aktuelle analysen | 77

Hanns Seidel Stiftung

Information Threats

Challenges for the European Information Space

Tabea Wilke

www.hss.de

Information Threats

Challenges for the European Information Space

IMPRESSUM

ISBN	978-3-88795-581-6
Publisher	Copyright 2020, Hanns-Seidel-Stiftung e.V. Lazarettstr. 33, 80636 Munich, Phone +49 (0)89 / 1258-0 E-Mail: info@hss.de, Online: www.hss.de
Chairman	Markus Ferber, MdEP
Secretary General	Oliver Jörg
Editorial Office	Barbara Fürbeth (Head of Office) Susanne Berke (Editor) Marion Steib (Design, Set, Layout)
Responsible person for the purpose of (German) press law	Thomas Reiner (Communication and Public Relations)
Cover Layout	Gundula Kalmer, Munich
Print	Hanns-Seidel-Stiftung e.V., Inhouse Print Office, Munich
Note	The views and opinions expressed in this text are those of the author and do not necessarily reflect the official policy or position of Hanns Seidel Foundation.

PREFACE

Markus Ferber, MdEP
Chairman of the
Hanns Seidel Foundation

The digital space increasingly shapes the private and work lives of young Europeans. The current Covid-19 pandemic accelerates the digital transformation. For Europe's younger generations it is hard to imagine a life without the world wide web, social media, or their smartphone.

The digital world gives mankind previously unknown access to information. Not only the access to information has radically changed, but also the opportunities to spread information. However, this still relatively new form of living with and in the digital space has not just beneficial consequences. With the rise of the internet as a medium for information and communication new dangers, threats, and challenges have developed and they are individual, collective, societal, and global.

For one thing, people with criminal energy use the digital space, mostly for personal gain. But it is not just "digital trickery"; the digitisation of our world is also being used to manipulate, to deceive, to damage, to scheme, to propagate or infiltrate. And this happens on different levels: it can happen on a private, personal level, on the level of civic or social groups, in companies, universities and other institutions, in political discourse as well as on state, interstate and international level.

Information plays a central role in all these challenges and threats: one can use, steal, manipulate, steer, and spread it. Information can be right, wrong, incomplete, incorrect, or inaccurate. It can serve as a weapon as well as a means of pressure or protection. This way information can become "disinformation" or so-called "Fake News".

Just as diverse as the intentions are the methods that can be used to ultimately influence or damage others in real life. The "ordinary citizen" in Germany and Europe is only slowly becoming aware of the manifold ways and kinds of threats that can lurk in the digital world.

Who actually knows what the term "hack and leak tactic" means? Where is the difference between a "silent" and a "cold" leak? What exactly are "social bots" and how do they work? In which way do they pose a threat? What is "narrative warfare" and how does it differ from "memetic propaganda"?

This White Paper seeks to present and explain the current most common threats in the digital information space. With this Aktuelle Analysen' edition we as Hanns Seidel Foundation want to contribute to a better understanding of the possible threats in the digital world and information sphere to deal with them more appropriately. That is why this publication, in both German and English, is aimed at all those in Europe who regularly enter the virtual space one way or another, retrieve information from it and possibly post, comment and spread it.

We wish you an interesting and informative read!

///

Contents

Tabea Wilke
is the founder and CEO of botswatch Technologies GmbH, Berlin. She is a member of the Association for Computing and Machinery "Special Interest Group Artificial Intelligence" (ACM SIGAI) and member of the Institute of Electrical and Electronics Engineers IEEE’s working group to develop a Standard for the Process of Identifying and Rating the Trustworthiness of News Sources. Wilke holds a Bachelor’s degree in Media and Communications and a Master’s degree in International Relations.

botswatch Technologies GmbH
Albrechtstr. 16, 10117 Berlin
www.botswatch.io

for
Hanns-Seidel-Stiftung e.V.

Tabea Wilke

Information Threats

Challenges for the European Information Space

Key Findings

- The identity of a society and the economic growth of liberal democracies are dependent on the authenticity, stability, and integrity of information, databases, and digital identities.

- The information space of liberal democracies is changing due to (1) rapid technological developments and (2) the erosion of people's trust in facts and scientific findings.

- Geopolitical conflicts are increasingly staged in the information space. Information warfare destabilizes information spaces around the world.

- Information threats can't be stopped by deleting accounts. Their architects will continuously search for ways to use the functionality and business models of relevant internet services for their own purposes. It is a daily competition between the attackers and the attacked, and the victor will be the side that has the best mastery of technology.

- The attribution of information threats is a substantial challenge. In the future, AI-enabled applications in speech and text processing, as well as in image processing, will remove the individual fingerprints of those that create information threats even as the threats are created. This will make reliable attribution even more difficult.

Recommendations

- The development of an understanding of the phenomena, risks, and dangers of threats to the information spaces of liberal democracies, free economic systems, and global political developments is one of the key competencies of policymakers in politics, society, and the economy.

- The development of appropriate measures to make people aware of threats in the information space.

- The implementation of a process for educating the target audiences of active information operations. An informed public is a resilient public. The more quickly the narrative, images, and goals of active information operations are known, the lower the odds that they will spread.

- Companies and organizational IT infrastructures that are secured according to industry standards, as well as multi-factor authentication for online accounts, contribute to the protection and authenticity of information, databases, and digital identities.

- Development of appropriate measures to enable people to recognize, process, and classify information from texts, images, videos, and feeds in a dynamic information space in the long term.

- Newsrooms need a shared code of ethics regarding covering information threats.

Background

Our information space is changing rapidly. People are connected globally, information is available worldwide and in real-time, the processing power of computers doubles every 3.5 months (Amodei & Hernandez, 2018), and smartphones offer the functions of powerful minicomputers. Our everyday lives are ruled by an ever-increasing amount of informational noise, in which it becomes more and more difficult to distinguish the relevant from the irrelevant and facts from falsehoods. The gray area in between is vast.

Even beyond technological developments, the way that people perceive information, process it, and react to it is changing. In the public discourse of liberal democracies, opinions and facts are becoming increasingly blurred, scientific findings are called into question, personal experience is given more weight than facts, and trust in established sources of information is dwindling (Mazarr, Bauer, Casey, Heintz, & Matthews, 2019). These societal phenomena are described with the terms "disruption of fact" (Lepore, 2016) and "truth decay" (Kavanagh & Rich, 2018). Today, credibility and trust have become among the most important currencies companies can possess.

While the information space of liberal democracies is changing, it is simultaneously becoming a place in which geopolitical conflicts and the battle for economic interests are staged. Terrorists stream their attacks on digital platforms in real-time (Stubbs, 2019). Individuals can use tweets to confuse and mislead security authorities (Backes, et al., 2016). International treaties are revoked as once-obvious alliances are called into question and new alliances form. Private actors are becoming a fixed component of international conflicts, which are increasingly carried out not with weapons, but with information (Lin & Kerr, 2019; Mazarr, Bauer, Casey, Heintz, & Matthews, 2019).

Information warfare is a type of war carried out without heavy weaponry or fallen soldiers. Technological developments and the global networking of humanity have provided information warfare with new tools. They can be seen in operations to influence the information space before elections and referendums, after natural disasters and acts of terrorism, in governmental crises, societal divisions, civil unrest and during protests and riots. The goal is to create doubts and mistrust in the minds of people, undermine faith in political order, stir national issues, weaken the identity of a society, generate false support, destabilize and confuse, drive apart existing alliances, and destroy the geopolitical and economic order of past decades.

The role of the authenticity of information

The previously discussed technological and social changes in the information space, and its increasing use as a place where warfare is conducted by means of information, are developments that we encounter every day.

In this white paper, information space is understood as the sum of all channels through which information is disseminated and can be provided to individual people or the public at large. This includes forms of media such as print, TV, radio, websites, and social media platforms, as well as blogs, apps, messenger services, emails, and the telephone (Mazarr, Bauer, Casey, Heintz, & Matthews, 2019). The focus of this white paper is on describing information threats on digital platforms, the social web, and internet services.

The information space is one of the most important systems of liberal democracies. Not only society, but also the economy and politics depend on a healthy information space in which information can be exchanged reliably between people and machines (Mazarr, Bauer, Casey, Heintz, & Matthews, 2019). The integrity of the information space is the basis for decisions made by people in their private lives, by people in companies, and by elected officials in politics.

They all rely on the stability, authenticity, and integrity of information, databases, and digital identities, which merge to create a mutually shared reality. It holds society and the global economy together. If the information space is manipulated, parallel realities are created that can endanger the stability and the growth of free societies and economic systems.

The scope of information threats

Information threats are strategies, instruments, and tactics that endanger the information space. They include disinformation, deepfakes, hack-and-leak tactics, social bots, account spoofing, and information operations.

Information threats exist on many platforms. Scientists, journalists, companies and the platforms themselves have proven and thoroughly documented operations on Facebook (DiResta, et al., 2018; Facebook, 2019; Facebook, 2018), Facebook groups (Facebook, Taking Down More Coordinated Inauthentic Behavior, 2018), Instagram (DiResta, et al., 2018; Facebook, 2019), Facebook Messenger (DiResta, et al., 2018), Twitter (DiResta, et al., 2018), YouTube (DiResta, et al., 2018), Wikipedia (Sharma & Scarr, 2019), Reddit (DiResta, et al., 2018), Soundcloud (DiResta, et al., 2018), Pokémon Go (DiResta, et al., 2018), Telegram (DiResta & Grossman, 2019), Gab.ai (DiResta, et al., 2018), Medium (DiResta, et al., 2018), VKontakte (DiResta, et al., 2018), Tumblr (DiResta, et al., 2018), Pinterest (DiResta, et al., 2018), Meetup (DiResta, et al., 2018), LiveJournal (DiResta, et al., 2018), Vine (DiResta, et al., 2018), Discord (Institute for Strategic Dialogue, 2019) and 4Chan (Institute for Strategic Dialogue, 2019). Operators of information threats select the platforms according to the current behavior of the target audience and the channel's opportunities and features in order to conduct the operation successfully. Therefore, the number of affected channels is constantly changing and may include additional platforms and applications in the future.

Almost every sector has already been a target of attacks. Targets include governments, parties, politicians, people in public life, journalists, activists, private citizens, network infrastructures, financial institutions, companies, NGOs, cities, schools, hospitals, airports, universities, sporting institutions, transnational organizations, and federations.

A challenge for the core values of liberal democracies

Threats in the information space are a daily competition between the attackers and the attacked, and the victor is the side that has the best mastery of technology. Internet companies can help by implementing appropriate information security measures for their platforms and users, which increases the effort and expense for attackers.

Attacks cannot be completely prevented. There are three reasons for this:

- Firstly, the architects of information threats are always looking for ways to use the functionality and business models of relevant platforms for their own purposes.
- Secondly, not only digital platforms and their applications, but also people's user behavior changes and develops every day. This opens up new opportunities for attackers.
- Thirdly, technology continues to develop, and this can create means of attack that were not previously technologically possible.

To effectively minimize threats in the information space without changing the shared values that underlie modern democracy and economic systems, is one of the greatest challenges of our age.

Even now, it is apparent that information threats are being used as an argument to limit the freedom of speech, as well as the access to the world wide web (Wakefield, 2019). For this reason, solid detection capabilities and the accurate attribution of harmful operations in the information space will become more and more important in the future.

Distinguishing between phenomenon and effect

This white paper will intentionally remain incomplete with regard to naming the threats in the information space. However, it will describe a number of strategies, tactics, and instruments that are currently relevant on digital platforms and which will continue to gain relevance in the future against the background of technological developments.

This white paper places great value on the distinction between the description of an existing phenomenon and the description of the effect of a phenomenon. This is important since a phenomenon may commonly occur, regardless of whether causal interdependencies between individual information threats to social or political changes have been identified and supported by scientific findings. This also – and particularly – applies to threats in the information space, which change daily.

This white paper describes various phenomena of information threats, their appearance, their use in various contexts, and their complex effects on the information space. For the question of the effect of information threats, we reference the research activities of Harvard University, Stanford University, Northeastern University, the University of Pennsylvania, the Oxford Internet Institute, and Princeton University, all of which have worked with this phenomenon in various scientific disciplines. Below, we will describe the threats, their importance, the various types of threats, and their actors using specific examples.

1. Information Operations

Information operations are military or news campaigns that seek to influence, control, confuse, deceive, change, or destroy the information space of a certain country or region (US-Army, 2003).

Information operations are carried out in times of war and armed conflicts, but also in times of peace (US-Army, 2003). They are part of psychological warfare and cognitive warfare. They are one of the strategies of hybrid warfare (Morris, et al., 2019) and generally stay below the threshold that would trigger a reaction from the adversary. As such, information operations are among the strategies in the military gray zone (gray zone conflicts) (Morris, et al., 2019). Meeting conflicts with information operations is called information warfare. Information war is a war without tanks and guns; it is a war with information.

Information operations are initiated and controlled by state actors. In the past 15 years, the execution of the operations has shifted to the private sector, meaning that non-state actors are also a component of hybrid warfare. The more complex and professional an information operation is, the more resources it requires.

Information operations make use of almost every channel that is used by the target audience in the respective information space. This includes platforms such as Facebook (DiResta, et al., 2018; Facebook, Taking Down More Coordinated Inauthentic Behavior, 2018; Facebook, Removing More Coordinated Inauthentic Behavior From Iran and Russia, 2019), Facebook Groups (Facebook, Taking Down More Coordinated Inauthentic Behavior, 2018), Facebook Messenger (DiResta, et al., 2018), Instagram (DiResta, et al., 2018), Twitter (DiResta, et al., 2018), Google Ad Sense (DiResta, et al., 2018), Gmail (DiResta, et al., 2018), YouTube (DiResta, et al., 2018), Wikipedia (Sharma & Scarr, 2019), Reddit (DiResta, et al., 2018), Soundcloud (DiResta, et al., 2018), Pokémon Go (DiResta, et al., 2018), Telegram (DiResta & Grossman, 2019), Gab.ai (DiResta, et al., 2018), Medium (DiResta, et al., 2018), VKontakte (DiResta, et al., 2018), Tumblr (DiResta, et al., 2018), Pinterest (DiResta, et al., 2018), Meetup (DiResta, et al., 2018), LiveJournal (DiResta, et al., 2018), Vine (DiResta, et al., 2018), Discord (Institute for Strategic Dialogue, 2019) and 4Chan (Institute for Strategic Dialogue, 2019). The actions of the information operations on digital platforms are complemented by state-backed alternative news sites (see Disinformation).

Information operations are not a new phenomenon. However, they have gained new opportunities for scale, speed, scope, and anonymity as the whole world has become connected through digital platforms (US-Army, 2003). Information operations are generally embedded in the larger concept of an influence operation (Lin & Kerr, 2019; US-Army, 2003).

Distinguishing between influence operations, astroturfing, and false flag operations

Information operations are targeted towards the information space of a country or a region. In contrast, influence operations use multiple tools to influence all aspects of a society through the economy, education, research, sports, the military, and diplomacy (US-Army, 2003). Information operations and influence operations are therefore distinguished by the spaces in which they operate.

Commercial PR campaigns by economic or political actors that, like information operations, seek to move through information space under disguise are called astroturfing. The common factor between information operations and astroturfing lies in the misleading intention of the operation and the professional execution of the campaign.

In past years, it has been increasingly common for some methods and tactics to be imitated by information operations. Campaigns coordinated by states that imitate an actor or method of a certain operation are called false flag operations. False flag operations are conducted to imitate another state actor and to simulate an activity that is not actually occurring. False flag operations that are conducted at a very high professional level are very difficult for the target public sphere and adversaries to identify.

Types of Information Operations

There are three different types of information operations: White, gray, and black (Lin & Kerr, 2019). The difference between the operations lies in the transparency of the information source and the client.

- **White information operations** are completely transparent with regard to the source and the client. The information space can clearly identify the author.

- **Gray information operations** disguise the origin of the information source and the client. They involve real third parties such as private citizens, foundations, NGOs, activists, and organizations as active actors to make the information seem authentic. Gray information operations are difficult for the civil information space to identify.

- **Black information operations** not only disguise the origin of the information source and the client, but are also first made visible by actors that come from the information space or appear to come from the information space. Black information operations are very difficult to identify and can only be exposed through forensic and intelligence-led capabilities. For the general public, it is hardly possible to connect an operation to its originator.

Narrative warfare and memetic warfare

Information operations appear in the digital space through (1) narratives and (2) viral images or short sequences of moving images (Graphics Interchange Format, GIFs), also called "memes". A society is connected by shared truths, shared narratives, and a consensus about its history. This forms the collective identity of a society. Information operations refer to this collective identity with the help of images and narratives in order to influence, shape, change, polarize or destroy it (US-Army, 2003). When this occurs using narratives, the tactic is called "narrative warfare" or "narrative propaganda". If it uses memes, the tactic is called "memetic warfare" or "memetic propaganda" (DiResta & Grossman, 2019).

Information operations use images and narratives for two purposes:

- Firstly, emotions such as fear, horror, disgust, surprise, dismay, schadenfreude, superiority or inferiority are evoked to create or stir societal discourse.

- Secondly, individual fringe groups of a society are connected through mutual images to create a new narrative.

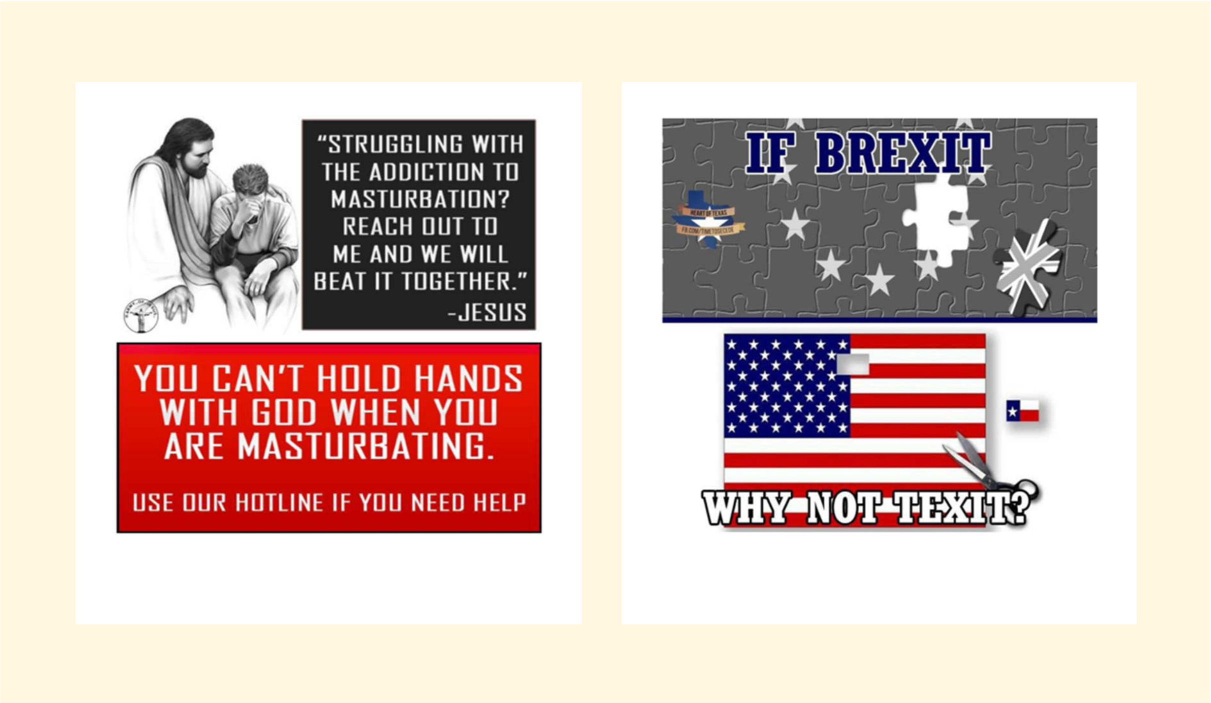
Beyond this, information operations use a variety of additional forms of information threats such as disinformation, hack-and-leak tactics, account spoofing, bots, deepfakes, shallow fakes, and many more.

Example of an Information Operation: DC Leaks

Influencing the US presidential election in 2016 was one of the most extensive and best documented information operations to date. The operation began in 2014 and lasted until the beginning of 2017. Some accounts remain active even today (DiResta, et al., 2018). The operation had three elements: (1) attacking and hacking voting systems, (2) hack-and-leaks of internal documents of the Democratic party (for an example, see "hack-and-leak tactics") and (3) extensive operations on digital platforms (DiResta, et al., 2018). All in all,

- approximately 10.4 million tweets were posted by more than 3,841 accounts,
- approximately 1,100 YouTube videos were posted by 17 accounts,
- approximately 116,000 Instagram posts were shared on 133 channels,
- approximately 61,500 Facebook posts were published on 81 Facebook pages (DiResta, et al., 2018).

Figure 1: "Army of Jesus" on Facebook and Instagram (left) and a visual that was posted in the Texit narrative (right, DiResta, et. al. 2018: 72). The figure at left received 5,436 likes and 284 comments in March and April 2017 (DiResta, et al., 2018: 40).


"STRUGGLING WITH THE ADDICTION TO MASTURBATION? REACH OUT TO ME AND WE WILL BEAT IT TOGETHER."
-JESUS
YOU CAN'T HOLD HANDS WITH GOD WHEN YOU ARE MASTURBATING.
USE OUR HOTLINE IF YOU NEED HELP
IF BREXIT
WHY NOT TEXIT?

On Instagram alone, the information operation achieved approximately 187 million engagements and on Facebook, approximately 77 million engagements (DiResta, et al., 2018). According to Facebook (DiResta, et al., 2018) the operation reached a total of approximately 126 million people. Its goals included the following (DiResta, et al., 2018):

- Demoralizing the black community and people of color in the US through extensive measures in approaching and influencing community leaders in churches, civil rights movements, the black media, self-defense courses, and protest movements with the intent of collecting sensitive private information, such as their sexual orientation or behaviors.
- Voter suppression. The goals of this campaign were (1) creating confusion about the electoral process and voting, (2) diluting votes by recommending people to vote for a third party, (3) demobilization of voters through calls to stay at home on voting day.
- Support for secession movements. In reference to Brexit, the information operation supported secession movements in the US, such as #Texit in Texas and #Calexit in California. They spread stereotypes and sensitivities against governments at the federal, state, and regional levels.

The challenge of attribution

Accrediting information operations to a certain actor (attribution) is one of the most significant challenges. In the future, it will even increase for two reasons:

- IP addresses, devices, technical services and operating systems can be easily spoofed or anonymized. This will make the solid detection and the accurate attribution of information operations more difficult.
- Individual language, individual grammatical errors, or styles in image processing will be more difficult to recognize in the future. As soon as highly developed AI-enabled translation and image processing are accessible for mobile devices, the individual characteristics that indicate the operator's individual digital fingerprint will be removed.

In addition, non-state actors such as activists, journalists, the private sector, and researchers are also imitating the methods and tactics of information operations for their purposes. This is another assault on the integrity of the information space and damages its authenticity.

The resilient public

The speed, agility, and rapidly changing nature of information operations pose significant challenges in countering them. One possibility is to inform the public immediately about active information operations and their narratives, images, and goals. Inorganic campaigns, images, narratives, and goals will become more obvious for the public. By informing the general public, the measures of the information operation would become ineffective.

In this context, reaction time plays an important role: Harmful and misleading narratives can be deployed within five minutes and amplified within 20 minutes. A subsequent correction of the narrative is hardly possible (Freedberg, 2019; Andrews, Fichet, Ding, Spiro, & Starbird, 2016). In the US, the Baltic States, Finland, Central Europe, and Sweden, these methods are already being used. This requires close collaboration between security authorities and experts in economics, science, and in NGOs to develop effective forensic capabilities for solid and accurate attribution. An informed public is a resilient public (US Director of National Intelligence DNI, 2019).

2. Deepfakes

A deepfake is video or audio material that looks real but which is created with the help of artificial intelligence. People do or say things that they never actually did or said. The word deepfake is a compound of the name of the technology with which deepfakes are produced (Deep Learning) and the goal of the change (fake).

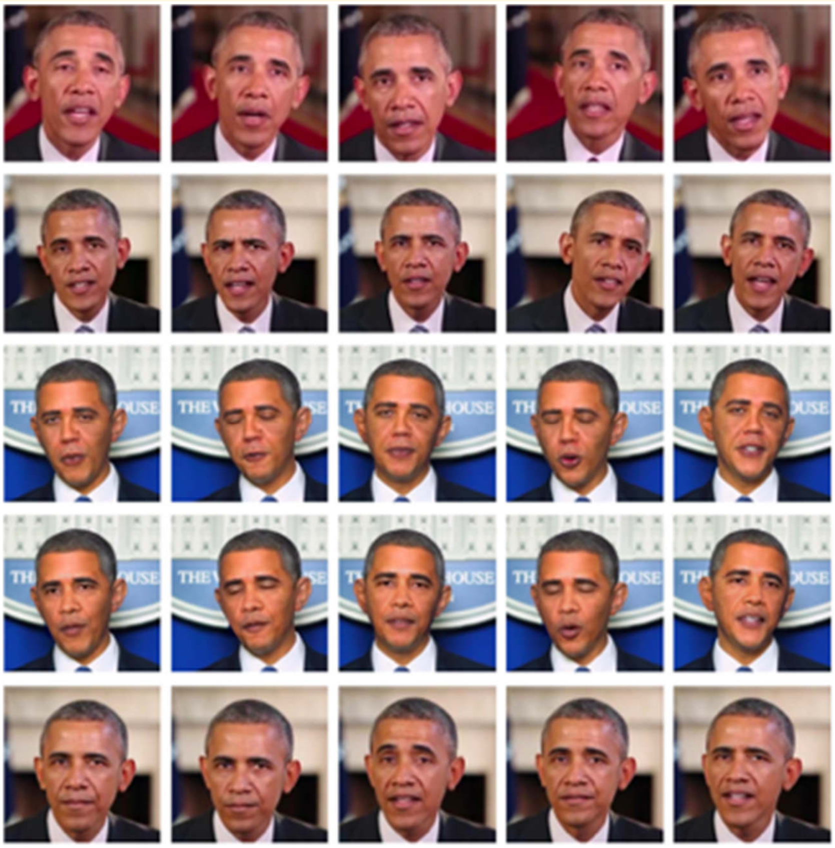
The underlying technology of deepfakes are deep learning models with generative adversarial networks (GAN). They have been used in developing text-to-speech models and improving the analysis of medical imaging data for years (Yi, Walia, & Babyn, 2019). High-quality deepfakes can hardly be distinguished from the original (Nelson & Lewis, 2019; Agarwal, et al., 2019).

Deepfakes may appear on almost all digital platforms through which audiovisual content is shared. This includes, for example, Instagram, Facebook, YouTube, Twitter, LinkedIn, Twitch, Vimeo, and Soundcloud.

Types of deepfakes

Currently, there are three different types of deepfakes: (1) face-swap, (2) lip-sync and (3) puppet master (Agarwal, et al., 2019). In a face-swap, the face in a video is automatically switched with another face (Harwell, 2018). In a lip-sync, the lip movements of a person are automatically adjusted to an audio frequency. A puppet master automatically changes all of a person's movements, such as head movements, facial expressions, and eye movements. In addition to these three types, there are countless variants, nuances, and new developments.

Figure 2: Five examples of a 10-second clip altered from the original (from top to bottom), lip-sync deep fake, comedic impersonator, face-swap deep fake and puppet master deep fake (Agarwal, et al., 2019)

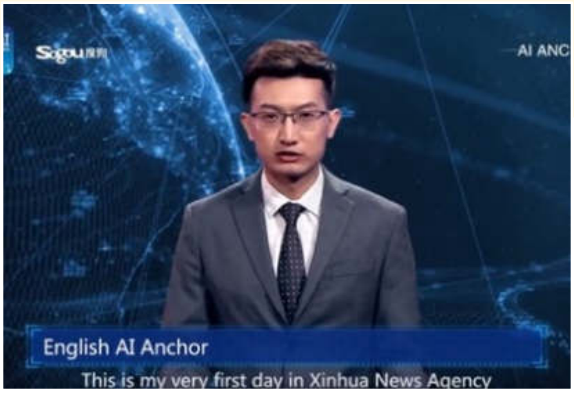


Commercial applications

In 2017, deepfakes became well-known in connection with pornographic content. The actor's faces were artificially switched for the faces of famous people. Commercial applications such as FaceApp from the Russian company Wireless Lab let the user's face age. The Chinese face-swapping app Zao integrates an upload into a popular blockbuster or streaming series like Game of Thrones. In its update in Fall 2019, the video platform Twitch integrated deepfake features into its livestream (Perez, 2019). In the summer of 2019, FaceApp won 12.7 million new users in only a few weeks (Sarwari, 2019). Zao quickly became one of the most popular apps in China (Ingram, 2019).

Equally relevant for the information space is the development of a deepfake news anchor for the Chinese state news agency Xinhua, which was introduced in November 2018 (Kuo, 2018). This deepfake is able to automatically read any kind of news, 365 days a year, at any time of day or night.

Figure 3: The first deepfake news anchor of the Chinese state broadcast station Xinhua



Deepfakes as a danger to the information space

- Rapid technological development. The technology that creates deepfakes is developing rapidly. New processes for high-end deepfakes appear on an almost weekly basis. The applications known today are only the beginning of a transformative technology.

- Potential impact. Deepfakes have a comparatively high potential for use in disinformation. They can create severe damage for individual people, political processes, or economies in a very short period of time (Nelson & Lewis, 2019).

- Access. Simple deepfake applications are available from many mobile apps. They can be created on a smartphone in only a few minutes – no programming skills required. Although created on a smartphone, this type of deepfake is sufficient to create confusion and draw attention in sensitive situations such as elections or terrorist attacks, and thereby has the ability to shape the outcome of major events.

Deepfakes have already influenced political processes in Malaysia. A possible deepfake of a man who claimed to have been intimate with the candidate for the office of prime minister was shared there. The video was disseminated quickly and led to confusion in Malaysian politics. Homosexuality is illegal in Malaysia. Another example is the fraud committed against a business enterprise with the help of a deepfake. In March 2019, employees of an energy company were deceived by an audio deepfake of their CEO and transferred payments totaling 220,000 euros to an external account (Stupp, 2019).

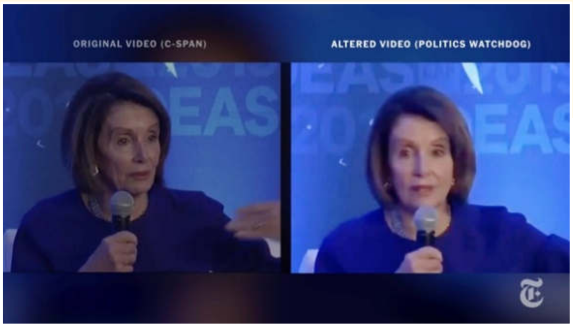
Differentiation from shallow fakes

Shallow fakes are also used in manipulating audiovisual content. Shallow fakes are not created with deep learning and therefore are not deepfakes. However, they are the result of an easy, generally minor manipulation of the material. This is the origin of the name "shallow", which can also mean "superficial". Despite the superficial manipulation of the material, a shallow fake can influence the course of political and economic processes.

Example of a shallow fake in politics

One such shallow fake made of the speaker of the US House of Representatives, Nancy Pelosi, was disseminated in May 2019 (Harwell, 2018). It showed a video of the politician at a public panel. By reducing the speed of the video, it created the impression that Nancy Pelosi was drunk or ill. The shallow fake spread quickly. On the "Politics WatchDog" Facebook page alone, the video was seen two million times in the first few hours, shared more than 45,000 times, commented on more than 23,000 times, and shared across platforms. Although it was soon clear that the video was manipulated, questions regarding the politician's health remained in the information space.

Figure 4: Original versus shallow fake of Nancy Pelosi (New York Times, 2019)



Challenges in detecting deepfakes

Detecting deepfakes is challenging. As the technology continues to develop, it becomes more and more difficult. A few weeks prior to the publication of this white paper, deepfakes could be detected by using image compression to identify hard edges, image errors, and shadows next to the person, unnatural blinking, or mouth movements. These phenomena have since been removed, meaning none of the high-quality deepfakes now have any of these errors. Researchers from the University of Berkeley expect that they will be able to automatically identify deepfakes in the future through the combination of facial expressions and head movements (Agarwal, et al., 2019). In coming years, the importance of deepfakes and shallow fakes as a threat to the information space and democratic and economic processes will continue to increase on pace with the development of the technology and commercial availability.

3. Hack-and-leak tactics

Hack-and-leak tactics disclose sensitive information (leaks) from an attack on a computer system or network (hacks) to create, shape, or stir public issues. The leak can occur immediately after the hack or at a later time. If the leak occurs at a later time, the disclosure corresponds to beneficial moments for the leak in politics, the economy, or society.

Types of hack-and-leak tactics

There are four different types of hack-and-leak tactics:

- **Hot leak.** A break into a computer system or network with access to sensitive data (hack) and the disclosure of the data either directly or by a third party (leak).

- **Silent leak.** A break into a computer system or network with access to sensitive data (hack) with no disclosure of the data (no leak).

- **Fake leak.** A break into a computer system or network with access to sensitive data (hack) and the spread of intentionally false or fabricated data (fake leak).

- **Cold leak.** A break into a computer system or network without access to sensitive data (no hack) and the spread of intentionally false or fabricated data (fake leak).

The methods of hack-and-leak tactics

For hack-and-leak tactics, the distribution of the data to the media is a decisive moment. As soon as information from hacks is published, public discourse generally focuses on the people, organizations, and content in the leaks. Very rarely, the way how journalists accessed the information or the credibility of sources is discussed. This is a weak point in the media coverage of leaks and is exploited by hack-and-leak playbooks.

Hack-and-leak tactics can also use the psychological effect of the hack. A hack can destabilize the person or organization that was attacked and lead them to take imprudent actions. Sometimes, these reactions have more significant effects and toxic outcome than the hack itself. At the same time, the full attention of the person under attack is focused on investigating the hack and limiting the alleged damage. In this time, the attacker can run additional operations which go almost unnoticed for the target.

Whether leaked data is legit or fabricated is of secondary importance to the success of hack-and-leak tactics. Any confusion or doubt created about a political leader, a political process such as an election, a presidential candidate, a mayor, a party, or the senior executive managers of a company has the potential to remain in the information space.

Hacks as a service: Hack-for-hire

An attacker gaining access to an email account poses the risk of compromising all the other services and connected contacts tied to that account as well. On the black market, the hacking of an email account is offered as a service. Currently, prices range from 100 to 400 euros (Mirian, 2019). However, these hack-for-hire services do not include leak campaigns.

A sophisticated and effective hack-and-leak operation is planned over months or sometimes years and is demanding in their operational execution. This makes them expensive and limits the originator mostly to state or state-backed actors.

Example of hack-and-leak tactics: DC Leaks

One of the most famous state-backed hack-and-leak operations was the DC leaks during the US presidential election campaign in 2016. The architecture of this operation included registering domains, email accounts with Microsoft and Gmail, and profiles and websites on the social web (US Department of Justice, 2019). Accounts on Facebook ("DCLeaks") and Twitter ("@dcleaks_") were used both to start the campaign and to contact journalists personally using direct messages.

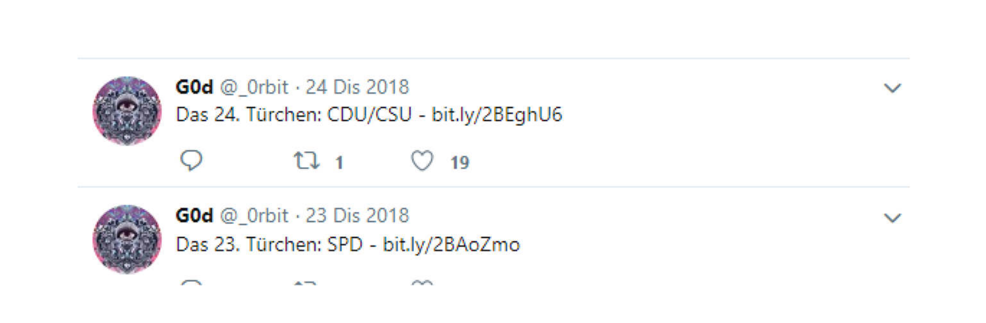
The active part of the campaign began with the registration of the domain (dcleaks.com) in April 2016, five months before the US presidential election. The domain remained active until March 2017. The website was used to spread thousands of documents obtained by hacking the Democratic party (Democratic Congressional Campaign Committee, DCCC and the Democratic National Committee, DNC), the Clinton campaign employees and volunteers, "including campaign chairman John Podesta, junior volunteers assigned to the Clinton Campaign's advance team, informal Clinton Campaign advisors, and a DNC employee" (US Department of Justice, 2019). The released material included personal identifying and financial information, internal correspondence related to the Clinton Campaign and prior political jobs, and fundraising files and information (US Department of Justice, 2019). Parts of the website were protected by a password to control the access to the documents by journalists and third parties (US Department of Justice, 2019).

The leak largely dominated media coverage for the last months before the presidential election. It also created the opportunity to spread fabricated information and conspiracy theories about the targeted people and Hillary Clinton ("Pizzagate"). This was not only used by the operators of the campaign, but also by commercial actors in other parts of the world who were substantially rewarded with advertising revenue through publishing false stories on their blogs and websites (Kirby, 2016).

Differentiation from doxing

Hack-and-leak tactics are different from doxing. The word "doxing" comes from the abbreviation of "documents" as "docs". In contrast to hack-and-leak tactics, doxing involves gathering sensitive information from websites and social media channels or with the help of social engineering tactics.

Figure 5: Tweets from the advent calendar doxing campaign of 2018



G0d @_0rbit · 24 Dis 2018
Das 24. Türchen: CDU/CSU - bit.ly/2BEghU6

G0d @_0rbit · 23 Dis 2018
Das 23. Türchen: SPD - bit.ly/2BAoZmo

The most famous case of doxing in Germany is the advent calendar leak in December 2018. In this case, both publicly available information, such as addresses and telephone numbers, as well as information that was hacked or acquired through social engineering, such as banking information and private chat protocols, was published (Eddy, 2019). The approximately 1,000 people affected included the Federal Chancellor, members of the Bundestag from almost every party, journalists, YouTubers, musicians, actors, and other people from public life. The data that were gathered were published on various accounts on Twitter step by step as an advent calendar, which initially remained unnoticed by German security authorities.

Challenges presented by hack-and-leak tactics

Both for hack-and-leak tactics and for doxing, there are countless variants, areas of overlap with other methods, and newly developed tactics. Hack-and-leak campaigns cannot be avoided. However, it is possible to increase the time and expense required on the part of the attacker. This is primarily achieved through IT infrastructure that corresponds to industry standards and by securing online accounts with multi-factor authentication.

4. Account Spoofing

Spoofing means "feigning" or "disguising" and is a method of capturing an existing digital identity for a certain period (hack) or imitating it with a similar-appearing identity (spoofing). Accounts are spoofed in order to use the stolen or imitated digital identity to spread false information, establish contact with connected people of this account, or convince them to do certain things. The goal could be transferring money, clicking on a link or on an attached file to download malware, or giving out a password.

Accounts can be spoofed on almost any platform. These attacks have the potential to create a high degree of global confusion at low cost with little skills or effort. However, a campaign with coordinated spoofed accounts on multiple platforms requires resources, advanced expertise in operation security, and professional planning and execution. Even though not every case of account spoofing has a malicious intention behind it, it has the potential to create confusion and mistrust in the integrity of the information space.

Types and methods of account spoofing

There are many different types of spoofing, such as email spoofing, text message spoofing, or IP spoofing. In the context of threats in the information space, account spoofing on digital platforms are most relevant. Currently, there are two common methods:

- Spoofing an individual's account to spread disinformation, spam, rumors or satire about the person or the institution to which the person belongs. In this method, a person's account is often hacked. The attacker then gains full access to the profile and to connected profiles and contacts.

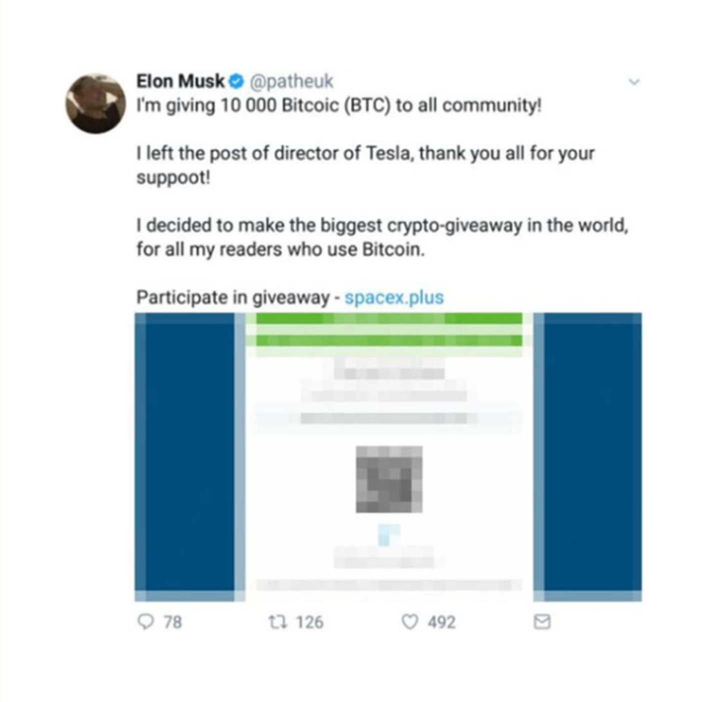
- Spoofing an organization's account, such as one belonging to journalists, a governmental authority, a news agency, or a company, in order to spread false information in sensitive situations such as terrorist attacks, civil unrest, natural disasters, riots or armed conflicts. In this method, a third-party account is used to mimic the targeted account.

In sensitive situations of public safety and security, account spoofing is especially harmful, since many people have severely limited awareness in such situations. They then overlook signals that the news report, image, or video comes from a fake account. The information is seen as credible and may be further distributed with retweets or shares. As soon as media outlets pick up this information, the attackers gain even greater coverage for their campaign. This is particularly a danger for journalists, authorities, politicians, and the communications departments of companies and organizations, all of which often feel under pressure to react promptly in such situations.

Example of account spoofing with Elon Musk's identity

An example of spoofing an individual account can be found in the scam campaign on Twitter in November 2018, which used the digital identity of the entrepreneur Elon Musk (Gerken, 2018). In this case, multiple accounts that were officially verified by Twitter were hacked and the profile names were changed to "Elon Musk". The spoofed accounts sent out spam tweets with a link to a website that would allegedly give out ten bitcoins for every one that was donated. Other hacked accounts replied to the tweet and thanked them for the bitcoins, which was intended to establish credibility. Indeed, the tweets were written like obvious scams ("Bitcoic" instead of "Bitcoin", "suppoot" instead of "support") and the accounts continued to have their specific user names on Twitter (Twitter handle). Despite that, the tweet looked like a tweet from Elon Musk to many people at first glance.

Figure 6: Spoofed account from the scam campaign that imitated the identity of Elon Musk



Elon Musk @patheuk

I'm giving 10 000 Bitcoic (BTC) to all community!

I left the post of director of Tesla, thank you all for your suppoot!

I decided to make the biggest crypto-giveaway in the world, for all my readers who use Bitcoin.

Participate in giveaway - spacex.plus

The challenge of protecting digital profiles

From a technical perspective, attackers will always find a way to get around account security and verification measures on digital platforms and to use details such as images and names of digital identities for their own purposes. Despite this, the role of multi-factor authentication of digital accounts is the first step to increase the cost for attackers (Mirian, 2019).

5. Bots

Bots are accounts on social media networks that are not controlled by people, but rather run automated by software. The name comes from an abbreviation of robot ("bot"). Bots interact with other accounts and are able to imitate human behavior (Ferrara, Varol, Davis, Menczer, & Flammini, 2016). Sophisticated programmed bots are difficult to detect.

Today, automation processes are a fundamental part of almost every digital service. Bots use automation that gives them the ability to control not just one account, but hundreds, thousands, or tens of thousands of accounts simultaneously. No humans are needed to control a bot account. The software's programming determines what activity the bot carries out at what time.

The most important platforms on which harmful bots are currently used are Twitter (Ferrara, Varol, Davis, Menczer, & Flammini, 2016), Facebook (Ferrara, Varol, Davis, Menczer, & Flammini, 2016) and Instagram (Maheshwari, 2018). According to Twitter, 8.5 % of accounts on the platform were automated in 2014 (Twitter Inc., 2014).

Differentiation from chatbots and comment bots

Bots should be distinguished from chatbots or comment bots. Chatbots allow conversations in an app or on a website to be automated. Although part of the name is the same, automation is all that connects the two. Chatbots are not sole and established accounts on social media networks. Comment bots post automated comments on products, photos, videos, or livestreams. Like chatbots, comment bots are not sole and established accounts on social media networks.

The manipulation of the information space at scale

Bots are used in the service sector to automatically answer customer questions, automatically post content such as tweets, images, or videos at a certain time, or to automatically favorite, like, or retweet certain accounts or words (Ferrara, Varol, Davis, Menczer, & Flammini, 2016).

However, in past years, bots were commonly used to manipulate digital platforms to distort the social or political reality (Howard, 2016; Ferrara, Varol, Davis, Menczer, & Flammini, 2016), to artificially boost the reach and amplification of tweets and accounts (Andrews, Fichet, Ding, Spiro, & Starbird, 2016), to scale campaigns meant to damage companies' reputations (Andrews, Fichet, Ding, Spiro, & Starbird, 2016), to influence elections (Howard & Kollanyi, 2016) and to diminish the impact of hashtags used by political activists with the help of spam (Finley, 2015).

In the field of disinformation, bots are used to flood digital platforms with misleading narratives at scale (Shao, et al., 2018). For this purpose, they share content at a high frequency and contact credible accounts on the platform deliberately and directly (Shao, et al., 2018) or use favorites and retweets to support real people who distribute their narrative. This increases the likelihood that the misleading narrative will be seen by credible accounts, accepted, and further distributed in their networks (Howard, 2018; Lazer, et al., 2017) and that uninformed journalists will include this narrative in their reporting and spread it even farther. Because they require little effort or expense, bots are a common instrument of disinformation, information operations, and hybrid warfare (see "Information Operations").

People use bots to create artificial majorities – whether for human rights and democracy or to create division in a society. The long and resource-intensive process of developing an organic sphere of influence and community is intentionally circumvented.

Types of bots

Researchers differentiate between two types of bots: bots and hybrids, also called cyborgs (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019). While bots are controlled completely automatically, hybrids are still controlled by people, either partially or for a certain period of time.

The characteristics of bots change constantly. That’s why there is no set definition of when an account is a bot. The Oxford Internet Institute defines it as highly frequent accounts that post more than 50 tweets a day (Howard, 2016). The disadvantage to this definition is that, for example, accounts held by news agencies or journalists that publish many tweets a day or that work with an automation software may be falsely categorized as bots. Although other researchers use different criteria for the definition for this reason, the Oxford Internet Institute's definition is a helpful approach in identifying automated accounts (Rinehart, 2017). The objectives pursued by bots cannot be determined solely by their automation function.

Effects of bots on the information space

- Artificially created majorities: People, companies, the media, and politicians think that this topic is important to many people in the country.

- Artificially created opinions: People, companies, the media, and politicians think that a certain opinion is now part of societal discourse or is the prevailing opinion.

- Polarization and increased social division of society: People, companies, the media, and politicians think that concepts of societal cohabitation are growing farther apart or that certain societal values and norms change.

Examples of the use of bots: Artificial majorities and damaging the reputation of businesses

In the French presidential election of 2017, bots were used to release and amplify leaks about the candidate Emmanuel Macron with the intent of influencing the outcome a few hours before the election (Volz, 2017).

In the US presidential election of 2016, bots boosted specific candidates. During debates, the percentage of bots was between 23 % and 27 % on Twitter that referred to the US 2016 elections. In the last week before the US presidential election in 2016, 18 % of tweets about the election were from bots (Kollanyi, Howard, & Woolley, 2016). Before the federal election in 2017 in Germany, the proportion of bot tweets on topics related to the election – in the same period and using the same criteria such as Kollanyi, Howard, & Woolley, 2016 – was almost 23 % (botswatch Technologies, 2017).

Figure 7: Activity of automated accounts during the week of the 2016 US presidential election (Kollanyi, Howard, & Woolley, 2016)

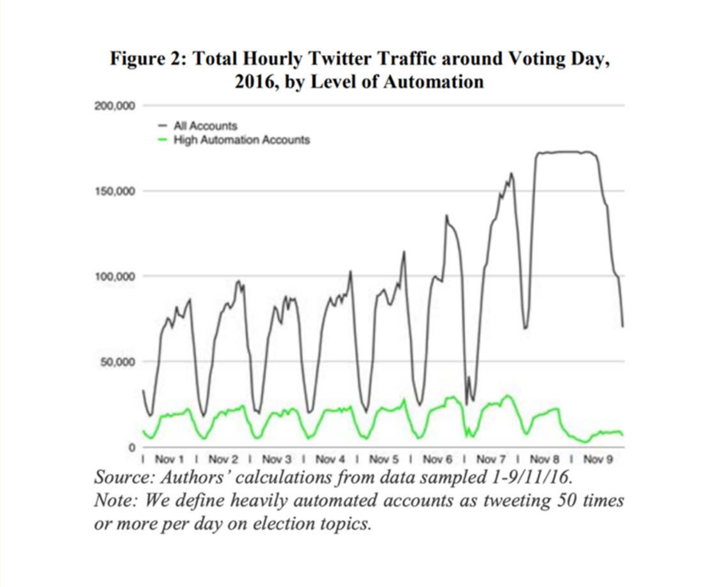
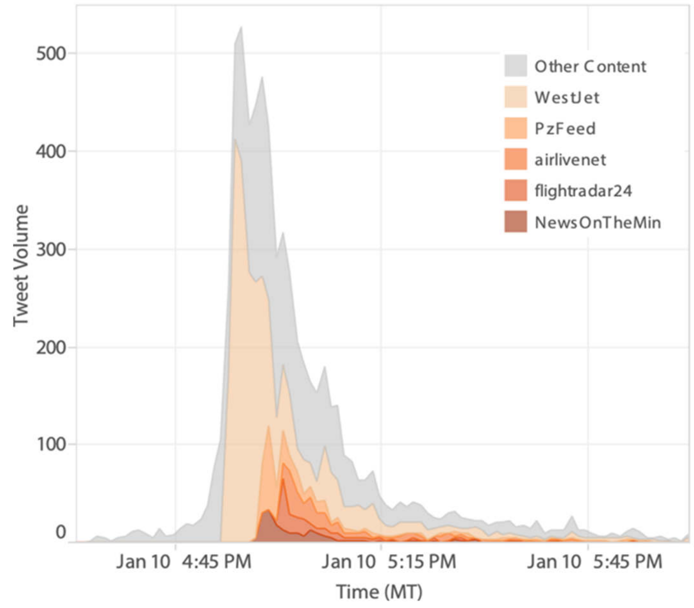


Figure 2: Total Hourly Twitter Traffic around Voting Day, 2016, by Level of Automation

Source: Authors' calculations from data sampled 1-9/11/16.
Note: We define heavily automated accounts as tweeting 50 times or more per day on election topics.

In 2015, bots amplified a rumor that a WestJet airplane sent an emergency signal on the way from Canada to Mexico (Andrews, Fichet, Ding, Spiro, & Starbird, 2016). The rumor was picked up by a flight-tracking website. Within 20 minutes, it was being disseminated through Twitter at a high frequency by bots. Due to the speed of the communication, the company was hardly able to refute the rumor quickly enough.

Figure 8: Tweet volumes of denials over the course of time (Andrews, Fichet, Ding, Spiro, & Starbird, 2016)



Bots as a service

Bots have the ability to influence the information space with little effort or expense. Developing or deploying a bot does not require deep programming skills. The service can be purchased inexpensively.

The future of bots

In coming years, the significance of bots as a threat to the information space will increase as AI-enabled technologies like natural language processing (NLP) become inexpensive and more accessible on mobile devices. With these technologies, it will be possible to customize and further adjust and synchronize bots to their specific target group and even to individual people.

6. Disinformation

Disinformation is the intentional planning, creation, and distribution of false, misleading, fabricated or deceptive information (Wardle, 2017). Its goal is to weaponize information in order to shape public opinion, destabilize and confuse, create doubts in the minds of people, undermine faith in trusted institutions, or to exploit societal divisions. This is particularly effective if official agencies remain silent in sensitive situations of public safety and security such as terrorist attacks, mass shootings, natural disasters, civil unrest, or riots (Runow, 2017). The tools of disinformation include bots, account spoofing, hack-and-leak tactics, and deepfakes.

Disinformation is not a new phenomenon. The manipulation of the information space is one part of psychological warfare since the beginning of the 21st century. Globally connected and digitalized societies, transnational public spheres, and smartphones with advanced capabilities in mobile image and audio processing have increased the speed at which content can be created and spread. The cost and barriers to disinformation have significantly decreased over the past few years.

Actors of disinformation are partisan citizens, activists, political parties, small and large organizations, commercial service providers, and governmental institutions. However, sophisticated disinformation campaigns require a cross-functional experienced team and accurate planning, technical equipment, and money. For this reason, advanced disinformation campaigns are often times initiated, backed, or financed by state actors.

Disinformation can appear on almost any digital platform. The channels currently being used include Facebook (DiResta, et al., 2018), Instagram (DiResta, et al., 2018), Facebook Messenger (DiResta, et al., 2018), Twitter (DiResta, et al., 2018), YouTube (DiResta, et al., 2018), Wikipedia (Sharma & Scarr, 2019), Reddit (DiResta, et al., 2018), Soundcloud (DiResta, et al., 2018), Pokémon Go (DiResta, et al., 2018), Telegram (DiResta, et al., 2018), Gab.ai (DiResta, et al., 2018), Medium (DiResta, et al., 2018), VKontakte (DiResta, et al., 2018), Tumblr (DiResta, et al., 2018), Pinterest (DiResta, et al., 2018), Meetup (DiResta, et al., 2018), LiveJournal (DiResta, et al., 2018), Vine (DiResta, et al., 2018), Discord (Institute for Strategic Dialogue, 2019) and 4Chan (Institute for Strategic Dialogue, 2019).

The selection of a platform is determined by the current behavior of the target audience and the technical opportunities offered by the platform for carrying out such an operation. The specific usage of channels is therefore constantly changing and may include additional platforms in the future.

Differentiation from misinformation

While disinformation always requires intentional planning and actions for distribution, misinformation is the unintentional spread of false information. Reasons for misinformation include poor journalistic skills (poor journalism), the intent to provoke (provoke or punk), or strong personal conviction in a specific matter (partisanship) (Wardle & Darakshan, 2017).

The phenomena and causes of disinformation and misinformation are often combined under the term fake news. In understanding the phenomenon and developing strategies for solutions, it is helpful to avoid the term "fake news" and furthermore differentiate between disinformation and misinformation.

Seven types of disinformation (Wardle & Darakshan, 2017)

- Satire or parody.
- Misleading content: Embedding information in a misleading way to put a topic or a person in a misleading context (framing).
- Impostor information: Authentic sources are imitated.
- Fabricated content: Manufactured and false information.
- False connection: The title of a post or article does not correspond to the content.
- False context: True information is placed in a false timeline or context.
- Manipulated information: The misleading manipulation of authentic information or images (Wardle & Darakshan, 2017).

State and alternative media as instruments of disinformation

Disinformation campaigns on digital platforms are often times complemented by alternative news websites. These include news blogs and alternative news websites with ideological, polarizing, highly partisan or extreme viewpoints (Newman, Fletcher, Kalogeropoulos, & Nielsen, 2019). Alternative news websites of state-backed disinformation campaigns not only approach the target public but also the citizens living overseas in the targeted country (diaspora).

Most alternative news websites mimic the look and feel of pages from credible media outlets. On their page, they emphasize disclosing the truth about society, politics, or companies and thereby lift themselves above the media, which they call the "lying media" (Lügenpresse) or "fake news media". Alternative news websites that distribute disinformation are mostly targeted towards a very specific audience in a very distinct regional area.

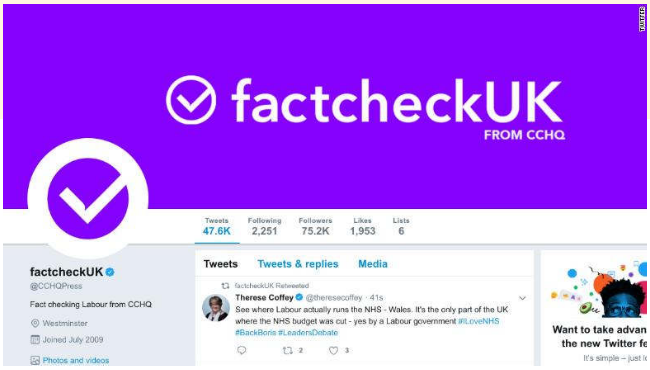
Significant usage of alternative or state- sponsored news websites has been measured in the US, the UK, France, Sweden, Norway, and Brazil (Newman, Fletcher, Kalogeropoulos, & Nielsen, 2019). In 2018, 22 % of the population of the US used an alternative, highly partisan, state-owned or state-sponsored news website such as Breitbart, Sputnik, RT, Daily Caller, Infowars or The Intercept at least once a week, while in the UK only 7 % usage was measured (Newman, Fletcher, Kalogeropoulos, & Nielsen, 2019).

Journalists often unintentionally become active actors of disinformation when they take up and spread narratives of certain operations. This gives the narratives additional credibility and increases their reach.

Example of disinformation: Renaming verified accounts

An example of disinformation can be found in the renaming of the Twitter account of the British conservative party @CCHQPress to "factcheckUK" during the debate between the candidates Boris Johnson and Jeremy Corbyn in the 2019 election campaign in the UK (Lee, 2019). After the televised debate was over, the account was renamed @CCHQPress again. Twitter warned the conservative party and referenced its Community Policy, which is intended to avoid and sanction misleading behavior, especially for verified accounts.

Figure 9: Renamed account of the British conservative party @CCHQPress on Twitter during the TV debate in the 2019 election



TWITTER

factcheckUK
FROM CCHQ

Tweets 47.6K | Following 2,251 | Followers 75.2K | Likes 1,953 | Lists 6

factcheckUK
@CCHQPress
Fact checking Labour from CCHQ
Westminster
Joined July 2009
Photos and videos

Tweets | Tweets & replies | Media

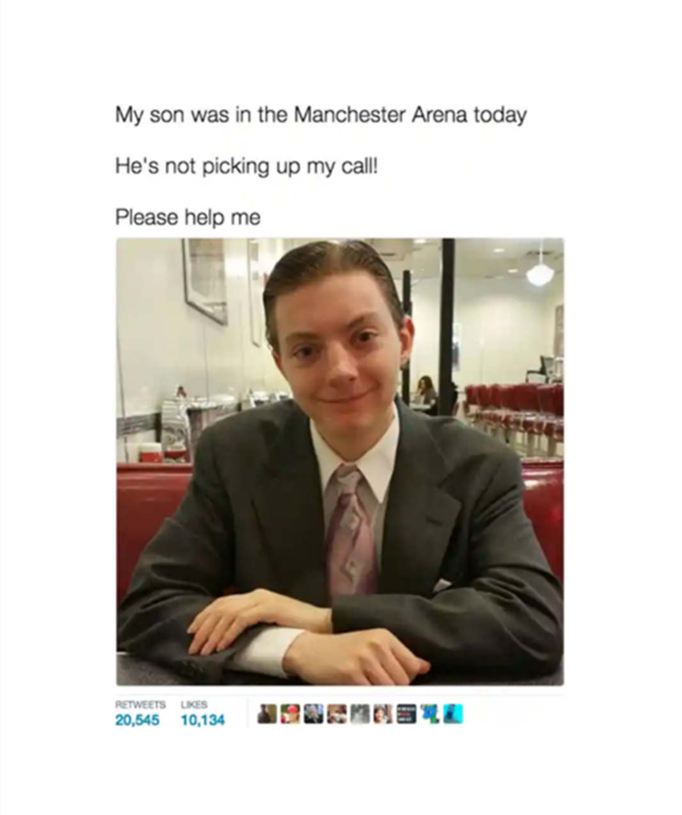
factcheckUK Retweeted
Therese Coffey @theresecoffey · 41s
See where Labour actually runs the NHS - Wales. It's the only part of the UK where the NHS budget was cut - yes by a Labour government #ILoveNHS #BackBoris #LeadersDebate
2 3

Want to take advan
the new Twitter fe
It's simple – just l

Example of disinformation as a tactic during terrorist attacks

In past years, disinformation has appeared particularly frequently during terrorist attacks. Shortly after the Manchester arena bombing in 2017, many inauthentic Twitter accounts coordinated in distributing the message that their friends or relatives were missing.

Figure 10:Fake tweet during the terrorist attack in Manchester in 2017



My son was in the Manchester Arena today

He's not picking up my call!

Please help me

They asked people for help to find their missing relatives or friends. The accounts used publicly accessible photographs of users, YouTubers, bloggers, and journalists, allowing them to reach additional communities and target audiences that were connected with these people. The YouTuber "The Report Of The Week" was among them. He reacted by explaining in a video that he was in the US and was still alive (Week, 2017).

The dismay felt by a young target audience and their parents in the social web about the supposed fates led to the terrorist attack spreading quickly and extensively (Cresci, 2017). The dismay and uncertainty that was created by the disinformation online reached far more people than the physical terrorist attack itself (Eder, 2017).

Key skills in fighting disinformation

The ability to identify, understand and process information from online texts, images, videos, and feeds is a key skill in countering disinformation and misinformation. Since 2017, many initiatives and projects have been created by volunteer organizations, companies, universities, and state-sponsored programs across the globe to educate people in media literacy. They approach children, adults, and journalists.

To support good journalism, it helps to have solid, practical training that develops advanced competence in online investigation and the proper handling of news sources. In daily business, it is essential to take the time to review information before it is going to be published. In addition, news outlets need to develop and implement a code of conduct regarding how and whether information threats should be covered.

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting World Leaders against Deep Fakes. Retrieved from The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 2019, pp. 38-45: https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19a.pdf

Amodei, D., & Hernandez, D. (2018, May 16). AI and Compute. Retrieved from OpenAI.

Andrews, C., Fichet, E., Ding, Y., Spiro, E., & Starbird, K. (2016, February 27). Keeping Up with the Tweet-dashians: The Impact of "Official" Accounts on Online Rumoring. Retrieved from Washington University: https://faculty.washington.edu/kstarbi/CSCW2016_Tweetdashians_Camera_Ready_final.pdf

Backes, T., Jaschensky, W., Langhans, K., Munzinger, H., Witzenberger, B., & Wormer, V. (2016). Timeline der Panik. Retrieved from Süddeutsche Zeitung: https://gfx.sueddeutsche.de/apps/57eba578910a46f716ca829d/www/

botswatch Technologies. (2017, September 21). Anteil der Aktivität von Social Bots kurz vor der Bundestagswahl 2017. Retrieved from https://twitter.com/botswatch/status/910863520035688449

Cresci, E. (2017, May 26). The story behind the fake Manchester attack victims. Retrieved from The Guardian: https://www.theguardian.com/technology/2017/may/26/the-story-behind-the-fake-manchester-attack-victims

DiResta, R., & Grossman, S. (2019). Potemkin Pages and Personas: Assessing GRU Online Operations 2014-2019. Retrieved from Stanford Internet Observatory Cyber Policy Center: https://cyber.fsi.stanford.edu/io/publication/potemkin-think-tanks

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Johnson, B. (2018, December). The Tactics & Tropes of the Internet Research Agency. Retrieved from https://cdn2.hubspot.net/hubfs/4326998/ira-report-rebrand_FinalJ14.pdf

Eddy, M. (2019, January 4). Hackers Leak Details of German Lawmakers, Except Those on Far Right. Retrieved from New York Times: https://www.nytimes.com/2019/01/04/world/europe/germany-hacking-politicians-leak.html

Eder, S. (2017, May 24). Fake News nach Manchester – In so einer Dimension gab es das noch nie. Retrieved from Frankfurter Allgemeine Zeitung: https://www.faz.net/aktuell/gesellschaft/kriminalitaet/fakenews-nach-manchester-in-so-einer-dimension-gab-es-das-noch-nie-15031082.html

Facebook. (2018, August 21). Taking Down More Coordinated Inauthentic Behavior. Retrieved from Newsroom: https://about.fb.com/news/2018/08/more-coordinated-inauthentic-behavior/

Facebook. (2019, October 21). Removing More Coordinated Inauthentic Behavior From Iran and Russia. Retrieved from Newsroom: https://about.fb.com/news/2019/10/removing-more-coordinated-inauthentic-behavior-from-iran-and-russia/

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016, July). The Rise of Social Bots. (Communications of the ACM, Vol. 59 No. 7, Pages 96-104) Retrieved from https://cacm.acm.org/magazines/2016/7/204021-the-rise-of-social-bots/fulltext

Finley, K. (2015, August 23). Pro-Government Twitter Bots Try to Hush Mexican Activists. Retrieved from Wired: https://www.wired.com/2015/08/pro-government-twitter-bots-try-hush-mexican-activists/

Freedberg, S. (2019, Otober 21). The Golden 5 Minutes': The Need For Speed In Information War. Retrieved from Breaking Defense: https://breakingdefense.com/2019/10/the-golden-five-minutes-the-need-for-speed-in-information-war/

Gerken, T. (2018, November 5). Twitter: Fake Elon Musk scam spreads after accounts hacked. Retrieved from BBC: https://www.bbc.com/news/technology-46097853

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019, January). Fake news on Twitter during the 2016 U.S. Presidential Election. Retrieved from Science, Vol. 363, Issue 6425, pp. 374-378: https://science.sciencemag.org/content/363/6425/374

Harwell, D. (2018, December 30). Fake-porn videos are being weaponized to harass and humiliate women: "Everybody is a potential target." Retrieved from Washington Post: https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/

Harwell, D. (2019, May 4). Faked Pelosi videos, slowed to make her appear drunk, spread across social media. Retrieved from Washington Post,: https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media

Howard, P. (2016, November 17). Pro-Trump highly automated accounts "colonised" pro-Clinton Twitter campaign. Retrieved from University of Oxford: http://www.ox.ac.uk/news/2016-11-17-pro-trump-highly-automated-accounts-%E2%80%98colonised%E2%80%99-pro-clinton-twitter-campaign

Howard, P. (2018, February 17). The Production And Detection Of Bots. Retrieved from University of Oxford: https://www.oii.ox.ac.uk/blog/the-production-and-detection-of-bots/

Howard, P., & Kollanyi, B. (2016, June 21). Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2798311

Ingram, D. (2019, September 5). A face-swapping app takes off in China, making AI-powered deepfakes for everyone. Retrieved from NBC: https://www.nbcnews.com/tech/security/face-swapping-app-takes-china-making-ai-powered-deepfakes-everyone-n1049501

Institute for Strategic Dialogue. (2019). The Battle for Bavaria: Online information campaigns in the 2018 Bavarian State Election. Retrieved from https://www.isdglobal.org/wp-content/uploads/2019/02/The-Battle-for-Bavaria.pdf

Kavanagh, J., & Rich, M. (2018). Truth Decay. An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life. Retrieved from RAND Corporation: https://www.rand.org/pubs/research_reports/RR2314.html

Kirby, E. (2016, December 5). The city getting rich from fake news. Retrieved from BBC: https://www.bbc.com/news/magazine-38168281

Kollanyi, B., Howard, P., & Woolley, S. (2016, October 5). Bots and Automation over Twitter during the U.S. Election. Retrieved from Oxford Internet Institute: https://com prop.oii.ox.ac.uk/wp-content/uploads/sites/89/2016/11/Data-Memo-US-Election.pdf

Kuo, L. (2018, November 9). World's first AI news anchor unveiled in China. Retrieved from The Guardian: https://www.theguardian.com/world/2018/nov/09/worlds-first-ai-news-anchor-unveiled-in-china

Lazer, D., Baum, M., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017, May 2). Combating Fake News: An Agenda for Research and Action. Retrieved from Shorenstein Center at Harvard Kennedy School: https://www.sipotra.it/wp-content/uploads/2017/06/Combating-Fake-News.pdf

Lee, D. (2019, November 20). Election debate: Conservatives criticised for renaming Twitter profile "factcheckUK". Retrieved from BBC: https://www.bbc.com/news/technol ogy-50482637

Lepore, J. (2016, March 14). After the Fact. In the history of truth, a new chapter begins. Retrieved from The New Yorker: https://www.newyorker.com/magazine/2016/03/21/the-internet-of-us-and-the-end-of-facts

Lin, H., & Kerr, J. (2019, May). On Cyber-Enabled Information Warfare and Information Operations. Retrieved from Oxford Handbook of Cybersecurity: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3015680

Maheshwari, S. (2018, March 12). "Uncovering Instagram Bots With a New Kind of Detective Work". Retrieved from New York Times: https://www.nytimes.com/2018/03/12/business/media/instagram-bots.html

Mazarr, M., Bauer, R. M., Casey, A., Heintz, S. A., & Matthews, L. (2019, October). The Emerging Risk of Virtual Societal Warfare. Social Manipulation in a Changing Information Environment. Retrieved from Research Report, RAND Corporation: https://www.rand.org/pubs/research_reports/RR2714.html

Mirian, A. (2019, December). Hack for Hire. Retrieved from Communications of the ACM, Vol. 62 No. 12, Pages 32-37, 10.1145/3359386: https://cacm.acm.org/magazines/2019/12/241053-hack-for-hire/fulltext

Morris, L., Mazarr, M., Hornung, J., Pezard, S., Binnendijk, A., & Kepe, M. (2019, July). Gaining Competitive Advantage in the Gray Zone. Response Options for Coercive Aggression Below the Threshold of Major War. Retrieved from Research Report, RAND Corporation: https://www.rand.org/pubs/research_reports/RR2942.html

Nelson, A., & Lewis, J. (2019, Ocotber 23). Trust Your Eyes? Deepfakes Policy Brief. Retrieved from Center for Strategic and International Studies (CSIS): https://www.csis.org/analysis/trust-your-eyes-deepfakes-policy-brief

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. (2019, June). Reuters Institute Digital News Report 2019. Retrieved from Reuters Institute, University of Oxford: http://www.digitalnewsreport.org/

Perez, S. (2019, November 12). Twitch publicly launches its free broadcasting software. Retrieved from Techcrunch: https://techcrunch.com/2019/11/12/twitch-publicly-launches-its-free-broadcasting-software-twitch-studio

Rinehart, A. (2017, June 22). Reporting on a new age of digital astroturfing. Retrieved from First Draft: https://firstdraftnews.org/latest/digital-astroturfing/

Runow, T. (2017, January 10). Wenn offizielle Stellen schweigen, sind Social Bots erfolgreich. Retrieved from Deutschlandfunk: https://www.deutschlandfunk.de/soziale-netzwerke-wenn-offizielle-stellen-schweigen-sind.807.de.html?dram:article_id=376020

Sarwari, K. (2019, July 19). You gave away the rights to your face. The one you use to unlock your phone. Retrieved from Northeastern University: https://news.northeastern.edu/2019/07/19/we-cant-get-enough-of-faceapp-but-should-we-be-giving-away-the-rights-to-our-faces

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini , A., & Menczer, F. (2018). The spread of low-credibility content by social bots. Retrieved from Nature Communications 9, Article number 4787: https://www.nature.com/articles/s41467-018-06930-7

Sharma, M., & Scarr, S. (2019, November 28). Wiki wars: Hong Kong's online frontline. Retrieved from Reuters : https://graphics.reuters.com/HONGKONG-PROTESTS-WIKIPEDIA/0100B33629V/index.html

Stubbs, J. (2019, March 15). 17 minutes of carnage: how New Zealand gunman broadcast his killings on Facebook. Retrieved from Reuters: https://www.reuters.com/article/us-newzealand-shootout-livestreaming/17-minutes-of-carnage-how-new-zealand-gunman-broadcast-his-killings-on-facebook-idUSKCN1QW294

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

Twitter Inc. (2014). 2Q 2014 Earnings Report. Retrieved from Financial Information: https://s22.q4cdn.com/826641620/files/doc_financials/2014/q2/2014_Q2_Earnings_Slides_-_Updated_NEW.pdf

US Department of Justice. (2019, March). Report On The Investigation Into Russian Interference In The 2016 Presidential Election. Retrieved from Volume I of II Special Counsel Robert S. Mueller: https://www.justice.gov/storage/report.pdf

US Director of National Intelligence DNI. (2019, November 5). Ensuring Security of 2020 Elections. Retrieved from Joint Statement from DOJ, DOD, DHS, DNI, FBI, NSA, and CISA: https://www.dni.gov/index.php/newsroom/press-releases/item/2063-joint-statement-from-doj-dod-dhs-dni-fbi-nsa-and-cisa-on-ensuring-security-of-2020-elections

US-Army. (2003, November). Information Operations: Doctrine, Tactics, Techniques and Procedures. Retrieved from Field Manual No. 3-13: https://fas.org/irp/doddir/army/fm3-13-2003.pdf

Volz, D. (2017, May 6). U.S. far-right activists, WikiLeaks and bots help amplify Macron leaks. Retrieved from Reuters: https://de.reuters.com/article/uk-france-election-cyber/u-s-far-right-activists-wikileaks-and-bots-help-amplify-macron-leaks-researchers-idUKKBN1820QJ

Wakefield, J. (2019, December 24). Russia 'successfully tests' its unplugged internet. Retrieved from BBC Technology: https://www.bbc.com/news/technology-50902496

Wardle, C. (2017, February 16). Fake news. It's complicated. Retrieved from First Draft: https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79

Wardle, C., & Darakshan, H. (2017, September 27). Information Disorder. Toward an interdisciplinary framework for research and policy making. Retrieved from Council of Europe: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Week, T. R. (2017, May 22). I am alive. Retrieved from YouTube Channel: https://youtu.be/0s70gbdf4AY

Yi, X., Walia, E., & Babyn, P. (2019, December). Generative Adversarial Network in Medical Imaging: A Review. Retrieved from Medical Image Analysis, Volume 58: https://doi.org/10.1016/j.media.2019.101552

///