RESEARCH BRIEF

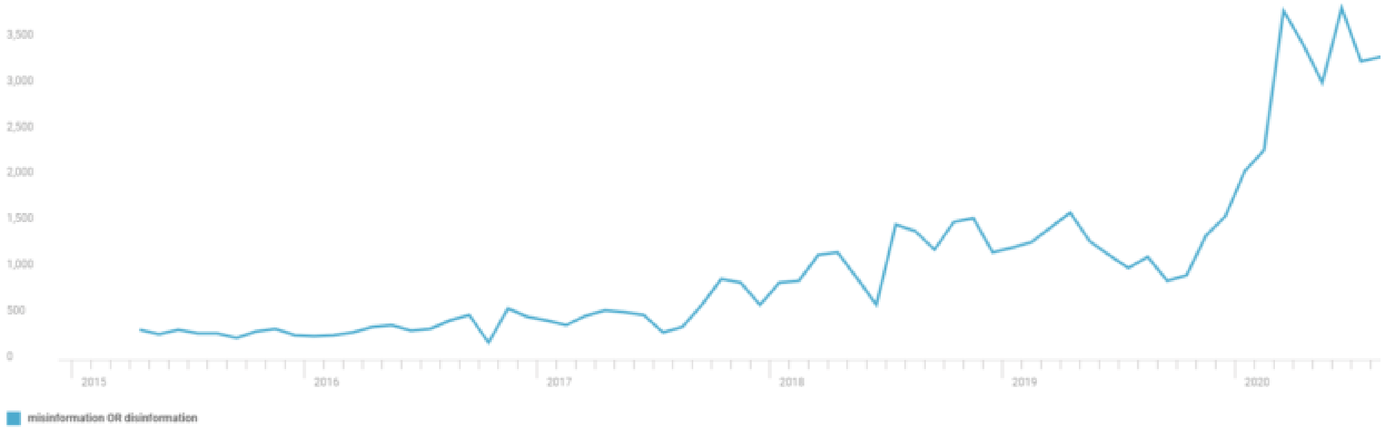
Disinformation That Kills: The Expanding Battlefield Of Digital Warfare

October 21, 2020

Artificial Intelligence Cybersecurity TMT

Since 2016, the digital battlefield has become more sophisticated and widespread across the globe. False information about major events from the Covid-19 outbreak to the 2020 US election is jeopardizing public health and safety. Here we dig into how modern warfare is being waged on the internet, and the steps being taken to stop it.

Since the revelations of Russia spreading false information via social media to interfere in the 2016 US election, the digital battlefield has expanded to include new players and more sophisticated techniques.



First, the Covid-19 outbreak spawned an infodemic, with false information about virus treatments and conspiracy theories proliferating online.

Then, ahead of the 2020 US presidential election, National Counterintelligence and Security Center director William Evanina warned of nations — including Iran, China, and others — seeking to "use influence measures in social and traditional media in an effort to sway U.S. voters' preferences and perspectives, to shift U.S. policies, to increase discord and to undermine confidence in our democratic process."

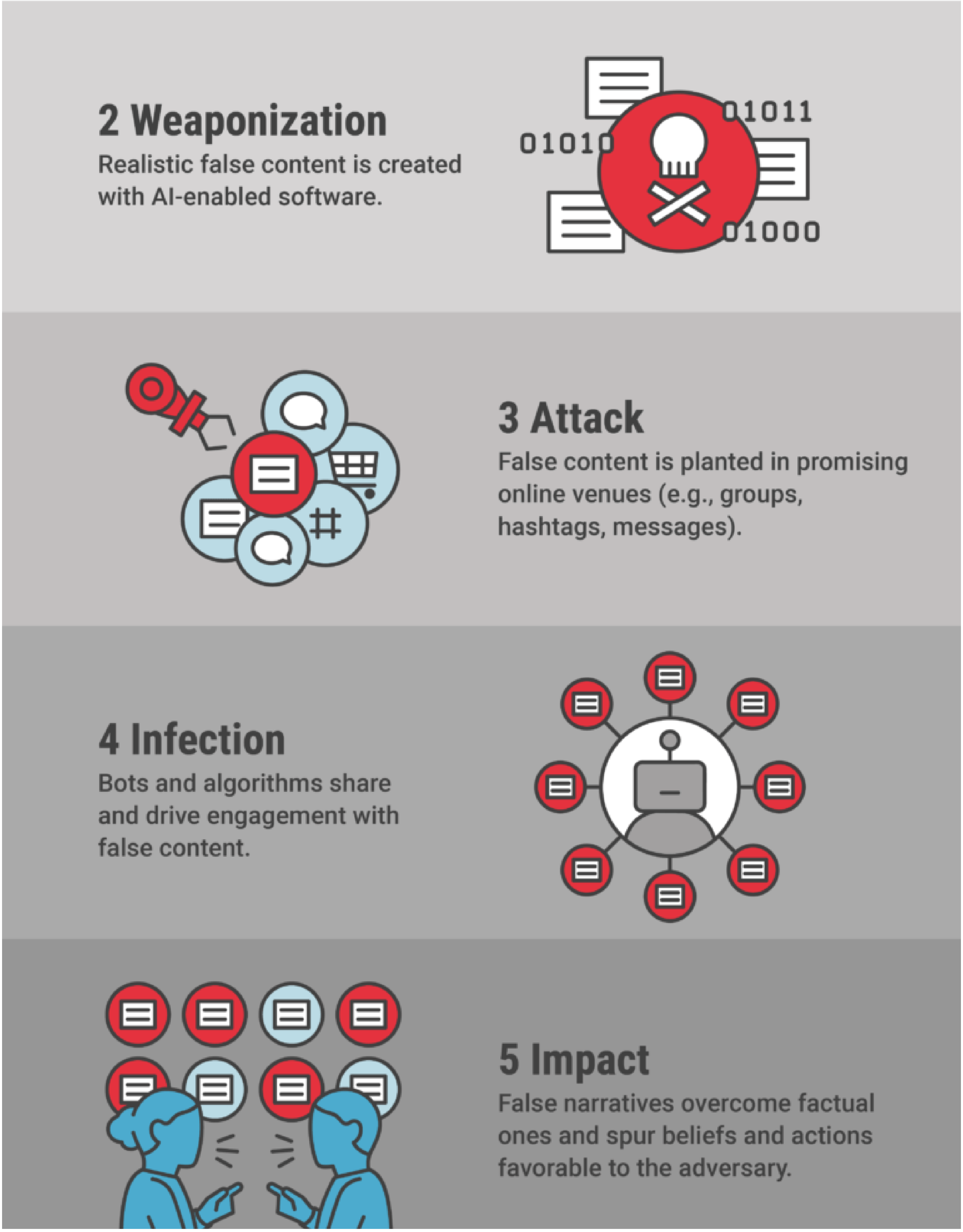
The stakes are high. Disinformation threatens the globe's ability to effectively combat a deadly virus and hold free and fair elections. Meanwhile, emerging technologies like artificial intelligence are accelerating the threat — but they also offer hope for scalable solutions that protect the public.

Below, we examine the evolving technologies and tactics behind the spread of disinformation, as well as how it impacts society and the next generation of war.

Future Of Information Warfare

Citizens will find themselves in the digital crossfire of international conflicts that exploit online platforms with convincing disinformation campaigns.

CBINSIGHTS



2 Weaponization

Realistic false content is created with AI-enabled software.

3 Attack

False content is planted in promising online venues (e.g., groups, hashtags, messages).

4 Infection

Bots and algorithms share and drive engagement with false content.

5 Impact

False narratives overcome factual ones and spur beliefs and actions favorable to the adversary.

The state of information warfare: Where are we today?

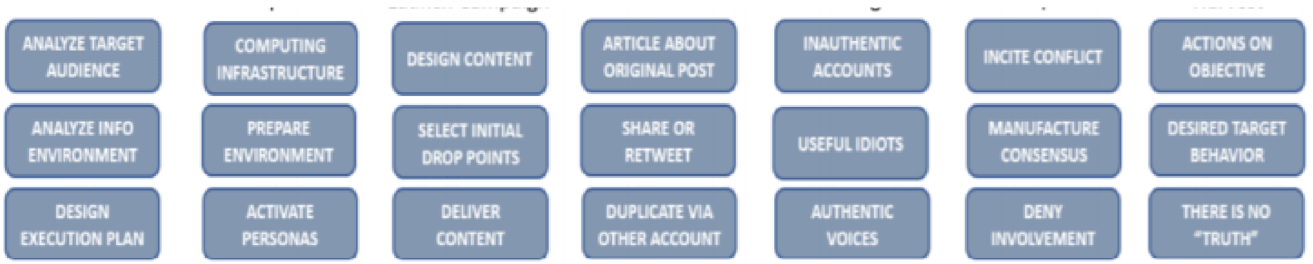
Though it’s tempting to view disinformation as a modern issue driven by world events and ever-evolving communications technology, it’s important to remember its historical context. Disinformation is not new — it has simply adapted to meet today’s technical and political realities.

> “Since the Cold War, propaganda has evolved in a direction opposite to that of most other weapons of war: it has become more diffuse and indiscriminate, not less.” — Joshua Yaffa, the New Yorker

Because the cost and technical skills required for executing an online disinformation campaign are remarkably low, the number of actors and the amount of malicious content have increased.

Further, the threat is not limited to foreign actors — it includes domestic entities as well. In fact, domestic players are expected to have a greater impact on the 2020 US election than foreign actors, according to a report by New York University (NYU).

Attackers execute disinformation campaigns using a 7-step process, which the US Department of Homeland Security outlines below:



ANALYZE TARGET AUDIENCE	COMPUTING INFRASTRUCTURE	DESIGN CONTENT	ARTICLE ABOUT ORIGINAL POST	INAUTHENTIC ACCOUNTS	INCITE CONFLICT	ACTIONS ON OBJECTIVE
ANALYZE INFO ENVIRONMENT	PREPARE ENVIRONMENT	SELECT INITIAL DROP POINTS	SHARE OR RETWEET	USEFUL IDIOTS	MANUFACTURE CONSENSUS	DESIRED TARGET BEHAVIOR
DESIGN EXECUTION PLAN	ACTIVATE PERSONAS	DELIVER CONTENT	DUPLICATE VIA OTHER ACCOUNT	AUTHENTIC VOICES	DENY INVOLVEMENT	THERE IS NO "TRUTH"

Note: A disinformation threat actor may skip steps in the kill chain process. However, doing so can reduce the effectiveness of the campaign and erode protections aimed at obfuscating the identity and objectives of the actor.

Source: The MITRE Corporation

Source: Department of Homeland Security

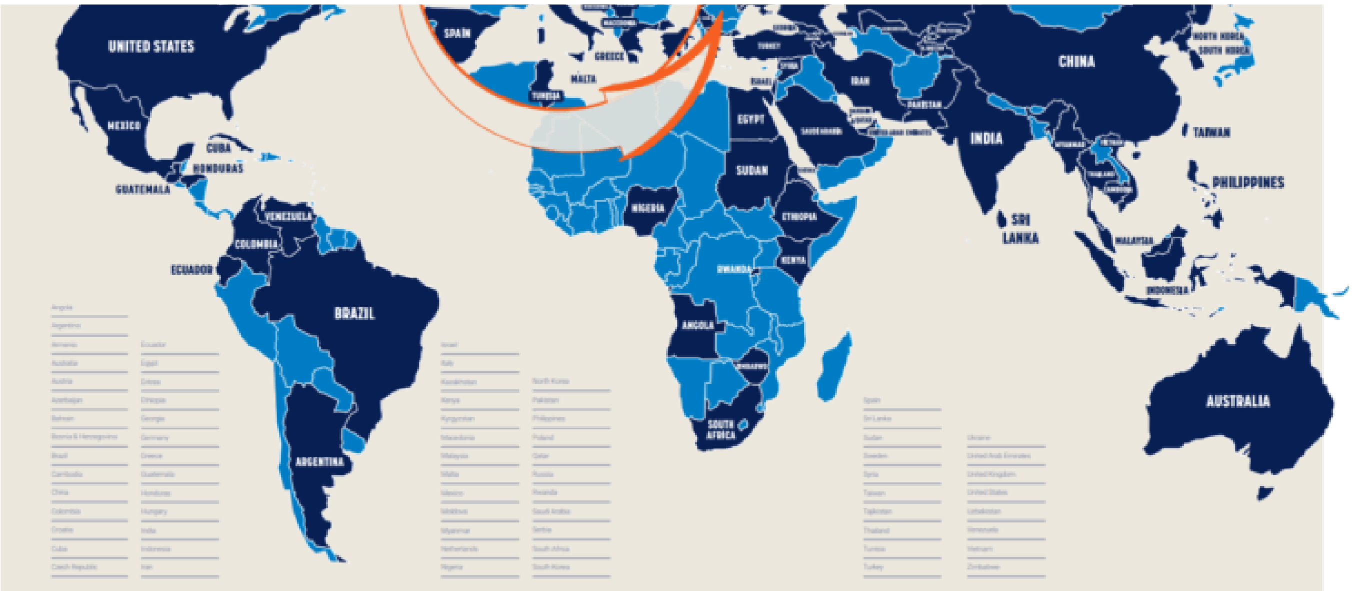
It all begins with an objective, such as generating support for removing sanctions. From there, attackers analyze their target audience's online behavior, develop online infrastructure (e.g. websites), and distribute manipulative content (e.g. social media posts). To make the narrative more compelling, attackers create supporting information like news articles and amplify the message with computer-controlled accounts or bots.

Over the past few years, this 7-step process has unfolded time and again as nations work to perfect it with new technologies and audiences. Attempts to get the problem of disinformation under control have often fallen flat.

One new technology hampering efforts to combat disinformation is WhatsApp. The communications app encrypts messages to prevent others from reading or monitoring them. The downside of this security is that it creates an environment where disinformation can thrive:

- In 2018, false rumors of a roaming group of kidnappers spread on the app, resulting in the killing of more than 24 people in India.
- In Brazil, sophisticated disinformation campaigns linking virus vaccination to death spread on WhatsApp and thwarted the government's efforts to vaccinate citizens against the spread of yellow fever in 2018.
- In 2020, Covid-19-related misinformation, from fake cures to false causes (such as 5G technology), spread across the platform.

Seventy countries used online platforms to spread disinformation in 2019 — an increase of 150% from 2017. Most of the efforts focused domestically on suppressing dissenting opinions and disparaging competing political parties. However, several countries — including China, Venezuela, Russia, Iran, and Saudi Arabia — attempted to influence the citizens of foreign countries.



Source: 2019 Global Inventory Of Organised Social Media Manipulation

The increase of countries involved in spreading disinformation and of domestically created content coincides with a deluge of false information online. In the case of Covid-19, an April 2020 poll found that nearly two-thirds of Americans saw news about the coronavirus that seemed completely made up.

Most Americans believe that fake news causes confusion about the basic facts of the Covid-19 outbreak. The World Health Organization has spoken about how crucial access to correct information is: “The antidote lies in making sure that science-backed facts and health guidance circulate even faster, and reach people wherever they access information.”

Key elements of the future of digital information warfare

> “We live in an **age of disinformation**. Private correspondence gets stolen and leaked to the press for malicious effect; political passions are inflamed online in order to **drive wedges into existing cracks** in liberal democracies; perpetrators **sow doubt** and deny malicious activity in public, while covertly ramping up behind the scenes.” — Thomas Rid, “Active Measures: The Secret History of Disinformation and Political Warfare”

Key tactics include:

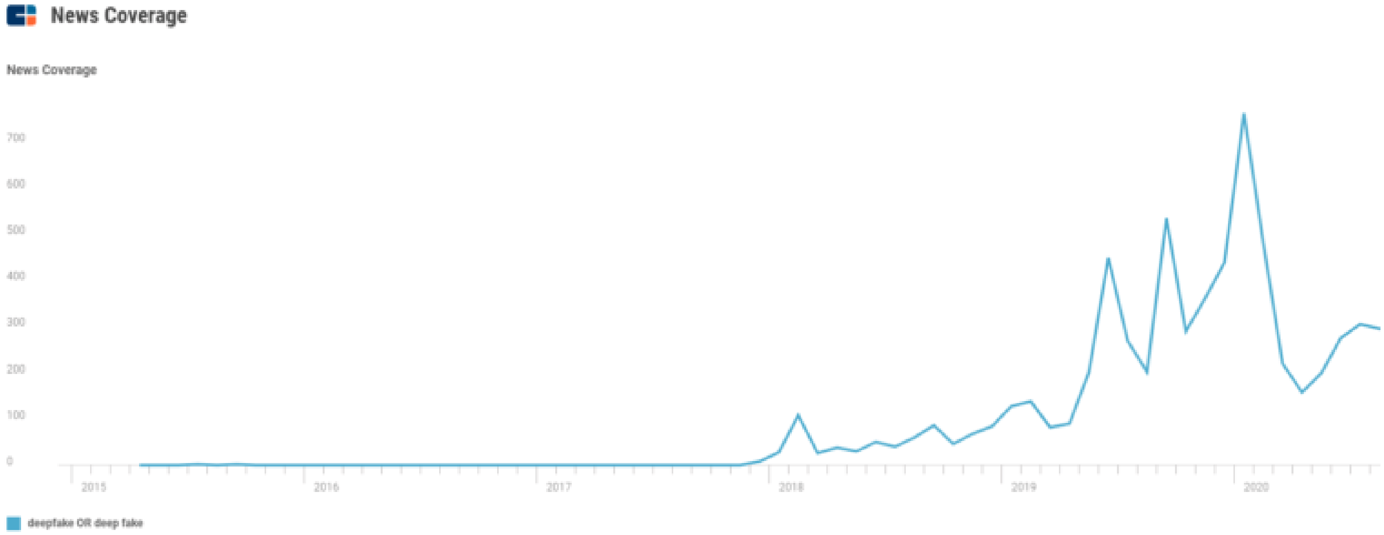
3. **Computational propaganda:** exploiting social media, human psychology, rumor, gossip, and algorithms to manipulate public opinion.

1. DIPLOMACY & REPUTATIONAL MANIPULATION: FAKING VIDEO AND AUDIO

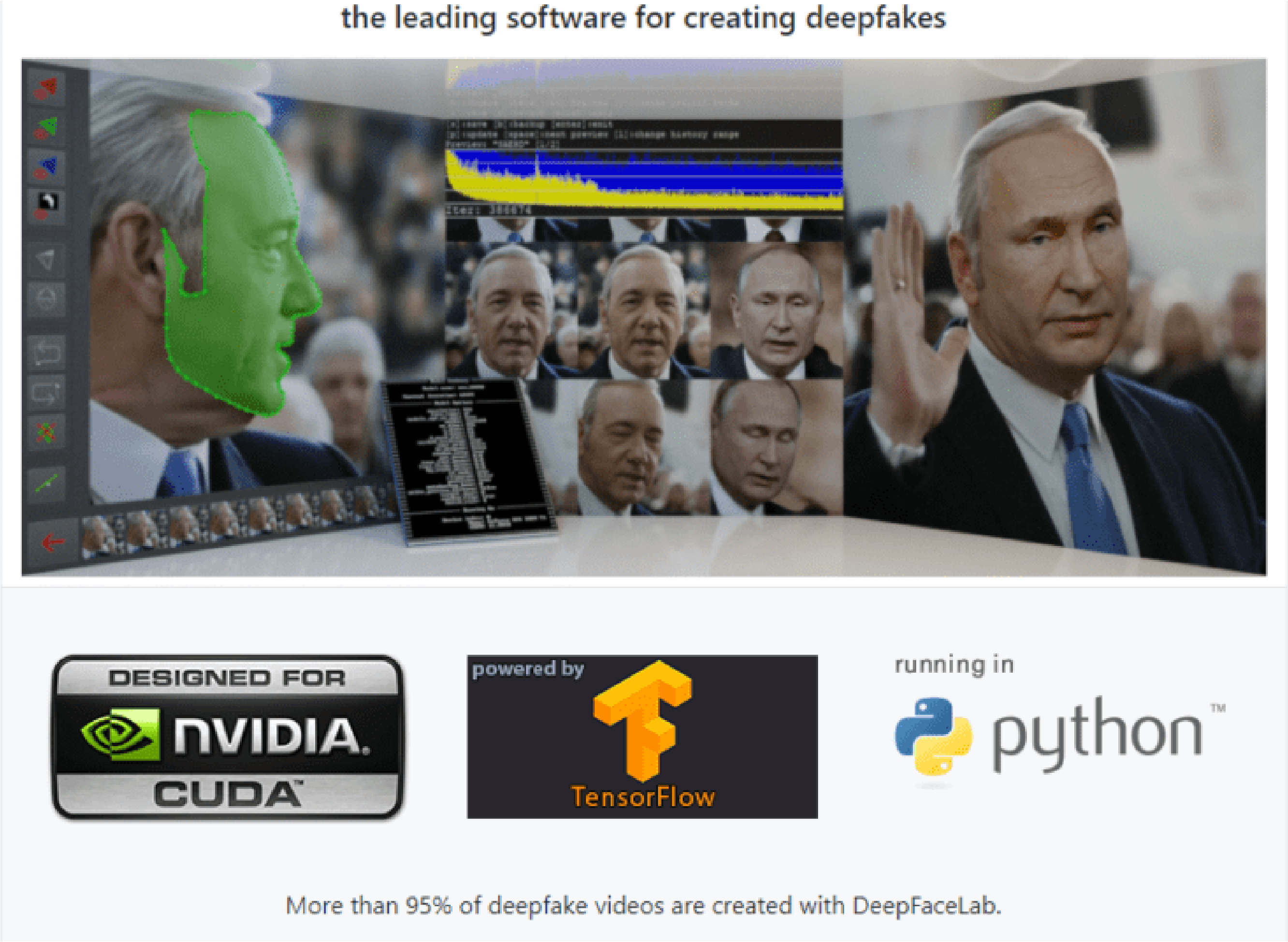
Correctly timing the release of a controversial video or audio recording could compromise peace talks, blow up trade negotiations, or influence an electoral outcome.

Similarly, malicious actors' ability to persuasively impersonate world leaders on digital platforms poses a reputational risk to these individuals. These threats may seem extreme or unlikely, but advances in technology are quickly bringing them closer to reality.

One crucial example is deepfakes, or hyper-realistic doctored images and videos. Deepfakes are a recent development, with news coverage only beginning to pick it up in 2018.



Historically, creating realistic fake images and videos required extensive editing expertise and custom tools. Today, GANs (generative adversarial networks), a type of AI used for unsupervised learning, can automate the process and create increasingly sophisticated fakes.

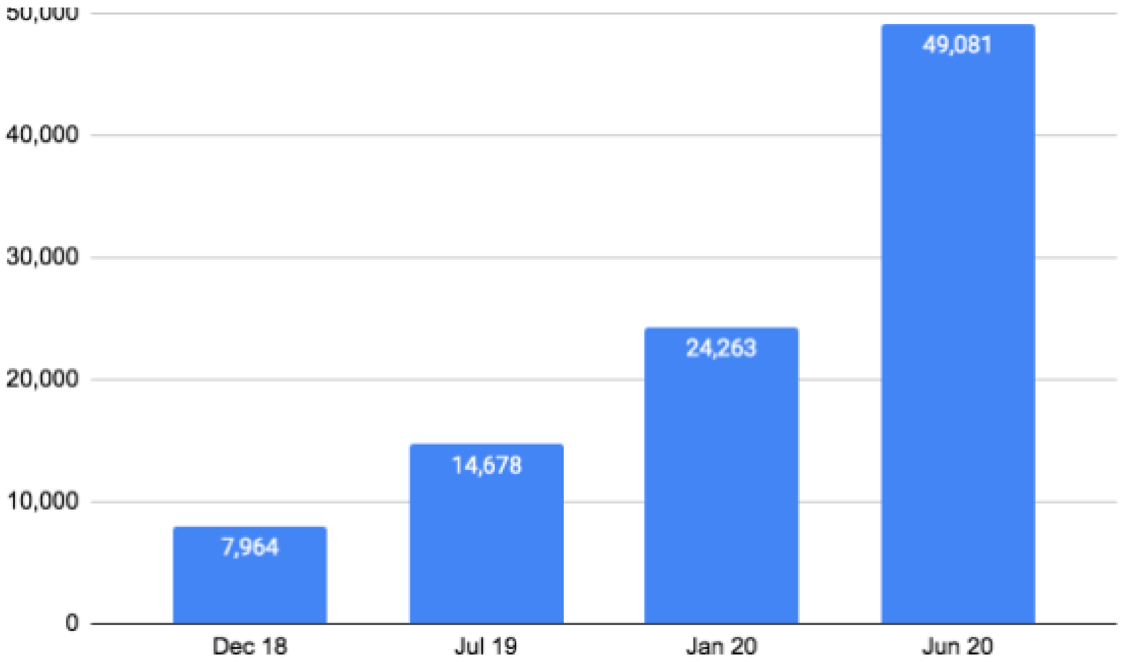


the leading software for creating deepfakes

More than 95% of deepfake videos are created with DeepFaceLab.

Source: DeepFaceLab

Much of the code for creating convincing deepfakes is open-source and included in software packages like DeepFaceLab, which is available publicly for anyone to access. This lowers the barriers for adoption, making deepfakes a viable tool for more hackers, whether or not they are tech-savvy.



Source: Sensity

Unsurprisingly, the number of deepfakes online has exploded over the last few years. According to Sensity, a startup tracking deepfake activity, the number doubles roughly every 6 months.

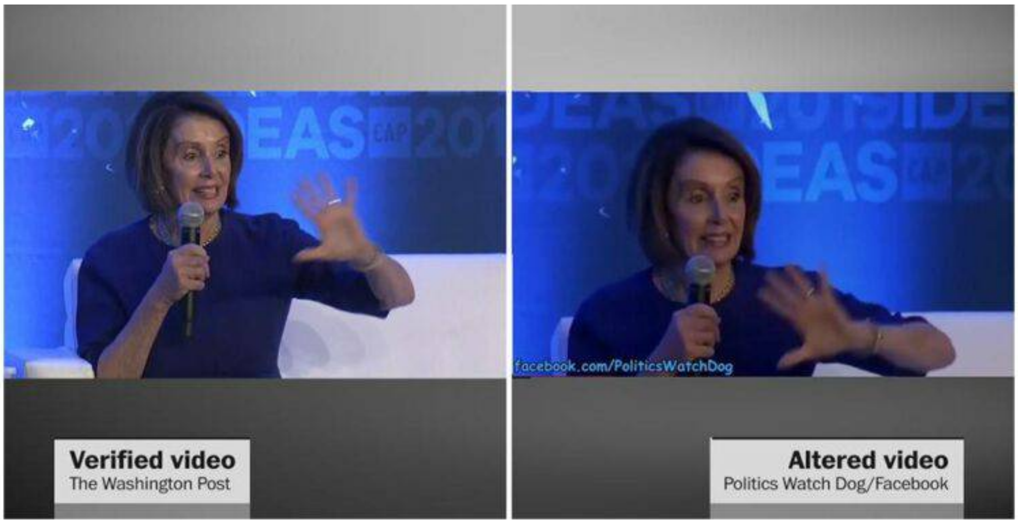
The vast majority of deepfake videos focus on pornography. However, a small percentage of them have political aims.

For example, a 2018 video of Gabonese president Ali Bongo contributed to an attempted coup by the country's military. The president's appearance and the suspicious timing of the video, which was released after several months during which the president was absent receiving medical care, led many to claim it was a deepfake. This perceived act of deception cast further doubt on the president's health and served as justification for his critics to act against the government.

Another instance occurred in June 2019, when a video depicting sexual acts by the Malaysian minister of economic affairs Azmin Ali created political controversy. In defense, Azmin Ali and his supporters delegitimized the video by calling it a deepfake.

In both cases, analyses of the videos to determine their authenticity were inconclusive — a fairly typical outcome for videos of lesser-known individuals or where only the manipulated version exists.

This uncertainty could give rise to an alarming phenomenon known as the "Liar's Dividend," where anyone can feasibly deflect responsibility by declaring an image or video as fake — a tactic aimed at undermining truth. Essentially, the fact that deepfakes exist at all creates an environment vulnerable to manipulation that threatens truth.

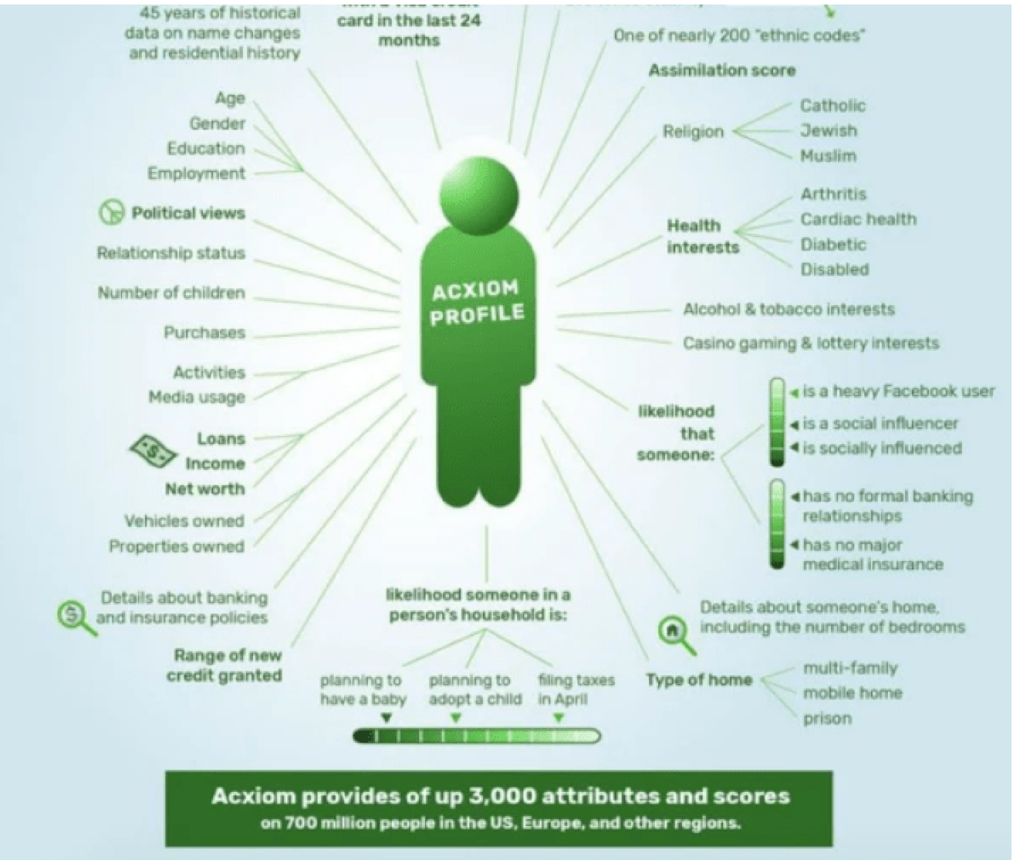


Source: Washington Post

For example, a shallowfake of the speaker of the House of Representatives Nancy Pelosi in May 2019 gave the impression that she was drunk and slurring her words. The video, which was retweeted by US president Donald Trump, received more than 2M views in 48 hours.

Whether a deepfake or shallowfake, the accessibility and potential virality of doctored videos threaten public figures' reputation and governance.

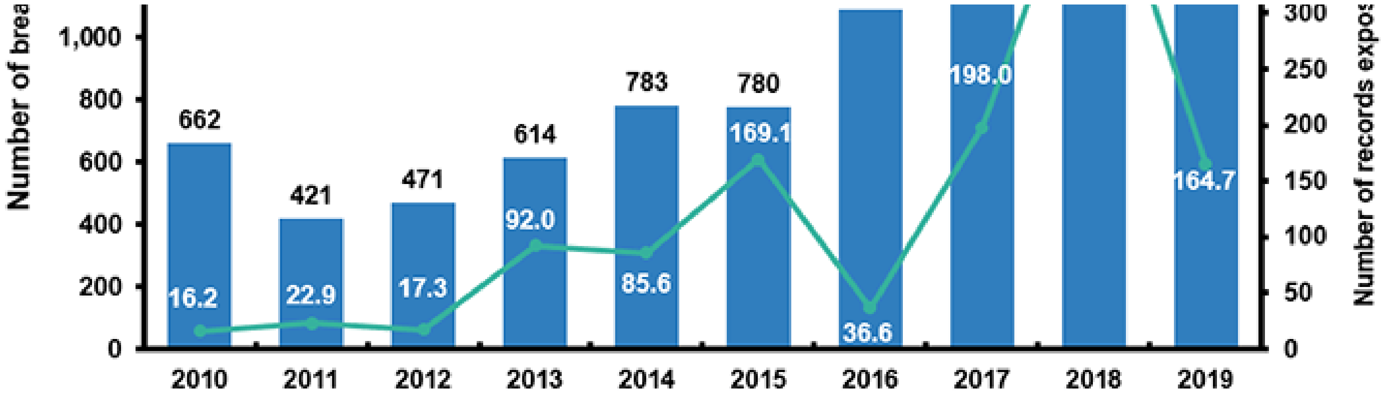
2. AUTOMATED LASER PHISHING: MALICIOUS AI IMPERSONATING AND MANIPULATING PEOPLE


45 years of historical data on name changes and residential history
card in the last 24 months
One of nearly 200 "ethnic codes"
Assimilation score
Age
Gender
Education
Employment
Religion
Catholic
Jewish
Muslim
Political views
Relationship status
Number of children
Purchases
Activities
Media usage
Loans
Income
Net worth
Vehicles owned
Properties owned
ACXIOM PROFILE
Health interests
Arthritis
Cardiac health
Diabetic
Disabled
Alcohol & tobacco interests
Casino gaming & lottery interests
likelihood that someone:
is a heavy Facebook user
is a social influencer
is socially influenced
has no formal banking relationships
has no major medical insurance
Details about banking and insurance policies
likelihood someone in a person's household is:
Details about someone's home, including the number of bedrooms
Range of new credit granted
planning to have a baby
planning to adopt a child
filing taxes in April
Type of home
multi-family
mobile home
prison
Acxiom provides of up 3,000 attributes and scores
on 700 million people in the US, Europe, and other regions.

Source: Cracked Labs via Fast Company

The users and applications of these personal data points vary. Advertising companies frequently use them to target individuals with personalized ads, while other companies use them to design new products and political campaigns use them to target voters.

Malicious actors also have uses for the information, and owing to frequent data breaches that resulted in the loss of nearly 500M personal records in 2018 alone, it's often accessible.



Source: Identity Theft Resource Center via Insurance Information Institute

Personal data plays a significant role in the early stages of a disinformation campaign. First, malicious actors can use the information to target individuals and groups sympathetic to their message.

Second, hackers may use personal data to craft sophisticated phishing attacks to collect sensitive information or hijack personal accounts.

Population targeting

Targeting audiences, whether through ads or online reconnaissance, is a crucial piece of the disinformation chain.

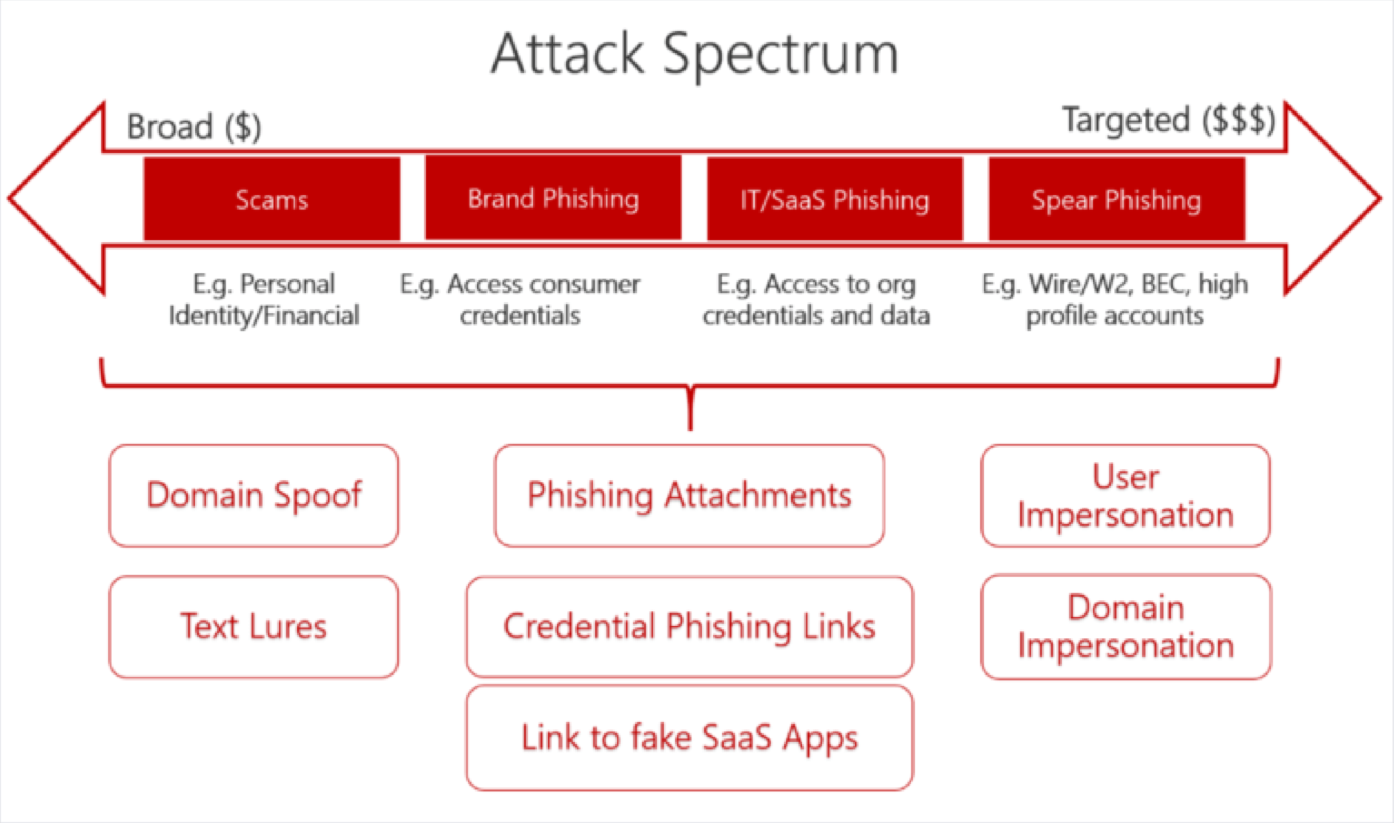
To exploit tensions within society and sow division, purveyors of disinformation target specific groups with content that supports their existing biases. This increases the likelihood that the content will be shared and that the foreign entity's goal will be achieved.

Russia's information warfare embodies this tactic. A congressional report found that the nation targeted race and related issues in its 2016 disinformation campaigns.

While online reconnaissance can identify the social media groups, pages, and forums most hospitable to a divisive or targeted message, buying online ads provides another useful tool for targeting individuals meeting a particular profile.

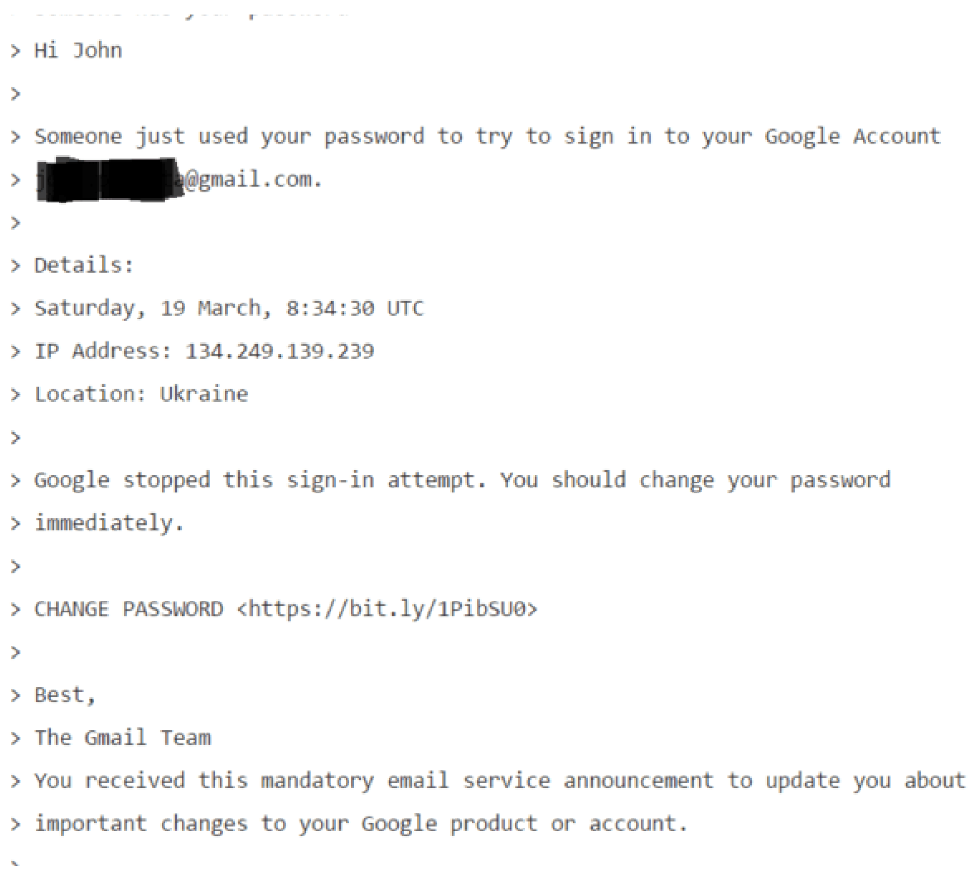
In the lead-up to the 2020 US presidential election, an unknown entity behind the website "Protect My Vote" purchased hundreds of ads that yielded hundreds of thousands of views on Facebook. Promoting fears of voter mail fraud, these ads targeted older voters in specific swing states that were more likely to be sympathetic to the message. The ads made unsubstantiated claims and, in one instance, misconstrued a quote by basketball star Lebron James.

Personalized phishing



Source: Microsoft

If a hacker's phishing attack successfully steals credentials or installs malware, the victim may face reputational damage. For example, ahead of the 2016 US presidential election, Russian hackers used a spear-phishing campaign to infiltrate Hillary Clinton's campaign chairman John Podesta's email and release collected information to the public.

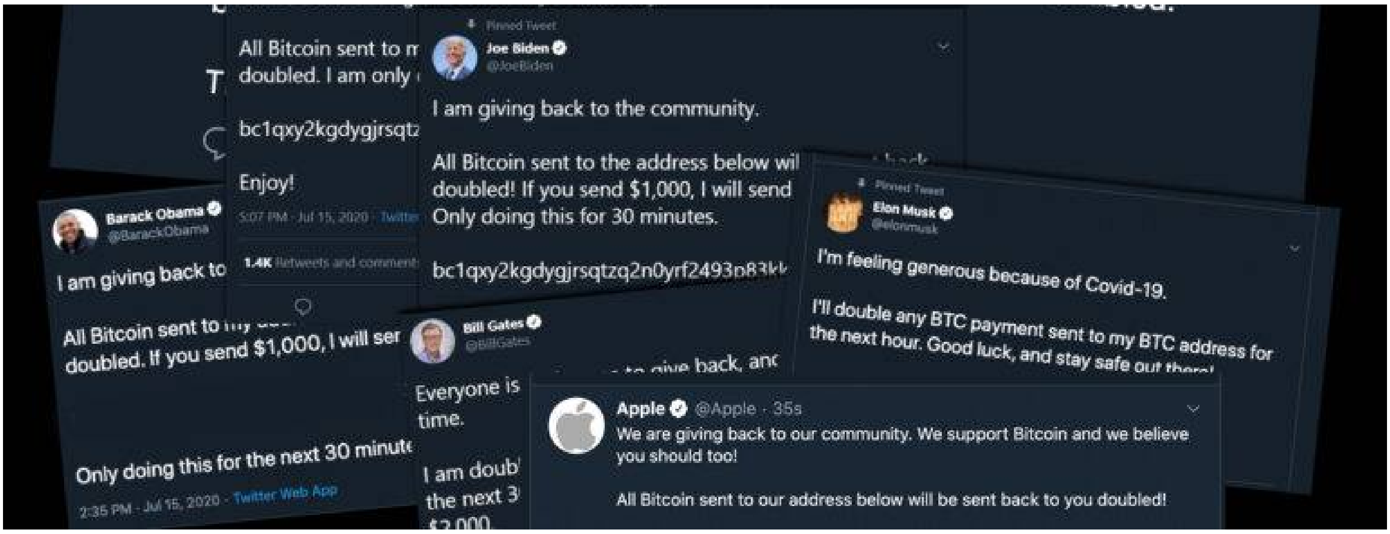


```
> Hi John
>
> Someone just used your password to try to sign in to your Google Account
> [redacted]@gmail.com.
>
> Details:
> Saturday, 19 March, 8:34:30 UTC
> IP Address: 134.249.139.239
> Location: Ukraine
>
> Google stopped this sign-in attempt. You should change your password
> immediately.
>
> CHANGE PASSWORD <https://bit.ly/1PibSU0>
>
> Best,
> The Gmail Team
> You received this mandatory email service announcement to update you about
> important changes to your Google product or account.
```

Source: CBS News

Selectively sharing personal or sensitive information provides disinformation campaigns with a sense of authenticity, and leaking the information ahead of significant events increases its impact.

Using phishing attacks to access high-profile individuals' email or social media accounts also poses a reputational and diplomatic threat. For example, in 2020, hackers gained access to the Twitter accounts of Joe Biden, Elon Musk, and Barack Obama, among others.



All Bitcoin sent to m
doubled. I am only
bc1qxy2kgdygjrsqtz
Enjoy!
5:07 PM · Jul 15, 2020 · Twitter
1.4K Retweets and comments

Pinned Tweet
Joe Biden @JoeBiden
I am giving back to the community.
All Bitcoin sent to the address below wil
doubled! If you send $1,000, I will send
Only doing this for 30 minutes.
bc1qxy2kgdygjrsqtzq2n0yrf2493p83kk

Barack Obama @BarackObama
I am giving back to
All Bitcoin sent to my
doubled. If you send $1,000, I will ser
Only doing this for the next 30 minute
2:35 PM · Jul 15, 2020 · Twitter Web App

Pinned Tweet
Elon Musk @elonmusk
I'm feeling generous because of Covid-19.
I'll double any BTC payment sent to my BTC address for the next hour. Good luck, and stay safe out there!

Bill Gates @BillGates
to give back, and
Everyone is
time.
I am doub
the next 3
$2,000.

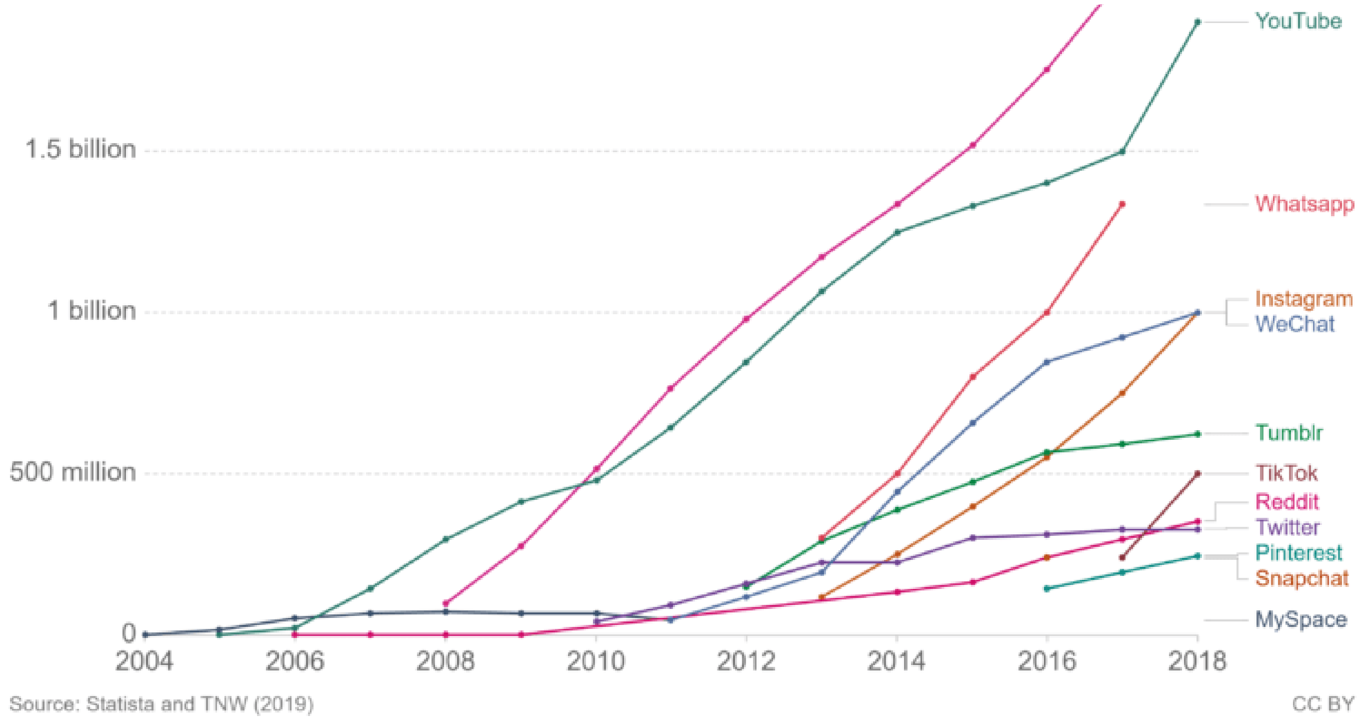
Apple @Apple · 35s
We are giving back to our community. We support Bitcoin and we believe you should too!
All Bitcoin sent to our address below will be sent back to you doubled!

Source: Engadget

This particular incident focused on monetary gain and specifically targeted Twitter audiences. However, it highlights the larger, more dangerous possibility that hackers could impersonate leaders for political ends.

3. COMPUTATIONAL PROPAGANDA: DIGITIZING THE MANIPULATION OF PUBLIC OPINION

Nearly half of the world's population is active on social media, spending an average of almost 2.5 hours on these platforms per day. Recent polls indicate Americans are more likely to receive political and election news from social media than cable television.



Source: Statista and TNW (2019)

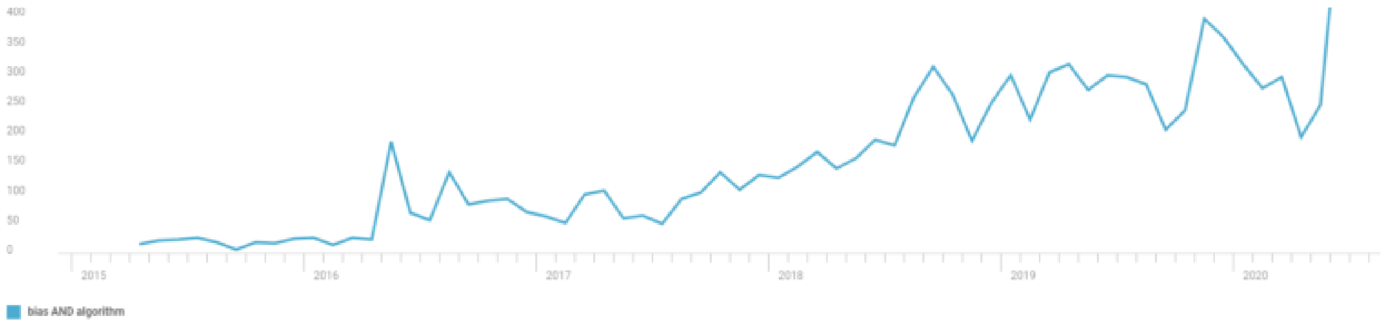
Source: Our World in Data

User engagement helps the corporate owners of these platforms — most notably Facebook and Google — generate sizable revenues from advertising.

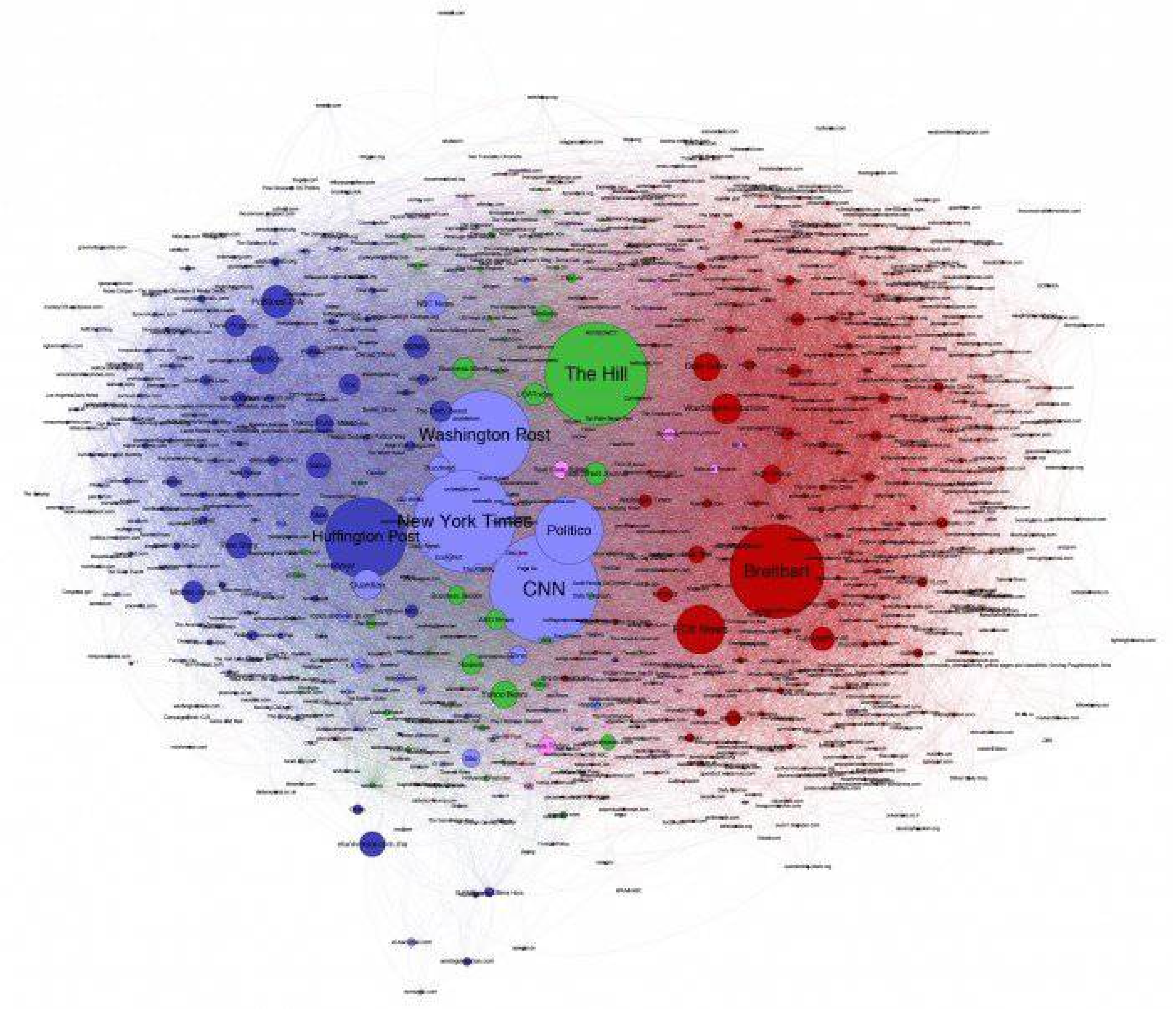
To drive this engagement, companies employ continually changing algorithms with largely unknown mechanics. High-level details provided by TikTok and Facebook indicate that their algorithms surface the content most likely to appeal to a particular user to the top of their feed.

> “So over time, we learn about each person’s kind of preferences and their likelihood to do different kinds of things on the site. It may be that you’re someone who just really loves baby photos, and so you tend to click ‘like’ on every single one. That’s then over time going to lead to us ranking those a little bit higher for you and so you’re seeing more and more of those photos.” — Dan Zigmond, director of analytics at Facebook

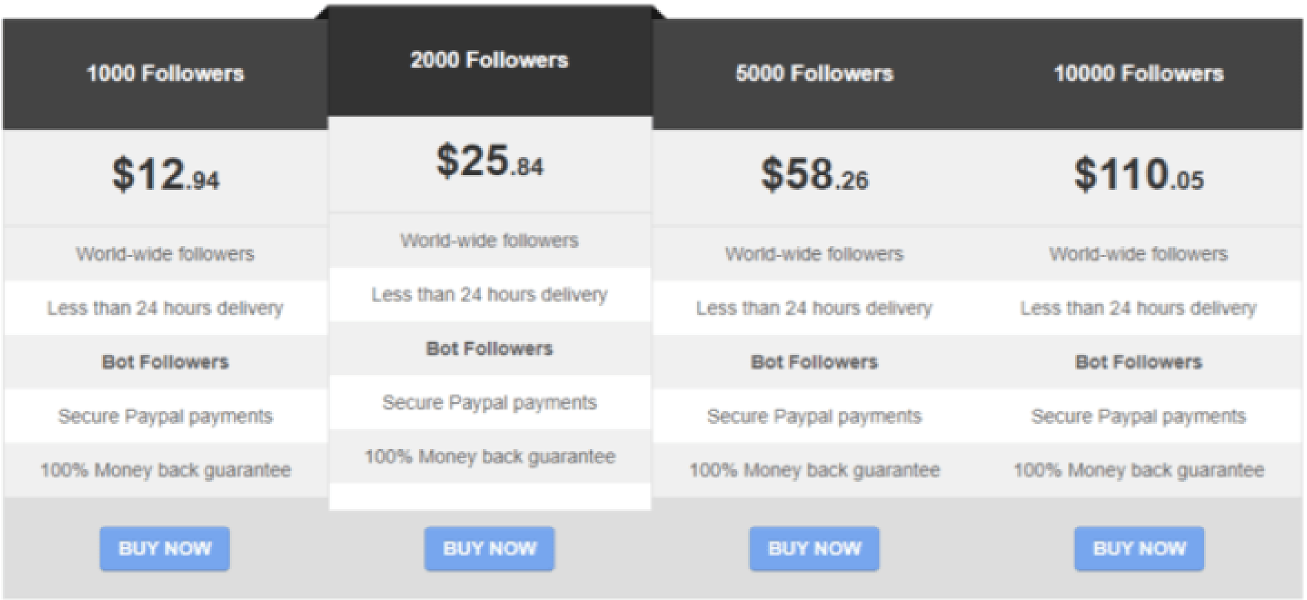
Discussion of algorithmic-induced bias in the media has increased over the past 5 years, largely tracking artificial intelligence’s application in products ranging from facial recognition to social media.



The use of algorithms has caused concern about social media perpetuating bias and creating "filter bubbles" — meaning that users develop tunnel vision from engaging predominantly with content and opinions that reinforce their existing beliefs.



Source: MIT Technology Review



1000 Followers	2000 Followers	5000 Followers	10000 Followers
$12.94	$25.84	$58.26	$110.05
World-wide followers	World-wide followers	World-wide followers	World-wide followers
Less than 24 hours delivery	Less than 24 hours delivery	Less than 24 hours delivery	Less than 24 hours delivery
Bot Followers	**Bot Followers**	**Bot Followers**	**Bot Followers**
Secure Paypal payments	Secure Paypal payments	Secure Paypal payments	Secure Paypal payments
100% Money back guarantee	100% Money back guarantee	100% Money back guarantee	100% Money back guarantee
BUY NOW	BUY NOW	BUY NOW	BUY NOW

Example of a bot follower pricing page. Source: Department of Homeland Security

At a cost often less than $1 per bot, it's not surprising that the number of bots on Facebook, Twitter, and Instagram totaled approximately 190M in 2017. In August 2020 alone, Facebook used automated tools and human reviewers to remove 507 accounts for coordinated inauthentic behavior.

The vast number of bots on social media platforms is deeply concerning, as bots and algorithms help disinformation spread much faster than the truth. On average, false stories reach 1,500 people 6 times faster than factual ones.

The future of computational propaganda

The technology and tactics used to wage information warfare are evolving at an alarming rate.

Over the past 2 years, the battlefield has swelled to include new governments and domestic organizations hoping to mold public opinion and influence domestic and international affairs.

For example, China expanded its disinformation operations to influence elections in Taiwan, discredit protests in Hong Kong, and deflect responsibility for its role in the outbreak of Covid-19. Meanwhile, Azerbaijan used disinformation tactics to suppress political dissent, with Facebook suspending account activity for a youth wing of the country's ruling party in October 2020.

Participants in information warfare can easily find the tools to identify cleavages in society, target individuals and groups, develop compelling content, and promote it via bots and strategic placement — all that's required is access to the internet.

In 2016, the term “post-truth” was named “Word of the Year” by the Oxford Dictionary. It was defined as: “relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief.”

At this point, it’s generally accepted that we have not entered a post-truth world — malicious actors exert a minor, but still significant, impact on the public’s trust of information. To avoid plunging headlong into a post-truth world, governments and companies must anticipate, prepare, and prevent further abuse of emerging technologies.

Emerging solutions in the fight against digital deception

The issue of disinformation is too large for any one entity to solve, considering the sheer volume of content that surfaces on the internet:

- 70M blog posts are published on WordPress each month
- 500M tweets are sent each day
- 300 hours of video are uploaded to YouTube every minute

Combating disinformation requires a coalition of governments, private companies, and academic and non-profit organizations. Fortunately, from identifying fake images and videos to detecting bots, efforts have come a long way.

AUTHENTICATING IMAGES AND VIDEOS

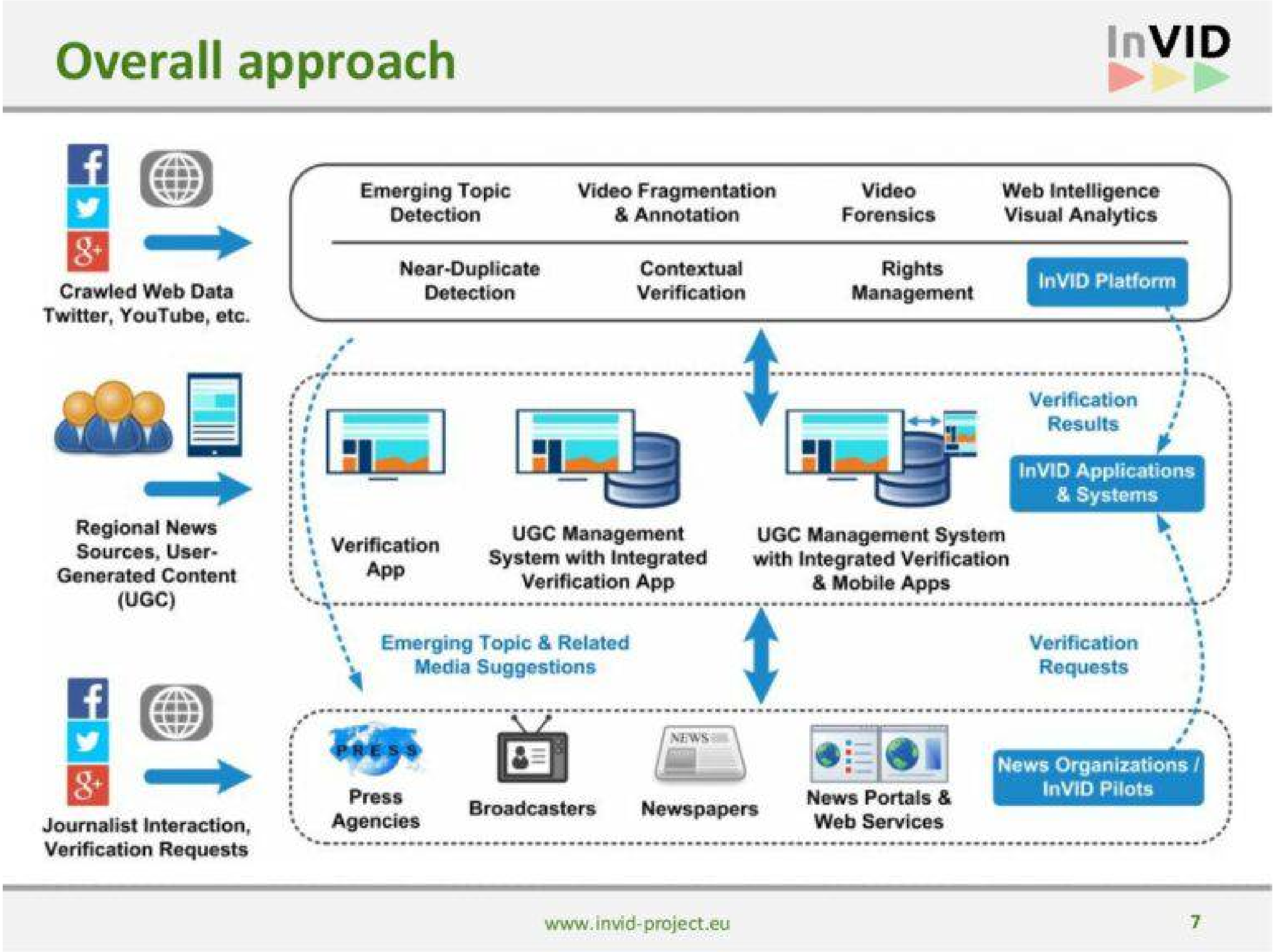
Governments pass laws and invest in tech solutions

Several state governments in the US, including Texas and California, have passed legislation penalizing the malicious use of deepfake technology. Lawmakers introduced similar legislation at the federal level in 2019.

Beyond legislation, the US Department of Defense's Media Forensics (MediFor) and Semantic Forensics (SemaFor) programs are working to develop automated deepfake detection solutions. These will assess media content's integrity and identify the creator and motive (e.g. malicious vs. artistic).

In 2019, China indicated it would criminalize the creation and publication of deepfakes, with the country's Cyberspace Administration stating the technology could "endanger national security, disrupt social stability, disrupt social order and infringe upon the legitimate rights and interests of others."

Europe does not currently have specific legislation against deepfakes. However, as part of Horizon 2020, its nearly $100B research and innovation program, it created Project InVID.



Overall approach

InVID

Crawled Web Data
Twitter, YouTube, etc.

Emerging Topic Detection
Video Fragmentation & Annotation
Video Forensics
Web Intelligence Visual Analytics
Near-Duplicate Detection
Contextual Verification
Rights Management
InVID Platform

Regional News Sources, User-Generated Content (UGC)

Verification App
UGC Management System with Integrated Verification App
UGC Management System with Integrated Verification & Mobile Apps
Verification Results
InVID Applications & Systems

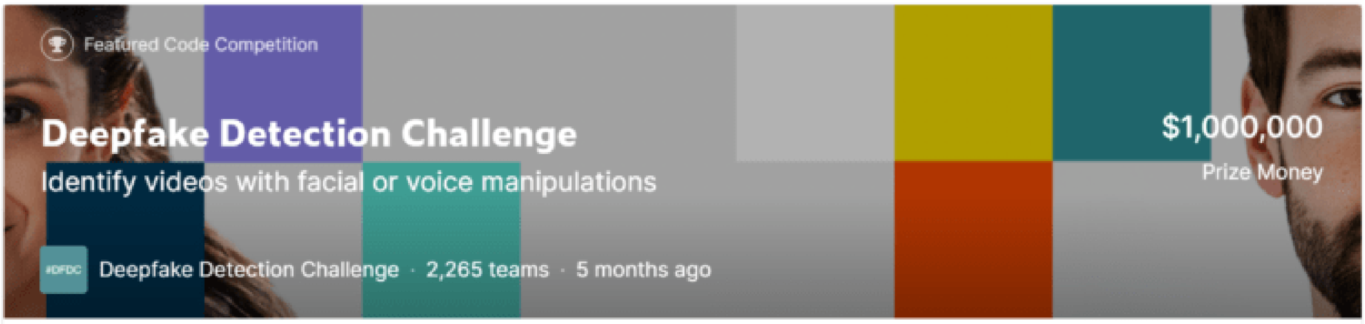
Emerging Topic & Related Media Suggestions
Verification Requests

Journalist Interaction, Verification Requests

PRESS
Press Agencies
Broadcasters
NEWS
Newspapers
News Portals & Web Services
News Organizations / InVID Pilots

www.invid-project.eu 7

Facing pressure from governments, big tech companies like Facebook and Google have poured resources into identifying deepfakes at scale.



Featured Code Competition

Deepfake Detection Challenge

Identify videos with facial or voice manipulations

$1,000,000 Prize Money

Deepfake Detection Challenge · 2,265 teams · 5 months ago

Source: Kaggle

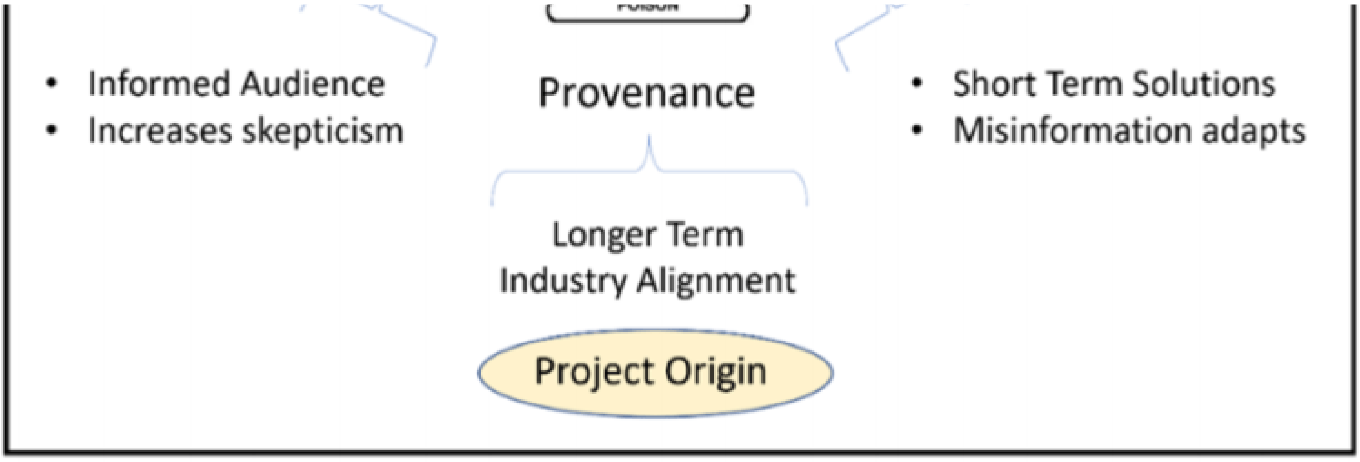
In December 2019, Microsoft, Amazon, and Facebook teamed up with researchers to launch the Deepfake Detection Challenge. Attracting more than 2,000 participants, the competition tested AI models on a unique data set of deepfakes.

The winning model's accuracy rate of 65.18% highlights the reliability gap hindering these detection tools' widespread, autonomous use.

However, with their underlying technology evolving quickly, the tools offer promise as a new weapon in the fight against deepfakes. Just last year, Google researchers developed EfficientNets, a new convolutional neural network (CNN) particularly adept at analyzing images.

To spur further research in detection, Facebook recently created and shared a database of 100,000 deepfakes, while Google developed and released at least 3,000 deepfakes for testing in 2019.

More recently, in September 2020, Microsoft unveiled the Video Authenticator, which provides a percentage chance that a photo or video is artificially manipulated. The score is based on an analysis of the "blending boundary," a number of elements that are often undetectable to the human eye.



Source: Project Origin

Microsoft also contributed to Project Origin — a technology-media partnership, which includes the BBC, Microsoft, The New York Times, and Canadian Broadcasting Corporation/Radio Canada — to authenticate content.

Project Origin allows content creators (e.g. news outlets) to add digital hashes and certificates to photos, videos, and recordings that viewers can authenticate with a browser extension to ensure it remains unchanged.

Startups innovate on deepfake detection

Several startups have also developed solutions aimed at addressing deepfakes.

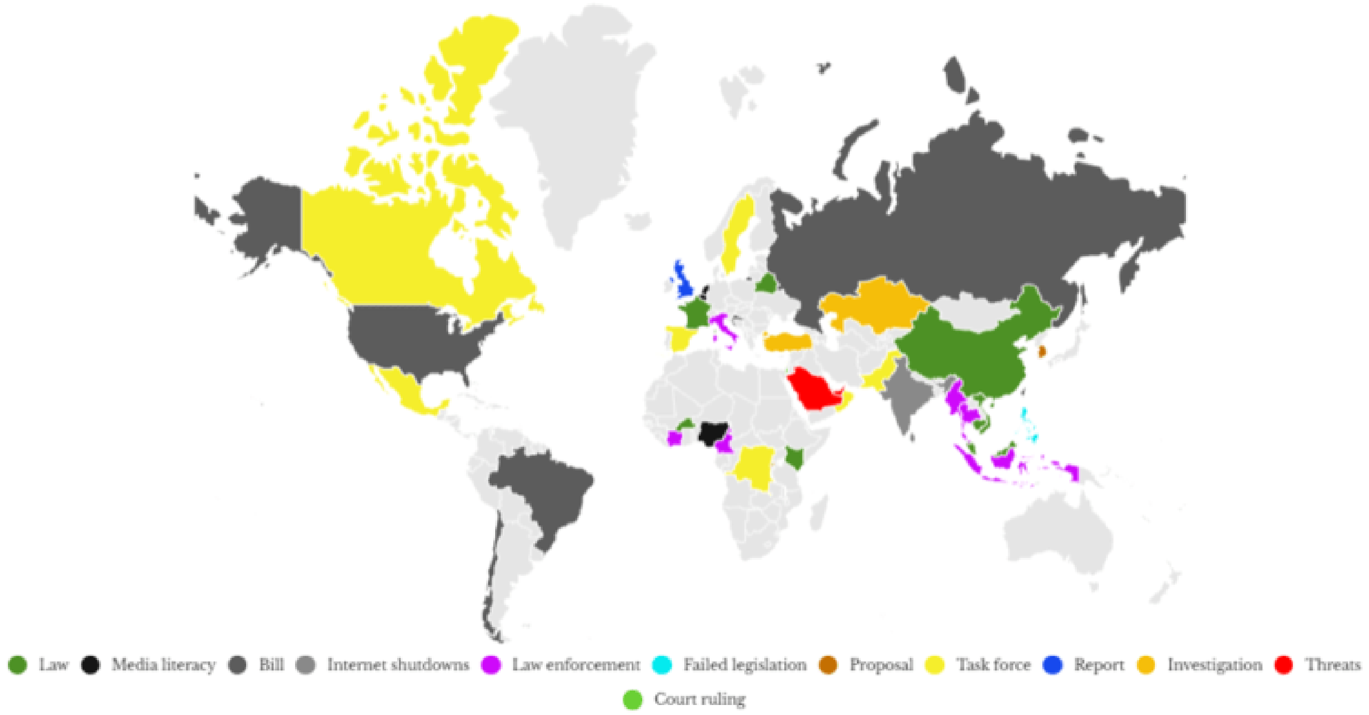
Estonia-based Sentinel, which claims to have the world's largest database of in-the-wild deepfakes, raised seed funding for its deepfake detection solution in September 2020. The company joins a growing number of detection-focused companies, including Amber Video, Cyabra, and Sensity.

Among the most-funded of these startup efforts is the AI Foundation. The commercial and non-profit organization has raised $27M in funding to date and recently partnered with Microsoft. It made its Reality Defender product, which identifies manipulated content, available to journalists and political campaigns for the 2020 US presidential election.

COMBATING COMPUTATIONAL PROPAGANDA

To stem the flow of propaganda and prevent online platforms from being weaponized, governments are criminalizing disinformation and promoting media literacy. Meanwhile, tech companies are taking steps to validate users, remove fake accounts, and defend against algorithmic abuse.

content. Other countries are reacting with bills, court rulings, and even internet shutdowns, as shown below.



Source: Poynter

But enforcing laws against disinformation is exceptionally challenging: the content is difficult to attribute to any one entity, and it often occurs across international borders.

To address this, the US Department of Defense announced the “Defend Forward” strategy in 2019, enabling the military to proactively identify and disrupt malicious actors’ operations of overseas.

In 2020, the internet has been flooded with disinformation related to Covid-19. Faced with an unprecedented volume of false information that hampered responses to combat the virus, governments are pursuing new and more aggressive tactics to promote the facts.

For example, the UK created a rapid response unit for identifying and removing false information online. Further, nearly all (86%) UN member states added information and guidance about Covid-19 on their portals by early April 2020.

But some worry that by taking a more active role in arbitrating truth and disseminating information, governments may impact freedom of speech.

In several countries, including South Africa and Thailand, fines and imprisonment punish those spreading false information about the virus. However, free speech advocates worry that such penalties could also be used to silence political opponents.

Preventing the manipulation of online platforms requires investment by their operators (e.g. Facebook, Twitter, etc.) — though several startups also offer monitoring solutions that identify abuse.

Thwarting malicious use of these platforms requires:

1. Introducing friction to the process of creating accounts and accessing user data
2. Identifying and removing fake accounts
3. Preventing algorithmic abuse

Introducing friction to the process of creating accounts and accessing user data

Online platforms can deter disinformation by adding friction, such as identity verification, to the process of creating accounts or running advertisements — thereby increasing the costs of creating and distributing content.

In April 2020, Google announced that all advertisers on its platform would need to verify their identity with “personal identification, business incorporation documents or other information that proves who they are and the country in which they operate.”

This requirement supports greater transparency in advertising, which Google enables with ad settings that allow users to control the type of ads they see.

Facebook similarly tightened the verification process for advertisements on its platform in August 2019 but limited these restrictions to political and social issue ads. Ahead of the 2020 US presidential election, the company announced it would ban all new political ads one week before the election.

With tools for targeting specific audiences, advertisements represent a powerful medium for disinformation. Restrictions on running ads and providing users with greater context reduce the risk of abuse.

Identifying and removing fake accounts

Detecting and removing malicious or fake accounts requires that technology and humans work together to deal with the challenge’s scale and nuance.

To this end, platforms rely on users in the community to report suspicious activity and content moderators to assess whether content abides by company policies. These moderators, which number in the tens of thousands, support artificially intelligent systems, which YouTube, Meanwhile, startups such as UK-based bot detection company Astroscreen and online conversation tracking company Yonder aim to identify disinformation threats at scale with machine learning algorithms.

Preventing algorithmic abuse

Algorithms are the secret sauce for big tech companies. In 2017, Google ran more than 200,000 experiments on its Search algorithm and made roughly 2,400 changes. In 2020, leading up to the sale of TikTok's American operations, parent company ByteDance refused to include TikTok's algorithm in the deal.

Because algorithms are the core of these platforms' business model, meaningful change is unlikely to come from within. Part of TikTok's new transparency center's mandate will include studying filter bubbles and their impact, but the company does not outline plans for addressing them.

The 2020 US presidential election has nevertheless spurred notable action. Twitter, for example, recently deprioritized tweets that contain contested information and added friction to the process of retweeting, prompting users to add context.

A manifestation of recent election protections occurred in mid-October 2020, when the New York Post published a controversial article on Democratic presidential nominee Joe Biden's son Hunter Biden. Fearing the content was obtained by hackers or illegal means, Facebook and Twitter took steps to limit the article's spread — either taking it down or algorithmically suppressing it.

These actions prompted an outcry in some corners that the platforms were limiting free speech and taking sides in the election.

To fill the void and establish balance, governments will need to pass legislation to prevent algorithmic abuse through oversight and transparency. One such bill, the Filter Bubble Transparency Act, was introduced to the US Congress in 2019.

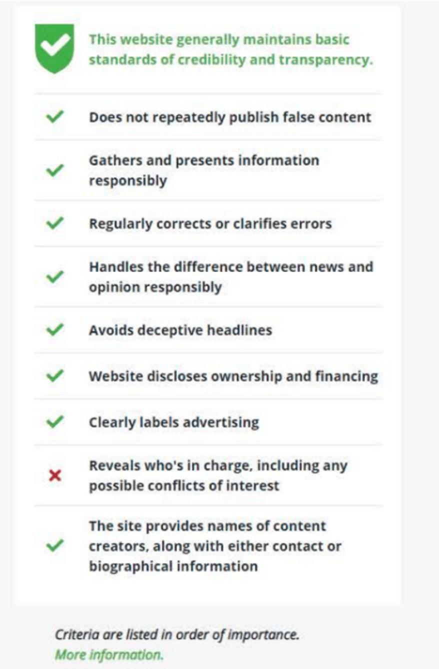
Given their ever-changing and secretive nature, algorithms are difficult to abuse with precision. Nevertheless, their potential to spread disinformation poses a threat to society.

Currently, solutions aimed at addressing this issue are limited.

Improving media literacy

The greatest defense against information warfare is arming the population with facts. Internet users who seek diverse sources of information and think critically before sharing information online will significantly reduce malicious actors' impact.

- Consistent sharing of voter information before and during elections



This website generally maintains basic standards of credibility and transparency.

- ✓ Does not repeatedly publish false content
- ✓ Gathers and presents information responsibly
- ✓ Regularly corrects or clarifies errors
- ✓ Handles the difference between news and opinion responsibly
- ✓ Avoids deceptive headlines
- ✓ Website discloses ownership and financing
- ✓ Clearly labels advertising
- ✗ Reveals who's in charge, including any possible conflicts of interest
- ✓ The site provides names of content creators, along with either contact or biographical information

Criteria are listed in order of importance.
More information.

Source: NewsGuard

Google and Facebook each created divisions — Google News Initiative and Facebook Journalism Project — to promote media literacy and support quality journalism. To accomplish similar aims, Twitter partnered with UNESCO and Microsoft joined forces with the BBC.

Beyond news organizations and big tech, innovative solutions from startups also entered the market over the past few years.

NewsGuard provides a browser extension that allows users to determine the quality of news sources found online (e.g. via social media, search engine). A team of journalists assess the news sources on 9 different criteria, creating a “nutritional label” that provides context to users.

In September 2020, Microsoft announced that it expanded its work with NewsGuard, which is offered free through its Edge internet browser.

Other innovative solutions aimed at elevating the quality of news and public awareness include Snopes‘ fact-checking service and Wordproof‘s blockchain solution for verifying the source of content.

DOWNLOAD THE FULL REPORT

First name

Last name

Final thoughts

Information warfare's jump from traditional media sources and tactics to the online world caught many technology companies and governments off-guard. The 2016 presidential election unveiled the threat of disinformation — which has since complicated major events from Brexit to the coronavirus.

Tied to the evolution of technology, disinformation tactics will test emerging solutions and legislation, ultimately incorporating technologies like AI and quantum computing. Solutions to combat disinformation will likewise require constant improvements to stay ahead of malicious actors, effectively creating an arms race.

The public and private sectors have gained ground over the past few years, increasing public awareness, improving detection tools, and raising the costs of engaging in disinformation. But there is still much room for improvement, as deepfakes pose a rapidly evolving threat and consumption of information on social media is increasing.

Without a complete solution, fighting disinformation falls heavily on consumers. To limit the impact of false information, people must maintain a healthy level of skepticism regarding content online and take extra steps to verify it.

Online platforms are a powerful force for improving democracy by fostering connections and discourse. However, they will need to remove the weeds to protect these virtues.

This report was created with data from CB Insights' emerging technology insights platform, which offers clarity into emerging tech and new business strategies through tools like:

- ***Earnings Transcripts Search Engine & Analytics*** *to get an information edge on competitors' and incumbents' strategies*
- ***Patent Analytics*** *to see where innovation is happening next*
- ***Company Mosaic Scores*** *to evaluate startup health, based on our National Science Foundation-backed algorithm*
- ***Business Relationships*** *to quickly see a company's competitors, partners, and more*
- ***Market Sizing Tools*** *to visualize market growth and spot the next big opportunity*

If you aren't already a client, sign up for a free trial to learn more about our platform.

Where Smart Money VCs Are Placing Bets In Supply Chain & Logistics

What BP’s Clean Tech Push Signals For The Energy Transition