscience.org/a/2019121801/.

107 Clea Simon, "How I Hacked the Government (It Was Easier Than You May Think)," *Harvard Gazette*, February 6, 2020, https://news.harvard.edu/gazette/story/2020/02/why-an-undergrad-flooded-government-websites-with-bot-comments/; and Weiss, "Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions."

108 Lakshmanan, "This AI Tool Is Smart Enough to Spot AI-generated Articles and Tweets."

109 Karen Hao, "An AI for Generating Fake News Could Also Help Detect It," *MIT Technology Review*, March 12, 2019, https://www.technologyreview.com/2019/03/12/136668/an-ai-for-generating-fake-news-could-also-help-detect-it/.

110 U.S. Senate Permanent Subcommittee on Investigations, "Abuses of the Federal Notice-and-Comment Rulemaking Process," October 24, 2019, https://www.hsgac.senate.gov/imo/media/doc/2019-10-24%20PSI%20Staff%20Report%20-%20Abuses%20of%20the%20Federal%20Notice-and-Comment%20Rulemaking%20Process.pdf?mod=article_inline.

111 Weiss, "Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions."

112 Beth Simone Noveck, "Astroturfing Is Bad But It's Not the Whole Problem," Nextgov, February 6, 2020, https://www.nextgov.com/ideas/2020/02/astroturfing-bad-its-not-whole-problem/162932/.

113 James V. Grimaldi, "Federal Agencies Found to Be Lax in Halting Fake Comments on Proposed Rules," *Wall Street Journal*, October 24, 2019, https://www.wsj.com/articles/federal-agencies-found-to-be-lax-in-halting-fake-comments-on-proposed-rules-11571909402.

114 Ajaka, Samuels, and Kessler, "Seeing Isn't Believing: The Fact Checker's Guide to Manipulated Video."

115 "CBS News Admits Bush Documents Can't Be Verified," Associated Press, September 21, 2004, http://www.nbcnews.com/id/6055248/ns/politics/t/cbs-news-admits-bush-documents-cant-be-verified/#.Xp45-FNKh0s; Hugh Schofield, "The Fake French Minister in a Silicone Mask Who Stole Millions," BBC, June 20, 2019, https://www.bbc.com/news/world-europe-48510027; and Ashley Feinberg, "The Pee Tape Is Real, but It's Fake," Slate, September 15, 2019, https://slate.com/comments/news-and-politics/2019/09/inside-the-convincing-fake-trump-pee-tape.html.

116 Emma Fletcher, "New Twist to Grandparent Scam: Mail Cash," Federal Trade Commission, December 3, 2018, https://www.ftc.gov/news-events/blogs/data-spotlight/2018/12/new-twist-grandparent-scam-mail-cash.

117 Thomas E. Kadri, "The Legal Implications of Synthetic and Manipulated Media," Carnegie Endowment for International Peace, November 15, 2019, https://carnegieendowment.org/2019/11/15/legal-ethical-and-efficacy-dimensions-of-managing-synthetic-and-manipulated-media-pub-80439#legal.

118 David Danks and Jack Parker, “The Un/Ethical Status of Synthetic Media,” Carnegie Endowment for International Peace, November 15, 2019, https://carnegieendowment.org/2019/11/15/legal-ethical-and-efficacy-dimensions-of-managing-synthetic-and-manipulated-media-pub-80439#ethics; David Danks and Jack Parker, “Ethical Analysis of Responses to Synthetic and Manipulated Media,” Carnegie Endowment for International Peace, November 15, 2019, https://carnegieendowment.org/2019/11/15/legal-ethical-and-efficacy-dimensions-of-managing-synthetic-and-manipulated-media-pub-80439#analysis; and Charlotte Stanton, “June 2019 Convening on Defining Inappropriate Synthetic/Manipulated Media Ahead of the U.S. 2020 Election,” June 19, 2019, Carnegie Endowment for International Peace, https://carnegieendowment.org/2019/06/19/june-2019-convening-on-defining-inappropriate-synthetic-manipulated-media-ahead-of-u.s.-2020-election-pub-79661.

119 Stanton, “November 2018 Convening Mapping Synthetic Media’s Problem and Solution Space.”

120 Amber Frankland and Lindsay Gorman, “Combating the Latest Technological Threat to Democracy: A Comparison of Facebook and Twitter’s Deepfake Policies,” January 13, 2020, Alliance for Securing Democracy, https://securingdemocracy.gmfus.org/combating-the-latest-technological-threat-to-democracy-a-comparison-of-facebooks-and-twitters-deepfake-policies/; and Kate Cox, “Twitter Wants Your Feedback on Its Proposed Deepfakes Policy,” *Ars Technica*, November 11, 2019, https://arstechnica.com/tech-policy/2019/11/twitter-wants-your-feedback-on-its-proposed-deepfakes-policy/.

121 Tim Bradshaw, “‘Tech Debt’: Why Badly Written Code Can Haunt Companies for Decades,” *Financial Times*, November 27, 2019, https://www.ft.com/content/d6822eb0-0fe0-11ea-a7e6-62bf4f9e548a.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

1779 Massachusetts Avenue NW | Washington, DC 20036 | P: + 1 202 483 7600

CarnegieEndowment.org