DEEPFAKE, CHEAPFAKE: THE INTERNET'S NEXT EARTHQUAKE?

FixFake Symposium Proceedings, Part 1. **Presented by DeepTrust Alliance**

DEEPTRUST
ALLIANCE

Authorization to reproduce material in this publication is granted, subject to full acknowledgement. It should be cited as: Harrison, Aros. (2020).
Defining the problems of deepfakes and disinformation. New York: DeepTrust Alliance.

This publication can be downloaded here.

For any questions, please contact info@deeptrustalliance.com

Design by Shae Riley.

TABLE OF CONTENTS

INTRODUCTION

Deepfakes burst into the collective imagination in 2018, portending serious consequences for our society. Until recently, media consumers could presumptively trust the authenticity of video and audio content - what they viewed was generally the same as what was originally captured by recording equipment.   Only Hollywood movie studios (and sovereign governments) had access to the processing power and computer-generated imagery (CGI) tools for advanced media creation, and the resulting content was labeled and consumed inside movie theaters and television programs. But those techniques have escaped the entertainment industry and are now being deployed as powerful new weapons in the long running battle of disinformation. Deepfakes threaten to create an alternate universe where what you see - or hear - may only be a spurious facsimile of reality. The implications of maliciously manipulated video, audio and text are staggering for the global community. How will decision makers trust inputs and data to have confidence in their conclusions? How will industry, government, courts, media and other institutions cope in a world where the authenticity of their communication is always suspect?

One of the first deepfake videos to bring wide exposure to the problem was created by Jordan Peele.[1]

Today, the development of deepfakes is only in its infancy; however, even long-standing tools for audio-visual manipulation are now widely accessible. Already, 3.5BN[2] humans carry advanced cameras, editing, filtering, synthesis and even AI tools on their smartphones. With smartphone ownership rapidly increasing worldwide, the scale of the risk is unprecedented.

Kathryn Harrison
CEO and Founder, Deep Trust Alliance

Dave Toomey
Policy Director, Deep Trust Alliance

3.5 **BILLION**

“Today, 3.5BN humans carry advanced cameras, editing, synthesis and AI tools in their pockets on their smartphones.

With the very real possibility that nothing we see on our devices may be exactly as it appears, the credibility of media, governments, companies, and technologies will be at stake. While today deepfakes represent only five percent of identified disinformation, we are rushing toward a cyborg future where human-machine interface will be the norm. The information landscape and human paradigms of trust are being transformed before our very eyes. How can we trust and verify the data that will stream from every human, machine, and device? Leaders of government and business as well as individuals across the globe are unequipped and ill-prepared to overcome the threats posed by deepfakes and synthetic media. Alarmingly, a fundamental erosion of trust is already underway. Hard as it may be to accept, the very viability of democracy may be at risk. Hyperbole, perhaps, but unrestrained “progress” can even outpace hyperbole.

DEFINITIONS

Below we lay out the definitions of some of the most common terms that were used throughout the FixFake Symposia and generally agreed upon by the disinformation community. There is still debate about the taxonomy of words used to describe the problems inherent to manipulated media as well as its solutions, and there is much work to be done.

AUDIO-VISUAL MANIPULATION *A series of techniques in which partisans create an image or argument that supports their particular interests by modifying video or audio with a range of tools and techniques from traditional video processing through to artificial intelligence (AI) enabled capabilities. Joan Donovan and Britt Paris define it as "any sociotechnical way for influencing the interpretation of media."* [3]

SYNTHETIC MEDIA *Refers to media that is algorithmically created or modified- ie it is made with a mix of humans and computers.*

DEEPFAKES *The word "deepfake" is a combination of "deep learning" and "fake." Most often, deepfakes refer to videos, images, audio or text created with artificial intelligence (AI) technologies such as Generative Adversarial Networks (GANs) or Recurrent Neural Networks (RNNs). These content synthesis technologies enable media representations of non-existent subjects as well as subjects doing or saying things they've never done or said.*

CHEAPFAKES *Audio-visual manipulations which use conventional techniques like speeding, slowing, cutting, re-staging, or re-contextualizing footage to change the meaning and interpretation of media.*[4] *They do not use AI enabled tools. Cheapfakes are by far the largest threat in the market today.*

MISINFORMATION *False or misleading information that is spread, regardless of whether there is intent to mislead. Misinformation can also include true or real information that is used erroneously out of context.*

DISINFORMATION *Deliberately misleading or biased information; manipulated narrative or facts; propaganda.*

FAKE NEWS *A term that has now been weaponized to disparage real news that is disliked by people in power.*

FixFake Symposia 2019

Clearly, the dangers of deepfakes must be confronted - and speed is paramount. We founded DeepTrust Alliance to help business, government and civil society adopt breakthrough technologies to mitigate the impact of the growing cybersecurity threat to virtually any internet transaction. As a community of technologists and technology-optimists, DeepTrust Alliance is fundamentally hopeful about the potential of AI technologies to transform and empower human communication in positive ways. However, we view the unchecked development and dissemination of deepfakes and disinformation as an insidious enemy to both national security and the entire digital economy. DeepTrust Alliance is committed to raising awareness, convening stakeholders, building open solutions and driving comprehensive adoption to counteract the dangerous and disruptive emergence of an all too powerful global threat.

To kick off our campaign, we convened the FixFake Symposia series in New York City, Washington, D.C and San Francisco in November and December 2019. The rapidly approaching 2020 election was the impetus for a key objective of the symposia: recognizing and confronting the doubts and dissension that can be seeded through synthetic media. The risks of disinformation must be on the radar of every media outlet, corporate, academic, government and civil society leader in the United States. To that end, we brought some of the world's leading experts on media, computer vision, forensics, policy and national security together with industry leaders for a wide-ranging conversation that sought to define the problems and to consider the array of tools to combat them. Across all sessions, the emphasis was on action as we examined the interplay of technology, policy and business.

CONVENING THE STAKEHOLDERS

WHO PARTICIPATED?

With nearly 200 attendees across the three sessions, the participants represented stakeholders from:

BUSINESS

- Financial Services
- Telecom
- Healthcare
- Technology
- Entertainment
- Lawyers
- Accountants
- Public Relations

GOVERNMENT

- Legislators
- Federal Agencies
- Politicians
- Law enforcement

MEDIA

- Publishers
- Journalists
- Social Platforms
- Fact-checking
- Brand safety companies

CIVIL SOCIETY

- Academia
- Non-governmental Organizations
- Inter-governmental organizations
- Foundations

We present this report as the first of a series in which we relay the outcomes of the FixFake Symposia sessions. In Part 1, we examine the threats that malicious audiovisual manipulations cause to society and industry. We also examine the obstacles to overcoming disinformation at scale. In the forthcoming Part 2, we review the existing and potential solutions to these problems and share the roadmap that DeepTrust Alliance is developing with its members to confront these issues.

To solve this labyrinth of problems, a network of dedicated and coordinated actors must come together to define and prioritize the issues; build a portfolio of solutions; drive adoption of solutions; and measure the impact of those solutions. With constantly evolving threats, solution builders find themselves in an arms race against nefarious actors. Long term success requires:

1. *Technical solutions that empower information consumers.*
2. *Legislative policy that balances privacy and security.*
3. *Education about the harms of disinformation, implications and solutions.*

Actual change will only come by deploying a portfolio of technology, legislation and education as there is not a 'silver bullet' solution to fix all the problems.

There is good news. Civil society, academia, and startups are stepping in to take action; however, there is still significant work that is needed with collaboration and incentives as two of the biggest challenges to overcome.

We hope that you will join DeepTrust Alliance as we build a global supply chain of trust.

Kathryn Harrison
Founder and CEO
DeepTrust Alliance

WHAT THREATS DO DEEPFAKES AND SYNTHETIC MEDIA POSE TO SOCIETY?

Throughout each of the three events, there was robust discussion about how deepfakes intensify existing threats against individuals, industry and society. All of these risks apply equally to cheap fakes, which is the more pressing issue in the short run. We noted a disparity in the level of awareness among the industries represented. We briefly summarize the broad harms, industry threats, and the financial implications below.

BOTTOM LINE Deepfakes and manipulated media introduce powerful new weapons for perpetrators of fraud and digital harm. Furthermore, the liar's dividend threatens the foundations of trust in society. Many companies are not yet aware of the scope of the threats that make them vulnerable targets.

LIAR'S DIVIDEND

The existence of deepfake technology allows an individual accused of wrongdoing (or any party who wishes to discredit true information) to sow doubt and deniability about the origins and authenticity of content. This corrosive capability fundamentally undermines confidence in the veracity and trustworthiness of the information ecosystem.

VIEW FULL VIDEO [6]

EXAMPLE [5]

In 2018, the President of Gabon, Ali Bongo, had been away from his country for several months seeking treatment for a serious illness in Saudi Arabia. Questions began surfacing about whether he had died and if the ruling government was trying to cover it up to retain power. The ailing president released a "proof of life" video to show that he was recovering from a stroke. Opponents of the president claimed that the video was a "deepfake" and the Gabonese military attempted an unsuccessful coup. While the video itself had some flags which gave forensics experts pause, it was never definitively proved to be manipulated content. This example shows how the existence of deepfake generating technology can unleash political destabilization.

SOCIAL ENGINEERING

Social engineering is a critical threat in which deceptions are deployed to manipulate individuals into divulging confidential or personal information that may be used to perpetrate fraud against them. Deepfakes and manipulation technologies create a new toolset for those who aim to profit through identity theft and impersonation.

EXAMPLE [7]

In March 2019, French insurance company Euler Hermes reported that criminals in Europe used commercially available, voice-generating AI software to impersonate the CEO of a German parent company in order to steal $243,000 from a UK energy company's bank accounts, which was never recovered.

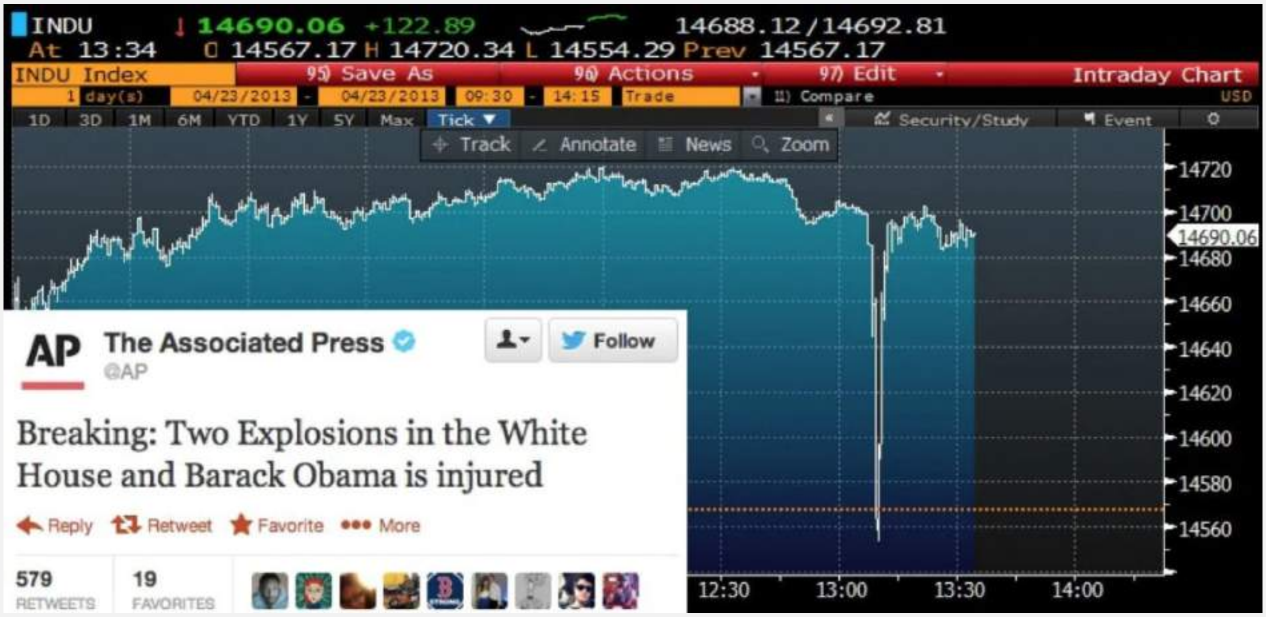
MARKET MANIPULATION

With the prevalence of algorithmic trading mechanisms, synthetic media provide the opportunity to manipulate public company share prices by shifting sentiment on social media platforms. By creating a momentary shift in the market perception of an executive, product or company, malicious actors can reap significant financial benefits and harm investors.

EXAMPLE [8]

On April 23, 2013, a single tweet from the Associated Press reported the White House had been attacked, and the President was injured. Within three minutes, automated trading algorithms responded aggressively to the sentiment signals coming from the social media platform. U.S. markets plunged, wiping out over $136 billion of value from the S&P 500 index in less than three minutes. The tweet that spawned the sell-off, in fact, came from Syrian hackers who had infiltrated the Associated Press Twitter account, aiming to harm the United States.

Fortunately, US officials corrected the disinformation, and markets recovered their losses quickly. However, this is a cautionary tale about the speed and scale of damage that can result from disinformation in the digital age.



The Associated Press
@AP

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

[9]

[12]

RANA AYYUB

JOURNALIST

EXTORTION + HARASSMENT

Finally, synthetic media can be used to extort money and harass individuals or organizations. This disproportionately affects women. Notably, 96% of deepfakes[10] *feature women, often actresses and musicians, in the form of non-consensual deepfake porn, made available through specialized porn websites.*

EXAMPLE [11]

Rana Ayyub, a journalist and outspoken critic of the Hindu Nationalist movement in India, had her likeness incorporated into a porn video using deepfake technology. The video was distributed widely across India through Facebook and WhatsApp groups, leading to death threats. Local law enforcement refused to assist until the United Nations intervened.

THREATS TO THE INDUSTRY

The FixFake Symposia exposed additional cases and examples of how these tools can be used across a diverse set of industries and organizations. Organizations may be aware of the threat of fraud, but not recognize the new tactics and manifestations that fraudsters are wielding.

There are also significant indirect harms which, though harder to quantify, can have a significant impact on our society. These include loss of trust in institutions and the internet, increase in fear, reductions in quality of life, and diversion of investment in education, training and innovation to risk management and fraud protection.

Below we have included multiple strategies nefarious actors can use to target industry, society and the digital economy –

POTENTIAL SCENARIOS	WHO IS IN THE CROSSHAIRS?			IMPACTS
	INDIVIDUALS	ORGANIZATIONS	ORGANIZATIONS	
Bogus identity authentication or creation with facial or audio recognition through AI generated audio, video, image or text	Consumers; Executives; Politicians; Voters	Financial Services; Telecom; Healthcare: Consumer Goods & Services; Technology	Government; Courts; Armed Services; Media and News; Academia	Identify theft, financial fraud, telecom fraud, robocalls, insurance fraud, medical malpractice, securities fraud, security breaches, national security risk, espionage, undermine trust in institutions, electoral fraud
Impersonation or false content of public figure, politican or individual	Consumers; Celebrities; Politicians; Executives; Judges; Journalists	Any Public Company; Entertainment; Financial Service; Public Relations; Consumer Goods & Services; Technology	Government; Military; Media and News; Academia; Justice System	Reputation, value, safety and security of person, product, company, industry or nation, securities fraud, compromised KYC/ AML processes, electoral fraud, bodily harm, undermined trust in institution, harm to minorities/ marginalized constituencies

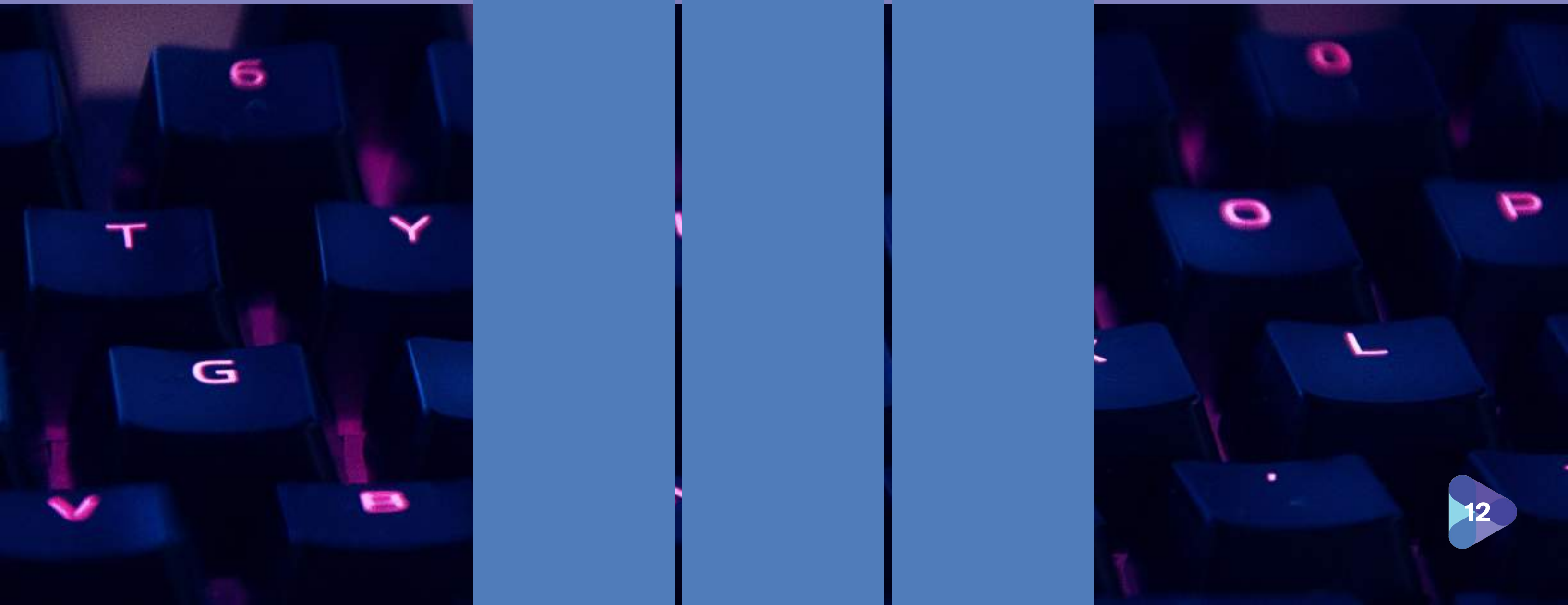
POTENTIAL SCENARIOS	WHO IS IN THE CROSSHAIRS?			IMPACTS
	INDIVIDUALS	ORGANIZATIONS	ORGANIZATIONS	
Misrepresentation of assets	Buyers; Sellers; Consumers; Agents	Financial Services; Real Estate; Technology; Healthcare; Consumer Goods & Services	Government; Justice System; Media and News; Academia	Financial loss, fraud, theft, financial securities fraud, compromised KYC/ AML processes, real estate fraud, harm to individual health and safety, harm to corporate asset health and safety, wrongful imprisonment, plagiarism/ synthetic papers in academic settings
Ad placement next to **false news and misinformation**	Consumers; Politicians	Advertising; Technology; Political Parties; Consumer Goods & Services; Any Digital Advertiser	Government; Elections; Media and News	Damage to credibility and brand equity of the advertiser, negative public outcry, risk of advertiser flight from technology platform, discredit credible news source
Plausible deniability / Liar's dividend / Difficulty in verification	Consumers; Voters; Journalists; Politicians; Public Figures	Corporations; Civil Society Organizations; Political Parties	News Media; Elections; Government; Military; Public Health; Academia	Loss of trust institutions, loss of trust in public figures, lack of foundational truths in society, disillusionment, inability to prove truth, spread of disinformation
Tampering with video, audio, images or text video of surveillance	Plaintiffs; Defendants; Passengers; Security Professionals; Patients; Doctors	Law Firms; Audit Firms; Energy & Utilities; Industrial; Automotive, Transportation; Supply Chain; Hospitals; Corporations	Justice System; Police; National Security; Military; Public Health	Wrongful conviction and/or imprisonment, fines, increased litigation, misdiagnoses, health insurance fraud, loss of life, pain and suffering, loss of confidence in institutions and technology, disruption or compromise of critical infrastructure, security breaches, fraud, failure in video enabled technologies like self-driving cars or drones, regulatory penalties

POTENTIAL SCENARIOS	WHO IS IN THE CROSSHAIRS?			IMPACTS
	INDIVIDUALS	ORGANIZATIONS	ORGANIZATIONS	

Manipulating public opinion	Voters;
	Minorities;
	Patients;
	Doctors
		Corporations;
		Hospitals;
		Pharma,
		Biotech
			Government;
			Military;
			Elections;
			News Media
				Divisive discourse, damage to social cohesion, domestic or foreign election manipulation, lack of foundational set of facts, ethnic violence, influence operations and other national security risks, Incorrect diagnoses and treatments, increased susceptibility and spread of disease, medical expenses, pain and suffering, loss of life, insurance fraud, increased costs, loss of trust in medical systems, panic

Erasing people from digital historical images	Politicians;
	Public Figures
			Government;
			Academia;
			Military
				Damage to social cohesion, lack of foundational history, distrust in institutions

False product reviews	Buyers;
	Sellers;
	Consumers
		Consumer Goods & Services;
		Travel,
		Tourism,
		Hospitality;
		Technology
			Marketplaces	Financial fraud, product safety problems, harm to consumer, violence against user, increased liability for companies, lack of trust in marketplaces, customer flight, negative public relations

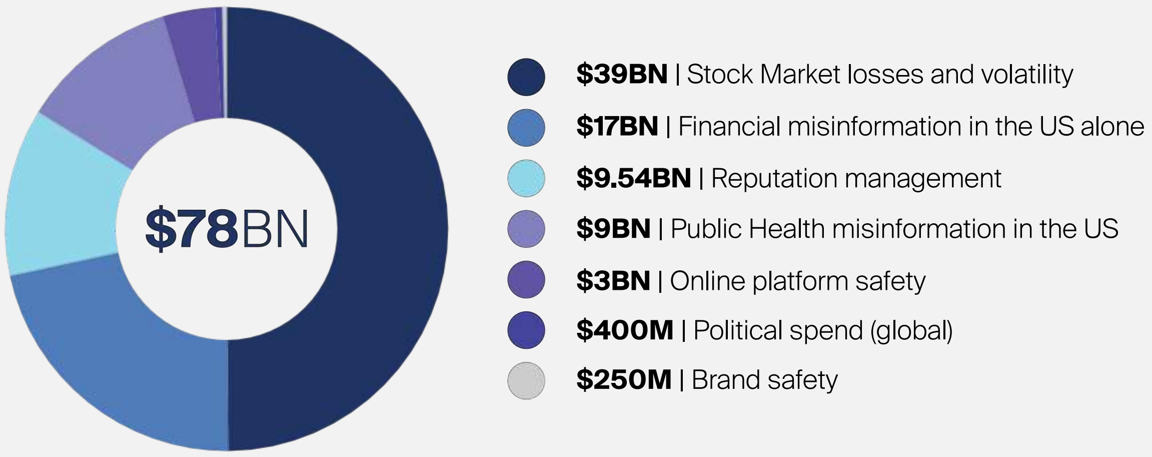


FINANCIAL COSTS
OF MISINFORMATION

Some participants from telecom, healthcare, and banking expressed a sense of frustration that their industries may not fully realize the risks that synthetic media may pose since these threats are hard to recognize individually.

In between the FixFake Symposia events, Israeli technology startup CHEQ and the University of Baltimore issued an estimate of the economic costs of misinformation.[13] The results below are staggering. Given the nascent nature of deepfakes in the commercial space, it is unlikely that companies are able to easily categorize the impacts that they may have already experienced due to misinformation.

While deepfakes and other types of synthetic media represent only a tiny fraction of this risk now, if growth trends of synthetic media mirror similar growth rates to AI enabled marketing technologies, then the market is looking at a 23-29% compound annual growth rate[14] over the coming years.



WHY ARE DEEPFAKES AND DISINFORMATION AN IMMINENT THREAT TO SOCIETY?

BOTTOM LINE The tools to create and disseminate disinformation are easier, faster, cheaper and more accessible than ever.

WHY NOW?

While the term deepfakes has only entered the national lexicon in the last 24 months, it does not take a historian to recognize that manipulating populations through mass media is not a new phenomenon. Even early papyrus scrolls from Egypt show disinformation related to the Pharaoh's exploits. Why the urgency around this issue today?

FixFake Symposia panelists overwhelmingly emphasized the urgency in finding solutions to the problem of manipulated media broadly. Society must take action before a mass proliferation of deepfakes, cheapfakes and manipulated media permanently undermines our ability to trust and be trusted in digital spaces.

1 BROAD ACCESS TO MANIPULATION TECHNOLOGY FOR THE GENERAL PUBLIC IS NEW

Movie studios have been using sophisticated graphics technologies to create special effects for several decades. However, new applications promise tools to create Hollywood caliber content by anyone - regardless of technical skill, budget or incentives. Wide availability of machine-learning algorithms, computer processing power, and data sets for training means barriers to deepfake development are rapidly receding. These technologies, coupled with the advancements in smartphone camera capabilities, are changing the game for content creation.

2 SPEED AND SCALE AT WHICH INFORMATION TRAVELS VIA THE INTERNET

Content can now be disseminated in targeted ways on a global scale through the use of social media and messaging platforms. All of the internet platforms enhance:

i. Discoverability of audiences likely to be receptive to deceptive content.
ii. Financial and social incentives through advertising revenue and social connection.
iii. Access to user-generated content with limited detail on its origins or provenance.

The human brain's innate desire to consume and share novel information fuels the spread of disinformation and stokes the outrage machine. This is particularly dangerous when marginalized populations who lack the tools and education needed to understand and combat the security threats are targeted.

3 MATURITY OF AI-ENABLED MANIPULATION TECHNIQUES

One of the main concerns cited throughout the FixFake Symposia was the lack of consistent standards, best practices and norms for identifying, labeling and responding to manipulated media across digital platforms. AI enabled technologies for content creation are still in the formative stages of development, meaning there is time to influence the development of the technology itself as well as norms around its outputs. This is a critical moment for stakeholders to set standards for how content is created, labeled and identified when built with AI enabled tools. Several technologists shared how their teams work to combat the unintended consequences of deepfake creation technology. These solutions ranged from access controls on the code to consent of those pictured to labeling the content as made with AI.

Today, detecting manipulation is a reactive pursuit rather than a proactive one, leaving the burden to fact-checking organizations at best and uneducated consumers at worst. Technical leaders in the forensics space at each event sounded the alarm that detection capabilities - already limited - will become increasingly difficult as media creation and computer vision technologies advance. Bad actors are already integrating these tools into their campaigns; adopting techniques to create fakes that leap-frog techniques to detect these fakes. Responsible stakeholders must act now to limit their reach and effectiveness.

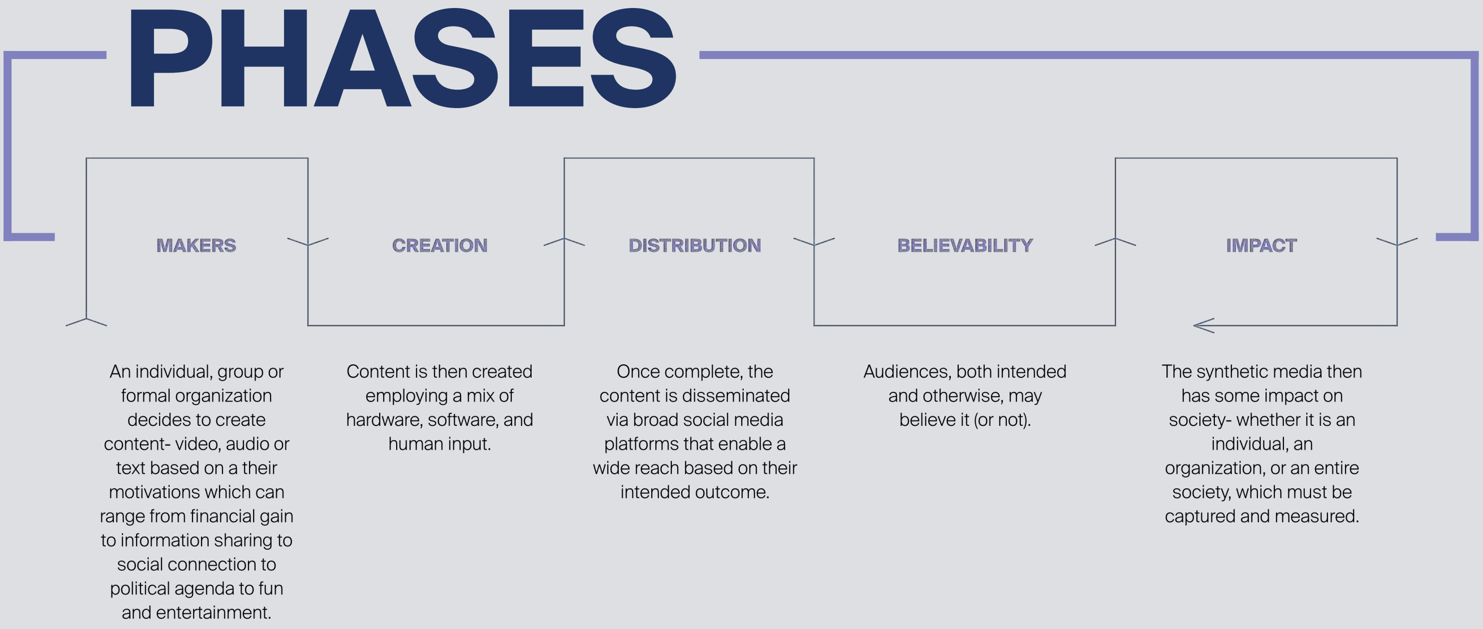
WHAT HAVE BEEN THE PROBLEMS TACKLING DISINFORMATION?

BOTTOM LINE Addressing the problem of disinformation requires convening diverse stakeholders, building tech and policy solutions that evolve as the tactics do, and increasing awareness and adoption of solutions across society broadly. Each layer contains structural and tactical dilemmas.

The FixFake Symposia focused on solutions to these problems. However, before moving to solutions, which we will tackle in Part 2 of our report, it is essential to understand why existing efforts to fight disinformation have faced difficulty. In summary, the majority of organizations focus on a specific industry or set of problems that most directly impact their business priorities. Policymakers fixate on the most visible stakeholders in the ecosystem. Yet, to date, the problem has been deemed too complex to consider in its entirety.

DISINFORMATION DISRUPTION FRAMEWORK

DeepTrust Alliance presents the Disinformation Disruption Framework,[15] a novel blueprint to help funders, researchers and operators assess the disinformation landscape and evaluate interventions. The framework, shown in part below, reveals the multi-layered nature of digital disinformation, demonstrating that no single entity is positioned to remedy the problem at every layer.



PHASES

MAKERS

An individual, group or formal organization decides to create content- video, audio or text based on a their motivations which can range from financial gain to information sharing to social connection to political agenda to fun and entertainment.

CREATION

Content is then created employing a mix of hardware, software, and human input.

DISTRIBUTION

Once complete, the content is disseminated via broad social media platforms that enable a wide reach based on their intended outcome.

BELIEVABILITY

Audiences, both intended and otherwise, may believe it (or not).

IMPACT

The synthetic media then has some impact on society- whether it is an individual, an organization, or an entire society, which must be captured and measured.

The discussions aimed to address these questions through three different lenses:

1. Challenges in convening stakeholders to define and address problems.
2. Challenges in building effective solutions to evolving threat tactics.
3. Challenges in creating awareness and driving adoption of solutions and best practices.

We will explore each in more detail next.

WHAT ARE THE CHALLENGES TO CONVENING A DIVERSE ECOSYSTEM?

BOTTOM LINE There is a core debate within open societies about the trade-offs between data privacy, freedom of speech, detection of harmful content and protection of vulnerable populations. Because content creation and dissemination are impacted by so many different factors, a broad set of stakeholders must work together to set standards and best practices. How is this achieved when trust and incentives for collaboration are lacking?

Disinformation has been recognized as a key threat to society, unleashing a diverse set of organizations and actors working to build solutions. The FixFake Symposia alone convened attendees from more than 100 organizations. DeepTrust Alliance has identified 1,000+ entities working on deepfakes, disinformation, brand safety, data privacy, influence operations, controlled capture, and forensics. We have conducted detailed analysis of over 300 of these global organizations and projects. In this report, we share a snapshot of how the top 100 projects fit into the disinformation disruption framework. While all lists are subjective, selections are based on:

- *Recognized industry leadership including publications, events & citations*
- *Funding/revenue*
- *Impact on the broader ecosystem**

Many organizations like Facebook and Google play multiple roles in the ecosystem from technology creator to grant funder. In this report, we share a snapshot of how the initiatives fit into the Disinformation Disruption Framework. We invite your feedback here. The complete analysis will be available to members of the DeepTrust Alliance.

HIGH-LEVEL FINDINGS

This information is based on DeepTrust Alliance analysis and accurate as of February 10, 2020.

56% are non- profit including foundations, think tanks, advocacy groups and academic institutions.

39% are for-profit including venture capital investors, established enterprise, and startups.

5% are government initiatives.

53% of organization profiled are focused principally on building tools.

24% of organizations profiled focus principally on research, policy or advocacy.

23% of organizations profiled focus principally on funding.

**Includes venture investing (25%), corporate investing and grant-making (8%) and non- profit/ government (67%).*

This is the number of organizations funding work in the disinformation space not the total number of dollars invested/granted. Corporate grant-makers represent a significant portion of the overall money spent in the space; however, total funding numbers are not available for all identified organizations, so results are not reported here.

TOP 100 DISINFORMATION DISRUPTORS

FUNDERS

Non-profit

- Craig Newmark Philanthropies
- Ford Foundation
- Hewlett Foundation
- Knight Foundation
- Luminate
- MacArthur Foundation
- Media Democracy Fund
- Omidyar Network
- Open Society Foundations
- Rita Allen Foundation
- The Atantic Foundation

Venture Capital

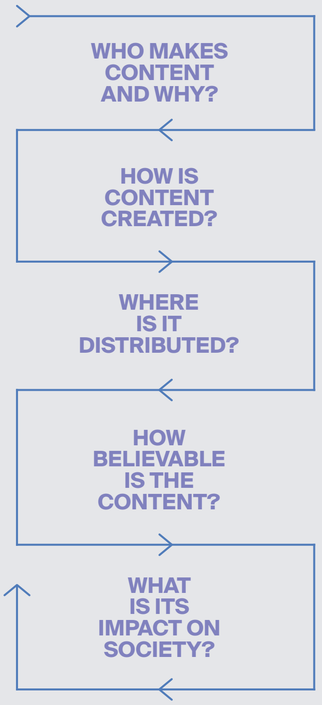
- Betaworks
- LuxCapital
- Samsung Next

Government Funders

- EU Commission
- In-Q-Tel
- National Science Foundation

Corporate Funders/Grant-Makers

- Facebook
- Google
- Microsoft
- Thomson Reuters


WHO MAKES CONTENT AND WHY?
HOW IS CONTENT CREATED?
WHERE IS IT DISTRIBUTED?
HOW BELIEVABLE IS THE CONTENT?
WHAT IS ITS IMPACT ON SOCIETY?

TOOLS

Media Attribution
- Content Authenticity Initiative (Adobe)
- Digimarc
- News Provenance Project
- Pressland

Photo/Video/Speech Integrity
- Amped
- DeepTrace Labs
- Serelay
- Truepic

Synthetic Media Creation
- Adobe
- Industrial Light and Magic
- Lyrebird
- Pinscreen
- Samsung

Fake News Detection/Protection
- AI Foundation
- Bellingcat
- Deepnews.ai
- Jigsaw
- Meedan
- Newsguard

Fact Checking
- Agence France Press
- AP Fact
- Poynter/Politifact
- Snopes

Brand Monitoring
- CHEQ
- Credibility Tech
- CREOPoint
- Edelman
- Factmata

Social Media Analysis
- botswatch
- Dataminr
- Graphika
- Storyful

Device/Hardware Vendors
- Apple
- Nvidia
- Qualcomm
- Samsung

Training/Best Practices (Media)
- Credibility Coalition
- First Draft
- NAMLE
- News Literacy Project
- Partnership on AI
- Reuters/Facebook
- Trust Project

Consulting
- Edelman
- Graphika
- Vidrovr

Media Platforms
- Facebook
- Google
- Instagram
- Microsoft
- Nextdoor
- NYTimes
- Reddit
- Snapchat
- Tiktok
- Twitter
- Vimeo
- Wall Street Journal
- Washington Post
- Whatsapp
- Youtube

Technical Standards
- DeepTrust Alliance
- Digital Object Identifier
- IEEE
- NIST
- W3C Credible Web Coalition

RESEARCHERS

	ACADEMIC	NONPROFIT	GOVERNMENT	CORPORATE
JOURNALISM SPECIFIC	– Duke Reporters Lab – News Integrity Initiative CUNY, Newmark School of Journalism – Reuters Institute – Nieman Lab at Harvard	– First Draft		– Google Digital News Initiative – Facebook Journalism Project
AI/COMPUTER VISION	– Carnegie Mellon University – MIT – Oxford Internet Institute – Stanford University – University of Washington, Allen Institute for AI – USC, Institute for Creative Technologies – Visual Computing Lab, Technical University of Munich		– Census Bureau – Defense Innovation Unit – Department of Homeland Security – NYPD	– Adobe – Facebook – Google – Jigsaw – Microsoft – OpenAI
MEDIA INTEGRITY, FORENSICS	– University at Albany, SUNY – UC Berkeley – New York University, Tandon Engineering	– Mozilla – Internet Archive	– DARPA	– Symantec – Cloudflare
SECURITY, POLICY, LAW, HUMAN RIGHTS	– Center for Media Engagement, UT Austin – Information Disorder Lab at Shorenstein Center (Harvard) – Stanford Cyber Policy Center – Stanford Internet Observatory – The Truthiness Collaboration at USC	– Access Now – Alliance for Securing Democracy – Anti-Defamation League – Avaaz – Center for Human Technology – Data & Society, Disinformation Action Lab – Digital Forensics Research Lab & Disinfo Portal at the Atlantic Council – Electronic Frontier Foundation – Pen America – Wiimedia – Witness	– EU High Level Expert Group on Fakes News and Online Disinformation – United Nations	– Jigsaw

COLLABORATIONS

In addition to the companies, nonprofits and academic institutions mentioned above, we recognize several prevalent modes of collaboration. This list is meant to be illustrative rather than exhaustive. If you have suggestions of other modes of collaboration that we should be aware of, please share them here.

ACADEMIC: RESEARCH FOCUS

Academic- Industry Collaboration

Example:

- Stanford Internet Observatory
- Duke Tech and Check Cooperative

Academic Journals and reports

Example:

- Harvard Kennedy School Misinformation Review
- NYU report on Disinformation and the 2020 Election: How the Social Media Industry Should Prepare

Disinformation Labs

- Duke Reporters Lab
- Indiana University Observatory on Social Media
- Nieman Journalism Lab-Harvard

ORGANIZATION TO ORGANIZATION

Corporate relationships with fact-checking services

- Poynter
- Snopes
- Agence France Presse

Corporate support of nonprofits

- Trust Project
- News Literacy Project
- First Draft

Civil society partnerships

- Witness and First Draft led workshop on synthetic media and deepfakes
- Knight Commission on Truth, Media & Democracy

INDUSTRY-FOCUSED

Journalism/Media:

- First Draft led projects Crosscheck
- EU High level working group on Fake news and Online Disinformation
- PAI Working group on Synthetic Media

Technology:

- Deepfake Detection Challenge
- Digital Object Identifier Standards work

National Security:

- DARPA's Medifor and Semafor Programs

DISINFORMATION EVENTS/ CONFERENCES

Disinformation specific events

- Misinfo Con
- FixFake Symposia
- WEF Misinformation 2020

Industry specific events

- BBC Trust Summit

General events with Disinformation topics

- World Economic Forum 2020
- SXSW
- Identiverse

These efforts represent critical starting points to tackle the enormity of the disinformation problem. However, even effective collaborations to date have been narrow in scope. While providing important learnings and trust building, protecting democracy and the Internet requires scaled-up collaboration, new incentives to engage, and the leadership to drive it all.

Building on this foundation, FixFake Symposia panelists identified three principal obstacles which a multi-stakeholder coalition must overcome in order to unleash collective power and action against disinformation:

Agreement on target goal

- Common understanding and definitions of the problem
- Common terminology to describe the issues were cited by panelists and participants as a fundamental challenge to a shared understanding of priorities.

Incentives

- Today neither regulation nor market conditions incentivize stakeholders to work together in systematic and sustainable ways.
- Funding mechanisms do not seem to privilege or drive collaboration across diverse stakeholders, and there is reluctance from some organizations to share data due to privacy and other business priorities.

Trust

- Organizations may feel that their fundamental values are at odds with other members of the coalition. Misaligned business incentives and priorities for individual contributors can obstruct collaboration.

WHAT ARE THE CHALLENGES TO BUILDING EFFECTIVE SOLUTIONS?

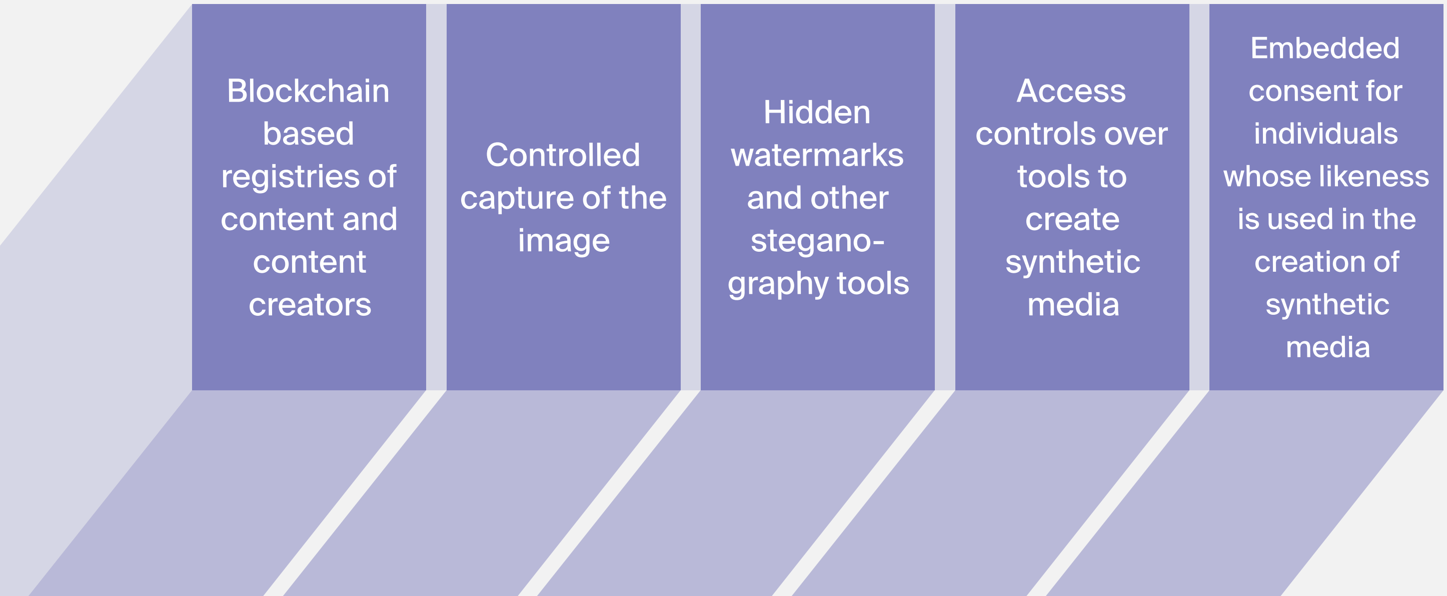
BOTTOM LINE With constantly evolving threats, solution builders find themselves in an arms race against nefarious actors. Long term success requires:

i. Technical solutions that empower information consumers.
ii. Legislative policy that balances privacy and security.
iii. Education about the harms of disinformation, implications and solutions.

Actual change will only come by deploying a portfolio of technology, legislation and education as there is not a 'silver bullet' solution to fix all the problems.

FixFake was inaugurated on the premise that solutions to malicious uses of synthetic media exist and are effective. Many comments began, "Couldn't we solve all of this, if we just..." So often, organizations approach these issues from their specific business problems and fail to see how their current challenges fit into a larger societal dynamic. This has led to a "whack-a-mole" style of problem-solving. The DeepTrust Alliance Disinformation Disruption Framework helps to clarify where and how work can be done in order to mitigate the harmful impact of malicious deep fakes in every sector at risk.

Many market participants bring a technology-first mindset in the pursuit of commercial solutions to mitigate the harm of disinformation. Prevalent solution proposals start at the digital edge with smartphone cameras and GPU hardware moving through to the software that enables creation of synthetic media. Some of these interventions include:



Blockchain based registries of content and content creators

Controlled capture of the image

Hidden watermarks and other stegano-graphy tools

Access controls over tools to create synthetic media

Embedded consent for individuals whose likeness is used in the creation of synthetic media

In terms of developing technological solutions, two hurdles were cited across the conversations:

i. Lack of data standards for identifying content.

ii. Lack of real-world synthetic media and data for building forensic models and tools.

There was broad agreement that the community has significant work ahead to achieve technical standards around the creation and identification of synthetic media. Frustration was expressed over the limited set of available deepfakes for training purposes. While recent initiatives from Google and Facebook have improved the pool, there is still a need for more “in the wild” deepfakes for forensic researchers. Data privacy and business priorities have been key limitations. A big part of the challenge is finding scalable solutions that respect individuals’ right to privacy and freedom of expression.

Despite the focus from entrepreneurs and venture capital on technological solutions, throughout all three events, computer vision and forensics experts adamantly asserted that technology alone will not solve the problem of disinformation. An arms race between “good” technology and “bad” technology will not succeed in sanitizing the playing field because the problem is ultimately one of human behavior. Policy and regulation are critical elements of controlling the creation, dissemination and monetization of synthetic media.

Not surprisingly, given the role of foreign actors in the 2016 elections as uncovered by the Mueller report,[16] legislators have also taken up the problem of synthetic media. At the Washington, DC event we heard from the legislative and committee staffs responsible for promulgating several proposed pieces of legislation. There are currently two main legislative approaches for addressing synthetic media:

i. Research, reporting, funding and general examination of the technology and its accompanying threats.

ii. Criminalization and civil penalties for the creation and the distribution of synthetic media.

There was significant debate about the role of regulation. Some participants felt that by outlawing deepfakes, Congress would immediately force social media platforms to adopt technological tools and policies on synthetic media intended to limit harm at scale. There was also recognition that government -mandated source verification, watermarks, content ID or scanning for harmful content could offer greater transparency and protection to consumers.

Others felt that an outright ban would yield unintended consequences including:

i. Infringement upon the rights of individual, privacy, free speech and expression especially for marginalized populations.
ii. Harm to the United States' competitive position and innovation in computer vision and AI especially relative to China.
iii. Moving disinformation to -end -to- end encrypted messaging platforms which are harder to detect.
iv. Reducing competition and harming startups and smaller organizations.
v. Impairing detection and threat deterrence.
vi. Inability to actually deliver on regulatory requirements given current technology maturity.

Attorneys on the panels advocated for relying on existing laws and statutes to the greatest extent possible in order to avoid controversial heavy-handed regulation. On both sides of the debate, stakeholders felt that greater government understanding of social media business models and manipulation technologies is essential before legislation of any kind should proceed.

Education of legislators, corporate leaders and consumers was another consistent theme across the sessions. Given the rate and pace of innovation in this space, coupled with recent academic research that shows populations over 65 are 7x more likely to share misinformation,[17] it was clear that media literacy programs need to be mandated, expanded and targeted to a wide range of stakeholders. Human-centered design was another important feature of the conversation as both technologists and policy experts advocated that solutions in all their forms must be as accessible as possible to as many individuals as possible. Strategies too complex to integrate and deploy are doomed to fail the threshold test of feasibility.

It's important to note that there was a feeling of frustration among disinformation experts who pointed out that many stakeholders assume new solutions are required when there is an inventory of existing approaches, solutions and laws that should be reviewed, evaluated and applied. In Part 2 of this report, we will review the inventory of possible interventions and prioritize these approaches in detail.

> **Research shows that populations over 65 are 7x more likely to share misinformation.**

WHAT ARE THE CHALLENGES TO AWARENESS AND ADOPTION OF SOLUTIONS?

BOTTOM LINE The implications of synthetic media and disinformation for the information economy are underappreciated, making it difficult to appeal to individuals, organizations and society to adopt existing solutions. There are challenges related to cost, interoperability, security, effectiveness and use at scale for all interventions.

Despite being an old problem, the risks of disinformation are still undervalued. While mainstream media players regularly alert the general public to the risks of synthetic media and disinformation, neither organizations nor individuals are well equipped to identify and deploy mitigations when consuming information on social media platforms. A significant burden falls on the individual information consumer to determine the credibility of content. While Pew Research indicates that consumers have grave general concerns about misinformation,[18] their online habits have not led to wide adoption of practices or tools that would counter these risks.

In order to overcome the risks of disinformation, policy makers must build a deep understanding of the technology and the implications of its use. Cybersecurity can be a model for valuable policy interventions. The Cybersecurity Information Sharing Act of 2015[19] helped to define key terminology and provide parameters to allow threat sharing amongst industry actors. Legislators should also be sure to war-game the unintended consequences of potential legislative proposals.

THE GREATEST CHALLENGE?

At a corporate level, the greatest challenge to overcoming disinformation for most enterprises is building the business case to take action. Marshalling resources to counter disinformation tactics, ensuring interoperability with existing systems and clearly articulating the risk is still a work in progress at most companies. Yet, AI-enabled technologies are already arriving in the cyber fraud units of multiple companies, and greater awareness will help companies to identify distinct threats and deploy mitigations against them. A clear set of use cases by industry along with the financial impacts (as outlined in the threats section of this report) are an essential element toward altering corporate positions and ensuring earlier action.

For individuals, there is a critical debate about the trade-offs between security and liberty. Society needs to have a robust debate about what freedoms and privacy they are willing to trade for greater security. The traditional forums for doing so - Congress, mainstream media and community institutions - are not currently well positioned to handle these debates because their credibility has been directly impacted by disinformation. In addition, they don't have enough information to evaluate the long term trade-offs and impacts on our society. Creating awareness and providing education to these stakeholders is an essential component of countering disinformation.

To effectively overcome the problem of disinformation, society must provide incentives and consequences to industries, organizations and individuals for not implementing available solutions.

CONCLUSION

The collective expertise in the rooms of the FixFake Symposia defined the most pressing problems threatening the credibility of our information ecosystem and generated ideas for the solutions to overcome these problems. Throughout this report, DeepTrust Alliance has aimed to provide a clear rendering of the workshops' proceedings along with background material that augments the expertise. A concise view of the deepfake disinformation problem is the first step toward solving them.

The landscape of disinformation has been clearly defined by the leaders who are shaping solutions and reacting to the threats they see across elections, news cycles and world events.

Now the agenda must address the current research and debate that will move the struggle against disinformation forward.

WHAT'S NEXT?

In Part 2, DeepTrust Alliance will expand upon the market analysis to walk through an inventory of active research and commercial projects as well as a list of solutions that were proposed during the FixFake Symposia. We will also describe our work to disrupt disinformation by fighting deepfakes, cheapfakes, and all manipulated media. DeepTrust convenes stakeholders, builds solutions and drives awareness and adoption as we work to empower decision-makers with truth and trusted information.

IN THE MEANTIME

In the meantime, we invite readers to share take this three minute survey on trust in our world today.

Finally, we invite organizations who are committed to overcoming these problems to join the DeepTrust Alliance to be part of this important work.

THANK YOU,

DEEPTRUST ALLIANCE

ADDITIONAL READING

At DeepTrust Alliance, we recognize that we stand on the shoulders of giants who have gone before us with important work in this space. Beyond the research we have cited throughout the report, we also wanted to share more resources for those interested in going deeper into the topic.

GENERAL

New York Times: Deepfakes: Is this video even real?

Washington Post: Seeing isn't believing

Brookings: Fighting deepfakes when detection fails

Deeptrace Labs: The State of Deepfakes 2019: Landscape, Threats and Impacts

Forbes: AI altered video is a threat to society

How we can protect truth in the age of disinformation

LEGAL/POLICY

WilmerHale: Deepfake Legislation: A Nationwide Survey

Information Disorder: Towards an Interdisciplinary Framework for Policy and Research

TECHNICAL RESEARCH

Generative Adversarial Nets

Protecting world leaders against deepfakes

A survey of protection and verification techniques

Defending against Fake News

Neural imagine pipelines: The Scourge or hope of forensics?

Detecting AI Synthesized Speech Using Bispectral Analysis

ACKNOWLEDGEMENTS

DeepTrust Alliance would like to thank the following individuals who participated as speakers and panelists in the FixFake Symposia events –

NEW YORK

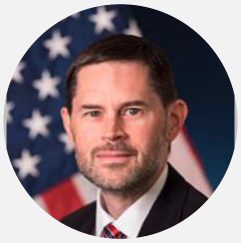
NOVEMBER 8, 2019

Pawel Korus

NYU Center
for Cybersecurity
Tandon School of
Engineering

Siwei Lyu

Computer Vision and
Machine Learning Lab
SUNY Albany Computer
Science



Matt Turek

Information
Innovation Office
Defense Advanced Research
Projects Agency (DARPA)

Marc Lavallee

Head of R&D
The New York Times

Claire Wardle

Executive Chair
First Draft News

Michael Casey

Chief Content
Officer
Coindesk

Jason Gonzalez

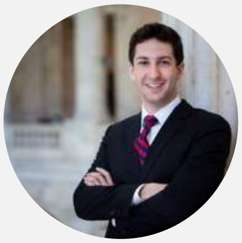
Partner
Nixon Peabody LLP

WASHINGTON D.C.

NOVEMBER 14, 2019

Rob Portman

Senator, (R-Ohio)
Senate Artificial Intelligence Caucus



Sam Mulopulos

Legislative Assistant
Senator Rob Portman

Zach Sorenson

Legislative Aide
Representative Adam Schiff, (D-NY)

Matt Ferraro

Senior Associate
Wilmer Hale

Jeremy Gilbert

Director of Strategic Initiatives
The Washington Post

Ben Sheffner

SVP & Associate General Counsel
Motion Picture Association

Marc Lavallee

Head of R&D
The New York Times

Chaitra Chandrasekhar

Partner
Oliver Wyman

SAN FRANCISCO

DECEMBER 10, 2019

Hany Farid

Center for Computational Science
UC-Berkeley

Maneesh Agrawala

Brown Institute for Media Innovation
Stanford University
Computer Science

Chris Guess

Duke Reporters' Lab
Duke University
Sanford School
of Public Policy

John Diaz

Editorial Page Editor
San Francisco Chronicle

Jason Gonzalez

Partner
Nixon Peabody LLP

Sherif Hanna

VP of Ecosystem Development
Truepic

Roger MacDonald

Director
TV News Archive at the Internet Archive

Leila Zia

Head of Research
Wikimedia Foundation

Jon Miller

San Francisco Giants

The views and recommendations contained in this report do not necessarily represent the views of individual workshop speakers, participants or their employers.

END NOTES

1. Peele, Jordan and Buzzfeed Video. “You won’t believe what Obama says in this video!” YouTube. April 17, 2018, https://www.youtube.com/watch?v=cQ54GDm1eL0

2. “Number of Smart Phones in the World,” Telecommunications, Statista, accessed January 31, 2020, https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

3. Donovan, Joan and Britt Paris, “Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence.” Report, Data & Society, September 18, 2019, https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf.

4. Donovan, Joan and Britt Paris, “Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence.” Report, Data & Society, September 18, 2019, https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf.

5. Breland, Ali, “The Bizarre and Terrifying Case of the Deepfake Video that Helped Bring an African Nation to the Brink,” Mother Jones, March 15, 2019, Politics, https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

6. Bongo, Ali Ondimba. “Adresse à la nation du Président de la République gabonaise du 31.12.18” Facebook. December 31, 2018. https://www.facebook.com/watch/?v=354552671762969

7. Stupp, Catherine, “Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case,” Wall Street Journal, August 30, 2019, Pro Cyber News, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

8. Fisher, Max, “Syrian Hackers Claim AP Hack that Tipped Stock Market by $136 Billion. Is It Terrorism?” Washington Post, April 23, 2013, WorldViews, https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackers-claim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/

9. Fisher, Max, “Syrian Hackers Claim AP Hack that Tipped Stock Market by $136 Billion. Is It Terrorism?” Washington Post, April 23, 2013, WorldViews, https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackers-claim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/

10. Ajder, Henry, Giorgio Patrini, Francesco Cavalli & Laurence Cullen, “The State of Deepfakes: Landscape, Threats and Impact.” Report, Deeptrace Labs, September, 2019, https://deeptracelabs.com/archive/.

11. “I Was Vomiting: Journalist Rana Ayyub Reveals Horrifying Account of Deepfake Porn Plot,” India Today, November 21, 2018, https://www.indiatoday.in/trending-news/story/journalist-rana-ayyub-deepfake-porn-1393423-2018-11-21.

12. Photo credit: Saumya Khandelwal/HT PHOTO https://www.hindustantimes.com/india-news/let-s-talk-about-trolls-women-are-molested-virtually-on-a-daily-basis-rana-ayyub/story-ehKczykSQcnhHe60eAujvN.html

13. Cavazos, Robert, “The Economic Cost of Bad Actors on the Internet.” Report, University of Baltimore CHEQ, 2019, https://www.cheq.ai/fakenews.

14. Columbus, Louis, “Roundup of Machine Learning Forecasts and Market Estimates for 2019,” Forbes, March 27, 2019. https://www.forbes.com/sites/louiscolumbus/2019/03/27/roundup-of-machine-learning-forecasts-and-market-estimates-2019/#61fb556e7695

15. The Disinformation Disruption Framework was developed independently by DeepTrust Alliance with feedback from Siwei Lyu and Matt Turek; as well as input from discussions during the FixFake Symposia events. This work builds on similar work done by Aviv Ovadya: https://docs.google.com/document/d/1HWNbavVoWc_kOxWybK-jCuRxem-FJeZWe2EFDz6EZ2o/edit#heading=h.o7mo62mpkly; Camille Francois: https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC_Framework_2019_Sept_2019.pdf.

16. Report on the Investigation into Russian Interference in the 2016 Presidential Election, Vol I, Special Counsel Robert S. Mueller, III, Washington, D.C., March 2019, https://www.justice.gov/storage/report.pdf.

17. Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. “Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook.” Science Advances, Vol. 5, no. 1, DOI: 10.1126/sciadv.aau4586, https://advances.sciencemag.org/content/5/1/eaau4586.

18. Mitchell, Amy, et al., “Many Americans Say Made-Up News Is a Critical Problem That Needs to Be Fixed,” Pew Research Center, Journalism & Media, June 5, 2019, https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.

19. Cybersecurity Information Sharing Act of 2015, S.754, 114th Cong. (2015). https://www.congress.gov/bill/114th-congress/senate-bill/754.

The author of this report is Kathryn Harrison (Executive Director, DeepTrust Alliance), with contributions from David Toomey, Sara Aros, Aviv Ovayda and Jane Lyons.

Responsibility for the final content of this report rests entirely with DeepTrust Alliance.

We are extremely grateful to Dmitri Mehlhorn of Investing in Us, Brooke Furst of Nixon Peabody, Sam Mulopulos of Senator Rob Portman, (R-Ohio) Christopher Hoven and Zach Sorenson of Representative Adam Schiff, (D-California)'s team, Sid Gandhi and Rashi Ranka of Deepfake Blue, and Ron Quaranta of the Wall Street Blockchain Association for your sponsorship and support of the FixFake Symposia series.

Thank you to all of our speakers for making time to join us and bring your unique insights to this important topic. In particular, thank you to Siwei Lyu, Pawel Korus, Matt Turek, Aviv Ovadya and Sam Gregory for your feedback on the Disinformation Disruption framework.

We thank our volunteers Jason Brett, Mike Leszcz, Ben Lyons, Christopher Harrison and those who helped to make the FixFake Symposia possible.

Lastly, we applaud the nearly 200 FixFake attendees and participants for your active engagement, energy and vital perspectives.

All graphic design work on this report was done by Shae Riley.

Copy-editing was done by Sara Aros.

ABOUT DEEPTRUST ALLIANCE

The DeepTrust Alliance is a non-profit membership organization bringing breakthrough technology to fight the malicious use of deepfakes and disinformation.