Can Social Media Save Truth?

A Study on Fighting Disinformation by Social Media Platforms
and the Impact on Freedom of Expression

Master Thesis for the Executive Master Cyber Security

Cyber Security Academy

Richard Rensman
Student ID: s2250756
December 9, 2019

Thesis supervisors:
Prof. dr. B. van den Berg, Universiteit Leiden
Dr. T. van Steen, Universiteit Leiden

Abstract

Social media were once seen as a liberating technology which empowered people around the world to speak out, allowing every voice to be heard. But it also allowed the thriving of misleading information, such as conspiracy theories and political influencing. After the 2016 US presidential elections it was evident that social media were also a very convenient tool for Russian disinformation campaigns. Social media had become a threat to the freedom of expression. Social media companies have since been under pressure to prevent disinformation from influencing public opinion. This gives them a tough struggle. Too much intervention affects the freedom of expression, whereas too little intervention lets disinformation drown out reliable sources of information.

This thesis examines whether fighting disinformation by social media platforms is causing a new form of censorship on the internet. This is done by first clarifying the definition of disinformation and explaining the elements of communication and phases of the life cycle of information. In order to understand how information spreads through social media platforms, the mechanisms of social media technology are explored. It is examined how algorithms and people's biases cause filter bubbles and echo chambers, explaining why users are susceptible for believing misleading information.

The efforts of social media companies to fight disinformation are discussed. Governments, too, have proposed measures against disinformation. While helpful, these measures cannot be applied blindly. The limitations of fact-checking, of transparency, of algorithms, and of media and information literacy, are considered in this research.

Finally the world of internet filtering is explored. Next to interventions by governments we look into the practices of restricting user behavior by social media platforms. Intervention by social media, by means of their algorithms, is done in such a subtle, personalized way, that it is a very opaque form of internet filtering.

We need social media to actively counteract the influencing by misleading content. To respect the right to freedom of expression we need the workings of social media platforms to be transparent in a comprehensible way. We need users to become active, critical thinkers, who recognize disinformation. To prevent social media companies from having to make all the decisions about intervention on their own we need a broad discussion between citizens, governments, academia, and tech companies together.

Keywords: social media, disinformation, freedom of expression, internet filtering

Preface

> *‘Who controls the past’, ran the Party slogan, ‘controls the future: who controls the present controls the past.’*
>
> ***George Orwell, Nineteen Eighty-Four***

The year is 1984. Citizens of Airstrip One are governed by the Party. The Ministry of Truth is continuously revising historical records to align them with the always changing views of the Party, depending, for example, on which country is considered an ally at that particular moment, and which is considered an enemy. Having ideas and thoughts of one’s own is a grave offense called thoughtcrime. People speak a constructed language called Newspeak, that has a limited vocabulary. The aim of Newspeak is to narrow the range of thought and make thoughtcrime literally impossible by eliminating the words needed to express divergent ideas.

The world depicted in George Orwell’s novel *Nineteen Eighty-Four* is the opposite of the ideal democracy with freedom of speech. It describes the ultimate authoritarianism, in which people are denied even the power to think for themselves. This goes way beyond the Germany of Adolf Hitler or the Soviet Union of Joseph Stalin. It serves us as a terrifying vision, instructing what we don’t want our world to be.

There is another threat to the freedom of expression. It doesn’t result from being limited in available information. On the contrary, it can be caused by overabundance. When people are overwhelmed by too much information their attention is diverted and limited to content they feel drawn to. They don’t discern real from fake anymore and choose whatever they like most. This also affects freedom of expression because one’s voice can be overpowered by information from unreliable sources.

If people seek only comfort and entertainment, they are occupied with distractions and loose interest in trustworthy news. In Aldous Huxley’s 1932 novel *A Brave New World* people no longer care about truth because they are genetically and psychologically conditioned to be happy with their lives. They blissfully accept their positions and fill their lives with little pleasantries. This novel is showing us an equally disturbing vision, in which people have lost the capability of judging information and thinking critically. Not because they are restrained by restrictions on speech or vocabulary, but simply by having lost care for truth.

Today social media platforms play a major role in the communication of a lot of people and have an impact on which information contributes to people’s ideas. Once heralded as the real heroes of freedom of speech through which anyone can express themselves, social media platforms are now a serious source of misleading information and threaten the very foundations of our society by affecting our ability to judge information. Fake stories already generated more engagement than mainstream media news during the 2016 US presidential elections (Silverman, 2016).

Our lesser judgment of truth is perfectly illustrated by the Oxford Dictionaries Word of the Year 2016, which, after the outcome of the 2016 US presidential elections and the Brexit referendum, was *post-truth*, 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.” (“Word of the Year 2016 is...,” 2016) The term post-truth had already been coined by Steve Tesich in 1992, referring to the attitude of the Americans after public administrations had withhold important information in various scandals (Kreitner, 2016).

We are losing trust in science and mainstream media, and we are drawn to conspiracy theories in which the traditional media are depicted as a liberal elite rather than the primary source of reliable information (D’Ancona, 2017). The sense of truth is declining. Tesich concluded that “we are rapidly becoming prototypes of a people that totalitarian monsters could only drool about in their dreams. All the dictators up to now have had to work hard at suppressing the truth. We, by our actions, are saying that this is no longer necessary, that we have acquired a spiritual mechanism that can denude truth of any significance. In a very fundamental way we, as a free people, have freely decided that we want to live in some post-truth world.” (Kreitner, 2016)

This thesis examines how social media, at one hand, are restricting what we see by means of their algorithms, while at the same time overwhelming us with irrelevant and misleading information. We start to recognize the nightmares of both Orwell and Huxley, which is a rather alarming observation. Can we fully understand the impact of social media platforms on our freedom of expression?

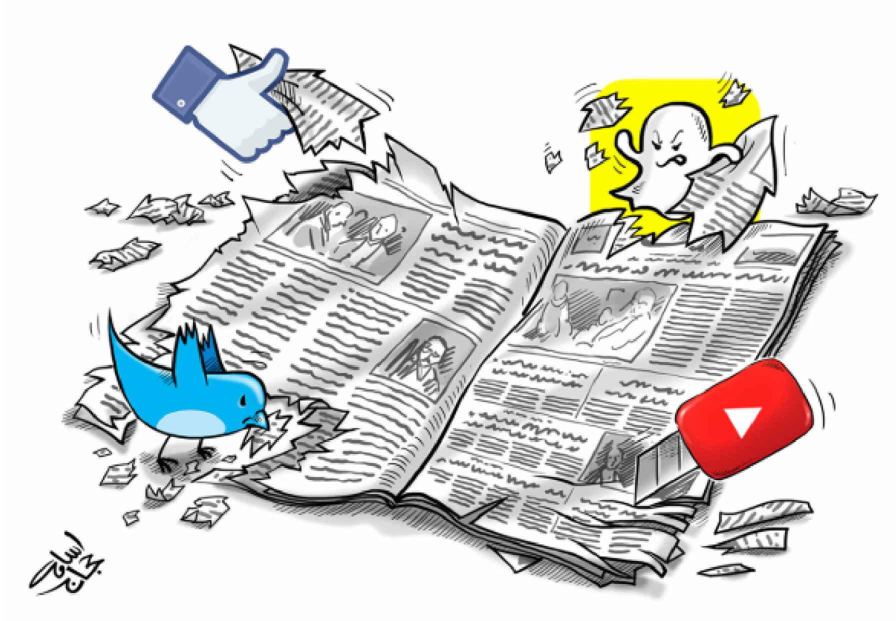


Figure 1: Cartoon by Osama Hajjaj, drawn for UNESCO for World Press Freedom Day 2019. Source: https://unesco.exposure.co/cartoons-for-freedom-of-expression-2019/photos/5820669

Table of Contents

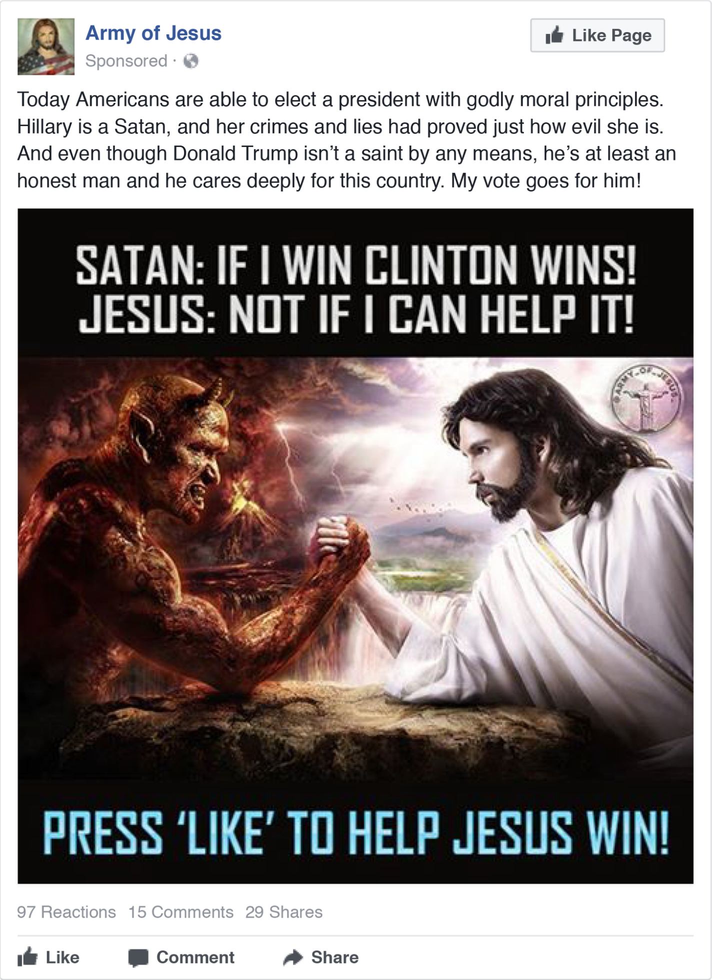
List of Figures

List of Tables

1. Introduction

> *Our success is built on getting people the stories that matter to them most. If you could look through thousands of stories every day and choose the 10 that were most important to you, which would they be? The answer should be your News Feed. It is subjective, personal, and unique — and defines the spirit of what we hope to achieve.*
>
> ***Facebook*** *(Mosseri, 2016)*

In 2016 two big events took place in which social media played a big role. The US elected a new president and the UK citizens voted for a future independent of the European Union. Among the information spread through social media was also misleading content intended to influence voters.



Army of Jesus

Sponsored ·

Today Americans are able to elect a president with godly moral principles. Hillary is a Satan, and her crimes and lies had proved just how evil she is. And even though Donald Trump isn't a saint by any means, he's at least an honest man and he cares deeply for this country. My vote goes for him!

Like Page
SATAN: IF I WIN CLINTON WINS!
JESUS: NOT IF I CAN HELP IT!
ARMY OF JESUS
PRESS 'LIKE' TO HELP JESUS WIN!
97 Reactions 15 Comments 29 Shares
Like
Comment
Share

Figure 2: Example of Russion disinformation via Facebook advertisement during 2016 US Presidential elections. Source: https://www.nytimes.com/2017/11/01/us/politics/russia-2016-election-facebook.html

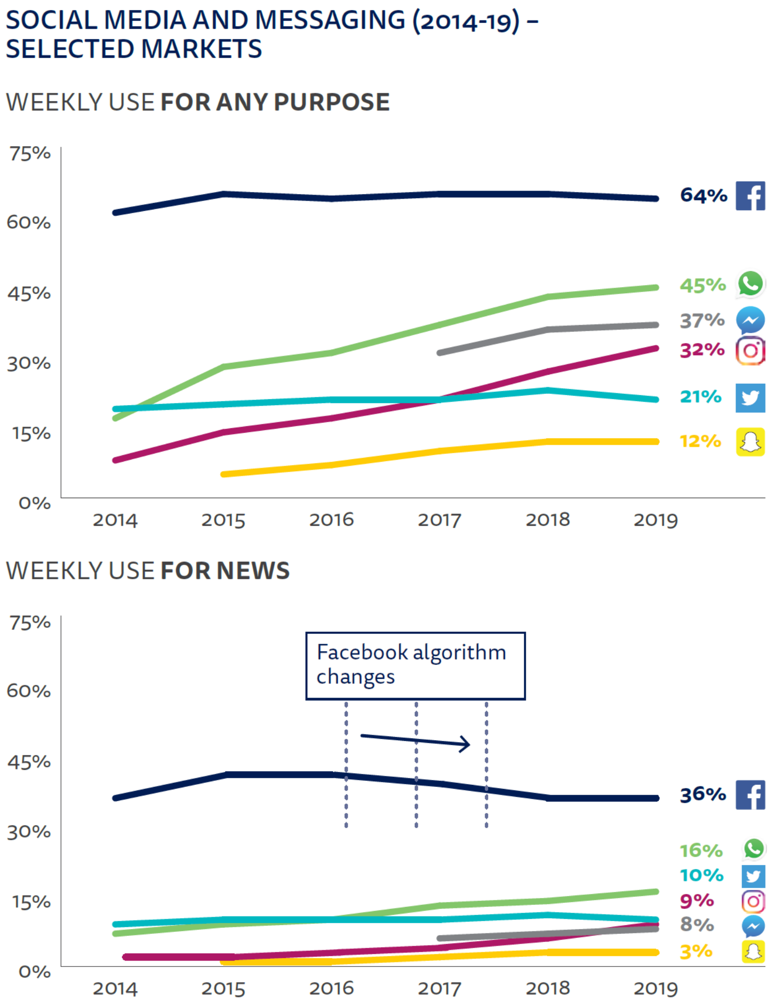
After the 2016 US presidential elections it became clear that not only the national campaign teams had spread information for influencing, but also that the Russians had used information on social media targeted at American voters. The use of misleading information was a tactic familiar to the Kremlin, who had a history of using propaganda, especially during the Soviet era. In 1992 the then KGB director Yevgeny Primakov admitted that the KGB was responsible for spreading the idea that Acquired Immune Deficiency Syndrome (AIDS), which emerged around 1980, was created by US government scientists (Boghardt, 2009).

The difference with the 2016 US presidential elections interference is the scale and the ease with which disinformation could be spread. The Russian Internet Research Agency (IRA) "conducted social media operations targeted at large US audiences with the goal of sowing discord in the US political system." (Mueller, 2019, p. 14) Facebook had identified fake accounts and Facebook pages run by the IRA, that had been posting over 80,000 pieces of content between January 2015 and August 2017, reaching 29 million people directly and 126 million people indirectly ("Hearing Before the United States Senate Select Committee on Intelligence," 2017). Twitter had identified 3,814 accounts run by the IRA, which had posted 175,993 tweets, approximately 8,4% of which were election-related ("Update on Twitter's Review of the 2016 US Election," 2018).

Even today new cases are being detected. Research on supposedly leaked documents trying to influence the debate during the UK elections of December 2019 strongly suspects Russian interference via social media (Nimmo, 2019).

The importance of social media nowadays is obvious. They have given us the opportunity to connect and communicate with anyone across the world and have empowered us to express ourselves. On average people spend more than two hours per day on social media (Salim, 2019). Although somewhat declined in the last three years, a significant number of people still use social media as a source of news, while the use of social media in general is still growing, according to Reuters (Newman, Fletcher, Kalogeropoulos, & Nielsen, 2019), see figure 3. Especially among young people social media has more impact than traditional news media. The Swedish gamer Felix Kjellberg owns PewDiePie, the YouTube channel with the most subscribers, over 100 million, and to people under 25, he has more exposure than any news channel (boyd, 2019).

This means that social media as a source of news and information, true or false, has become a phenomenon that cannot be ignored. The European Commission states that "new technologies can be used, notably through social media, to disseminate disinformation on a scale and with speed and precision of targeting that is unprecedented, creating personalised information spheres and becoming powerful echo chambers for disinformation campaigns.“ (*Tackling Online Disinformation: a European Approach*, 2018, p. 1)



Q12a/b. Which, if any, of the following have you used for any purpose/for news in the last week? *Base: Total 2014-19 sample in each country: 18,859/23,557/24,814/24,487/24,735/24,146. Note: From 2015-19 the 12 countries included are UK, US, Germany, France, Spain, Italy, Ireland, Denmark, Finland, Japan, Australia and Brazil. In 2014, we did not poll in Australia or Ireland.*

Figure 3: Use of social media in general and for news. Source: Reuters Institute Digital News Report 2019

1.1 Freedom of Expression

Social media companies have been called upon to take their responsibility in fighting disinformation on their platforms. This has resulted in social media platforms increasing their focus on removing inappropriate content and reducing the distribution of disinformation. The question arises when exactly to intervene without affecting the users' ability to express their opinions and beliefs. Social media have come to play a big role in the right to freedom of expression and the right to information, which is described as a fundamental human right in various documents. The UN has declared in Article 19 of the Universal Declaration of Human Rights: ("Universal Declaration of Human Rights," 1948)

> *Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.*

Article 10 of the European Convention of Human Rights reads: ("Convention for the Protection of Human Rights and Fundamental Freedoms," 1950)

> *1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.*
>
> *2. The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.*

Article 11 of Charter of Fundamental Rights of the European Union is called the 'freedom of expression and information' and reads: ("Charter of Fundamental Rights of the European Union," 2012)

> *1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.*
>
> *2. The freedom and pluralism of the media shall be respected.*

In all three documents the right to freedom of expression and the right to information is part of the universal human rights. The question is whether social media platforms can be entrusted with the burden of making the right decisions. Are they the ones to decide what content is to be allowed? Are they capable to intervene in a way that respects the freedom of expression? Can they save truth?

That is the dilemma.

1.2 Research Question

To get a better understanding of this dilemma the main question of this research has been formulated:

> **When does fighting disinformation by social media platforms become a form of internet filtering?**

Before being able to answer this question a number of steps need to be taken. First the term disinformation must be clarified. It also has to be determined what social media are and how the distribution of information on their platforms happens. With the definition and the understanding of the mechanisms of social media platforms the fight against disinformation by social media can be looked into. To judge the effectiveness of this fight the limitations of current practices need to be examined. Finally the term and the practices of internet filtering require exploring to be able to answer the research question.

This leads to the following subquestions for this research:

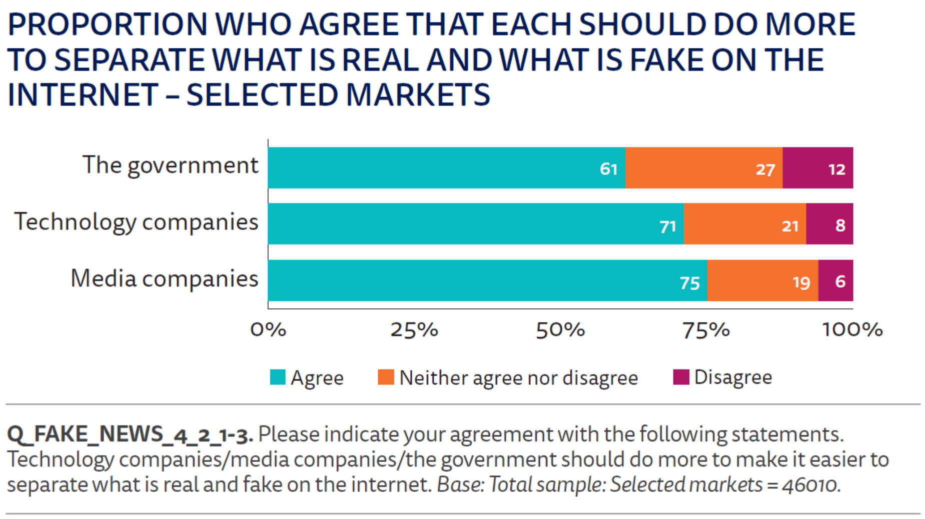
- What is disinformation?
- How does (dis)information spread via social media?
- How are social media platforms combating disinformation?
- What are considerations when balancing the right to freedom of expression and the fight against disinformation?
- What is internet filtering?

Chapters 2 – 3 and 5 – 7 are each dedicated to one of the subquestions. In chapter 4 the conspiracy theory of the flat earth is used as a case study to illustrate the definitions and mechanisms explained in chapters 2 and 3.

1.3 Relevance

People are concerned about fake news and disinformation and are losing trust in news media, journalism, and social media companies (Nielsen & Graves, 2017). Because of this concern a majority of people in several countries agree upon the need to combat disinformation by media and technology companies. Whether governments should act to combat disinformation is less agreed upon, see figure 4. It differs per country, with the lowest figure in the US, probably because the First Amendment of the US Constitution (Newman, Fletcher, Kalogeropoulos, Levy, & Nielsen, 2018).

People are now starting to see the drawbacks of new technology. The internet gives them a kind of anonymity that changes their behavior. “Online political discussions (often among anonymous strangers) are experienced as angrier and less civil than those in real life; networks of partisans co-create worldviews that can become more and more extreme; disinformation campaigns flourish; violent ideologies lure recruits.” (Haidt & Rose-Stockwell, 2019)



PROPORTION WHO AGREE THAT EACH SHOULD DO MORE TO SEPARATE WHAT IS REAL AND WHAT IS FAKE ON THE INTERNET – SELECTED MARKETS

Q_FAKE_NEWS_4_2_1-3. Please indicate your agreement with the following statements. Technology companies/media companies/the government should do more to make it easier to separate what is real and fake on the internet. *Base: Total sample: Selected markets = 46010.*

Figure 4: Who should be doing more to combat disinformation? Source: Reuters Institute Digital News Report 2018

Social media are taking action against disinformation, but there is limited knowledge about the risk of restricting the freedom of expression. It threatens to become a new form of internet filtering.

The intent of this research is to have a clearer understanding of the fight against disinformation by social media companies and the concept of internet filtering in order to be able to understand whether interventions by social media are actually not violating the right to freedom of expression and the right to information.

1.4 Methodology

This is a qualitative study. References have been made to quantitative research to illustrate whether findings correspond to theoretical concepts.

The definition of disinformation is needed as a starting point. It also delimits the problem. To comprehend the role of social media in fighting disinformation we need an understanding of the characteristics and workings of social media. With this knowledge the efforts of countering disinformation by social media platforms can be judged. We do this by looking into the efforts by Facebook and Twitter. We also make clear what limitations need to be taken into consideration when fighting disinformation. Finally the practices of internet filtering are discussed. This enables us to judge when the fighting of disinformation by social media platforms can be called a form of internet filtering.

Examples are given to illustrate the concepts and definitions. Examples are described within the framework to illustrate. A case study shows how the mechanisms of social media platforms have given rise to the spread of disinformation.

This research has been based on available literature from several sources. Academic articles and books have been used to substantiate the research with definitions and frameworks. Non-academic books, articles from newspapers, and online news or opinion sites have been consulted for past events and opinions. Governmental reports and documents give us insight in the motives and goals of governments towards dealing with disinformation. Lastly the information on the websites of social media platforms has proven to be an invaluable source of understanding the viewpoints of the social media themselves and the efforts they undertake.

1.5 Scope

There are a lot of social media platforms. This research has been focused primarily on Facebook and Twitter. Examples of studies on YouTube have been used where relevant. The reason for limiting to Facebook and Twitter has to do, on the one hand, with the size and reach of these platforms, and on the other side, with available literature. Facebook is by far the biggest platform with 2,375 billion monthly active users, of which 1,49 billion daily active (Smith, 2019a). Twitter has only 330 million active users and ranks lower than platforms like YouTube (1,9 billion users) or WhatsApp (1,6 billion users) or Instagram (1 billion users) (Smith, 2019b). However, Twitter is popular among politicians and influential people, and has an impact on public opinion. Both Twitter and Facebook encourage its users to interact actively, by means of retweet, share, comment and like functions.

Furthermore a lot of research has been done on Twitter and Facebook. They have been in the news regularly and both companies have testified in US Congress. Because of their visibility they have been the target of public attention and of the pressure to become more transparent, which has resulted in more information available to use in this research.

2. Disinformation

> *It is hardly possible to overrate the value, in the present low state of human improvement, of placing human beings in contact with persons dissimilar to themselves, and with modes of thought and action unlike those with which they are familiar. . . . Such communication has always been, and is peculiarly in the present age, one of the primary sources of progress.To human beings, who, as hitherto educated, can scarcely cultivate even a good quality without running it into a fault, it is indispensable to be perpetually comparing their own notions and customs with the experience and example of persons in different circumstances from themselves: and there is no nation which does not need to borrow from others, not merely particular arts or practices, but essential points of character in which its own type is inferior.*
>
> ***John Stuart Mill, Principles of Political Economy with Some of Their Applications to Social Philosophy, 1848***

Information is a powerful tool. It is used in communication between people and enables people to connect to each other and share ideas. It can also be used for persuasion or deception. Figure 5 shows an example of misleading information form a Facebook ad, posted by the Leave.EU campaign team during the Brexit referendum. It depicts the 5 countries that are candidates for EU membership, suggesting it to be very costly and to only result in extra migrants. The membership discussion is in progress for years and no decision has yet been made. There is no substantiation for the amount of £ 2 billion and the checkbox after the word “Fact” suggests checked facts, which makes this a misleading advertisement.

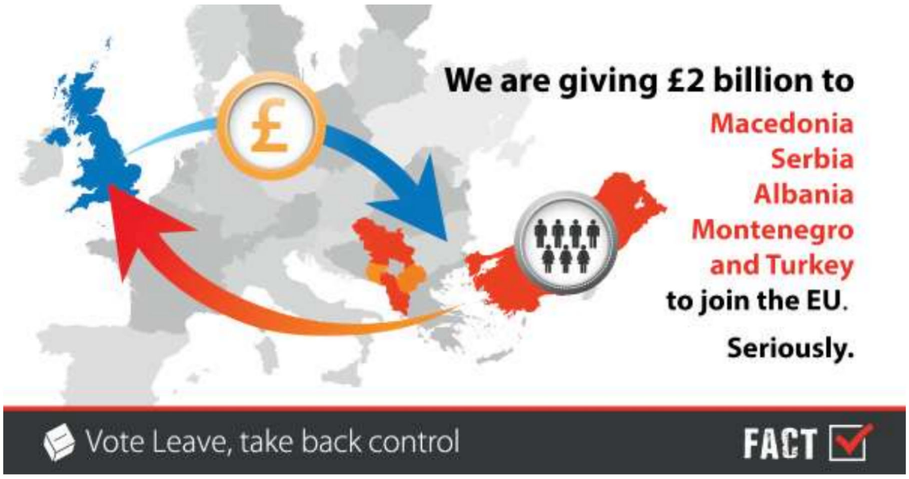
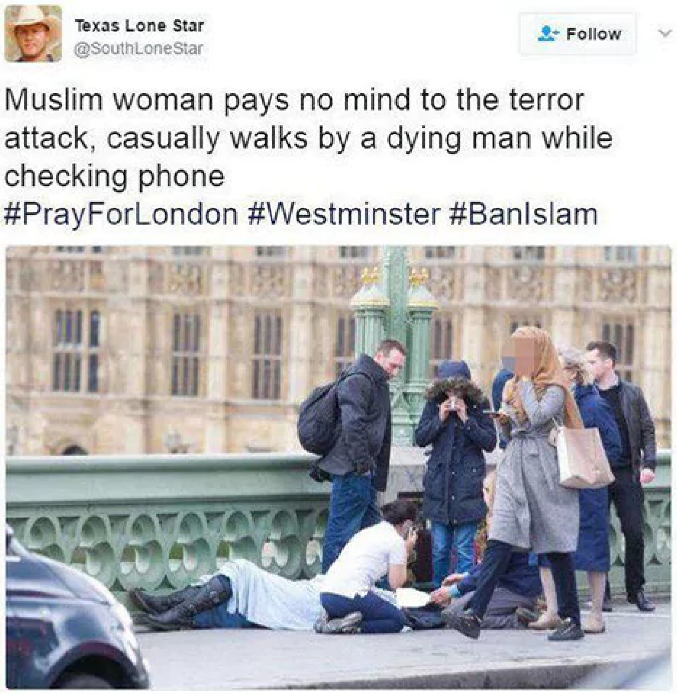


Figure 5: Misleading information by Leave.EU campaign
Source: https://www.parliament.uk/documents/commons-committees/culture-media-and-sport/Fake_news_evidence/Vote-Leave-50-Million-Ads.pdf

Figure 6 shows an example of disinformation, posted by Russian trolls on Twitter, wrongly accusing a Muslim woman of ignoring victims of the Westminster attack in 2017. The woman in the photograph later said she had been trying to help before calling her family to tell them she was fine (Seidel, 2017).

People can be persuaded to believe false information, which can have effect on opinions about political subjects. When there is a lot of political information available while trust in politics is low, people can easily fall into the trap of conspiracy thinking (Miller, Saunders, & Farhart, 2016). Or people can be manipulated by misleading information, which was clearly the intent of the examples in Figure 5 and Figure 6.



Texas Lone Star
@SouthLoneStar

Muslim woman pays no mind to the terror attack, casually walks by a dying man while checking phone
#PrayForLondon #Westminster #BanIslam

Figure 6: Example of Russian disinformation. Source: https://www.news.com.au/technology/online/social/us-congress-told-how-russia-weaponised-this-photo/news-story/8135ca050976761ab476c04c32d44fd2

This chapter examines the use of misleading information. The term propaganda used to be applied for the use of deceiving information with political motive. Recently the term fake news has surfaced. It is explained why this term is not adequate to cover the problem. Existing definitions of the term disinformation are evaluated in order to come up with a definition to be used in this research. To be able to understand the workings of disinformation in communication the elements of communication and the phases in the life of information are explored. This provides us with a framework which can be used to explore disinformation.

2.1 Propaganda & Fake News

Using information to deceive is by no means a recent phenomenon. Two examples show how powerful information can be in shaping the course of history.

In 33BC Rome was ruled by the second triumvirate. In a growing tension between Octavian, heir of Julius Ceasar, and Mark Antony, who had served as a general under Julius Ceasar, Octavian got hold of Antony's will and read it aloud in the Senate. It is debated whether the will was real, but it was an effective way of discrediting Mark Antony (Johnson, 2013).

In 1941 the US citizens didn't want to be involved in the European war. When president Franklin Roosevelt delivered the Navy Day Speech, he mentioned the attack of a US navy destroyer by a German U-Boot and informed the public of a map of South America, divided into four countries. The map with German annotations, clearly showing evidence of the Nazi's expansion plans into America, later turned out to be forged by the British Security Coordination to lure the US into the war (Boyd, 2014).

The use of information as a means of persuasion or manipulation in a political context has been called propaganda. Information operations has been used as a weapon in wars and has been attributed the term 'information warfare', with greatly varying definitions (Gioe, Goodman, & Wanless, 2019). This term has also been attributed to the fight between news media and public and political opinion (boyd, 2017a). In recent years the term 'fake news' has become popular. This term isn't suitable for our research because of three reasons. Firstly it does not cover all types of information. For our research we are not only interested in news, but in information in general. Secondly we not only want to include information that is evidently fake, but also information that has been manipulated or used outside its context. And thirdly the term is misleading because of its use by politicians who are trying to express their discord with news media (Wardle, 2018).

2.2 Definition of Disinformation

In spite of the long history of information campaigns, the word disinformation only surfaced in the 20th century. It is a translation of дезинформация, the KGB department of the Soviet Union dedicated to black propaganda (Jowett & O'Donnell, 1999). In the context of the KGB it means "false, incomplete, or misleading information that is passed, fed, or confirmed to a targeted individual, group, or country." (Jowett & O'Donnell, 1999, p. 28)

Many definitions of disinformation have been made. It has been classified as a subset of the broader concept of misinformation, which could result from mistakes and need not be intentional (Floridi, 2011). Disinformation on the other hand has the intention to mislead and can be called "misinformation with an attitude." (Fetzer, 2004, p. 231)

Don Fallis has argued that some definitions are too broad or too narrow. He does not include mistakes in his definition. Accidental truths, on the other hand, should be included, by which he means statements intended to be false that are actually true. On the other hand sarcasm, or satire, is not included because there is no intention to mislead. Fallis arrives at the rather constructed, theoretical definition of “misleading information that has the function of misleading.” (Fallis, 2015, p. 413)

The European Commission defines disinformation as “verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm. [...] Disinformation does not include reporting errors, satire and parody, or clearly identified partisan news and commentary.” (“Code of Practice on Disinformation,” 2018, p. 1)

Claire Wardle and Hossein Derakhshan write about ‘information disorder’ and categorize t into three categories, based on truthfulness and the intention to harm, see figure 7: (Wardle & Derakhshan, 2017)

- **Misinformation** is false information created without any intention to harm. This can result of error or inaccuracy. Satire and parody also belongs to this category.
- **Disinformation** is false information created with the intention to harm. This comprises misleading, fabricated or manipulated content. The examples of figures 5 and 6 at the beginning of this chapter belong to this category.
- **Malinformation** is true information that is meant to harm. Harassment and hate speech fall into this category. Leaks of classified information is another example, e.g. the publications by Wikileaks.

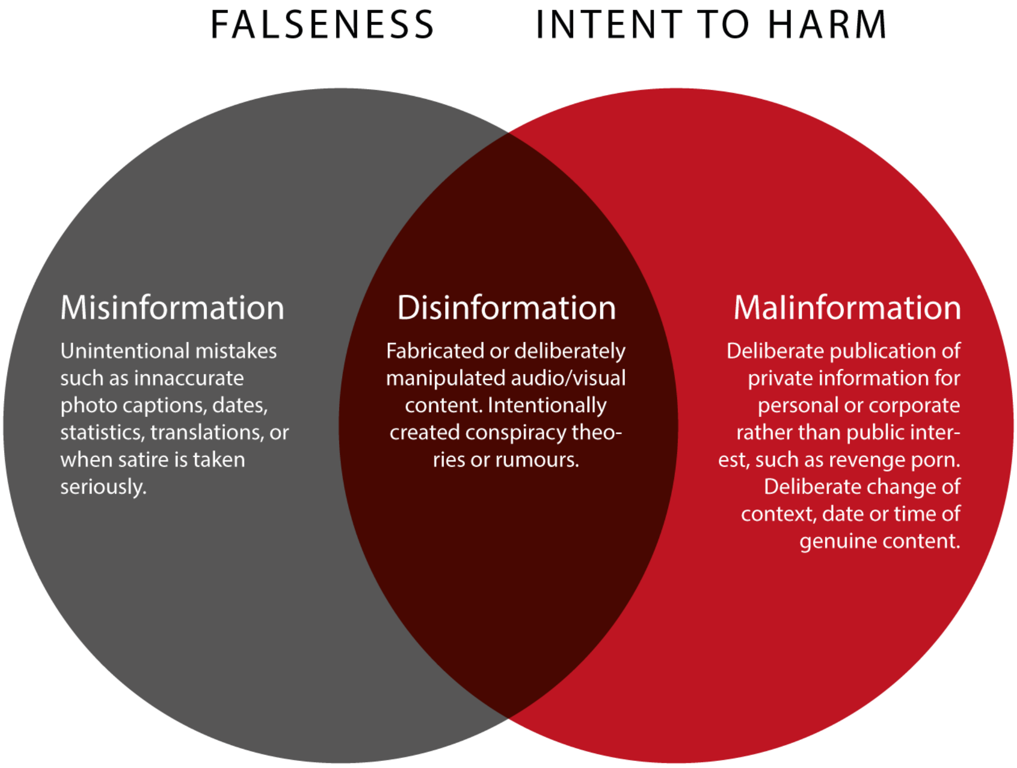

FALSENESS
INTENT TO HARM
Misinformation
Unintentional mistakes such as innacurate photo captions, dates, statistics, translations, or when satire is taken seriously.
Disinformation
Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.
Malinformation
Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of genuine content.

Figure 7: Three Types of Information Disorder. Credit: Claire Wardle & Hossein Derakhshan, 2017. Source: https://firstdraftnews.org/wp-content/uploads/2018/07/Types-of-Information-Disorder-Venn-Diagram.png

In all of these definitions of disinformation the actual intent to harm is present. Note that these definitions lack a statement whether this intent should be a conscious motivation to harm people or institutions. People can also spread information out of conviction or belief that they are telling the truth, without a conscious intention to mislead. This is the case with numerous conspiracy theories. People intend to tell the truth from their belief system, while actually distributing false information. Spreading this kind of disinformation undermines trust in governments or science, and is therefore to be considered harmful. The unconscious act of deception by spreading information one beliefs in, is to be included in the definition of disinformation, because it is still causing harm. For information on social media to be harmful it is also important that it reaches other people, so we include the distribution of information into the definition, which now reads:

> ***Disinformation is false information that has been distributed that can be wittingly or unwittingly harmful.***

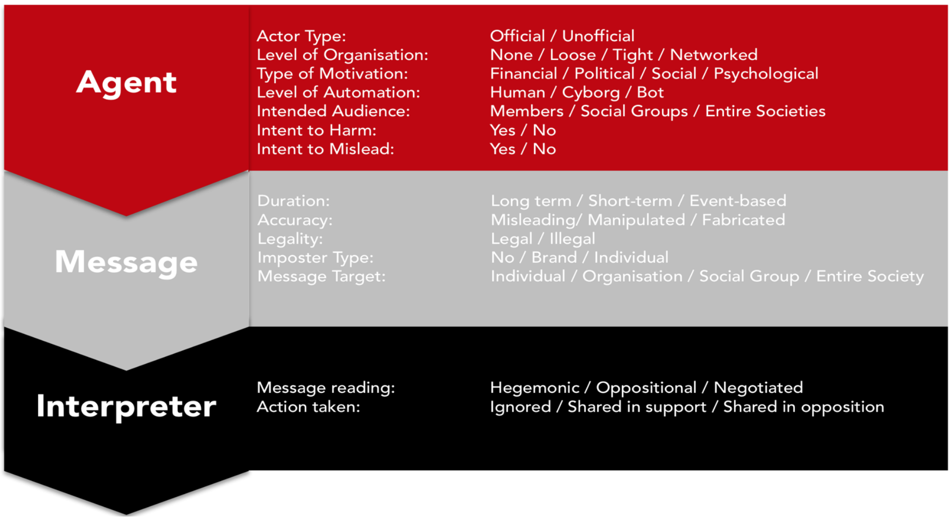
2.3 Communicating Information

To determine if harm is done by disinformation we need to understand the system in which it is used. Next to the definition of disinformation we have to describe how disinformation is distributed. This requires an understanding of the use of information in communication. We need to look at the parts that form communication and the steps that makes information spread from one place to another. This is done by adapting the framework of Claire Wardle and Hossein Derakhshan.

Elements of information in communication

The elements in the transfer of information distinguished by Wardle and Derakhshan as agent, message and interpreter, see figure 8: (Wardle & Derakhshan, 2017)

- **Agent**. This the one who has created and distributed the information. It could be a single person or an organization. It is important to consider their motivation, e.g. a financial or political goal, and the intention of the message. The agent may intend to mislead, or has an intention to harm.
- **Message**. Communication may go in the form of speech, text, audio, or video. It can be distributed in conversation, through newspaper, television, or social media. It can also be considered whether a message contains false or harmful information, and what audience is targeted.
- **Interpreter**. The message can provoke a rational or emotional reaction in the receiver. This may result in an external reaction, e.g. answering, or, in the case of social media, liking, sharing or commenting.



Agent

Actor Type: Official / Unofficial
Level of Organisation: None / Loose / Tight / Networked
Type of Motivation: Financial / Political / Social / Psychological
Level of Automation: Human / Cyborg / Bot
Intended Audience: Members / Social Groups / Entire Societies
Intent to Harm: Yes / No
Intent to Mislead: Yes / No

Message

Duration: Long term / Short-term / Event-based
Accuracy: Misleading/ Manipulated / Fabricated
Legality: Legal / Illegal
Imposter Type: No / Brand / Individual
Message Target: Individual / Organisation / Social Group / Entire Society

Interpreter

Message reading: Hegemonic / Oppositional / Negotiated
Action taken: Ignored / Shared in support / Shared in opposition

Figure 8: Three Elements of Information Disorder. Credit: Claire Wardle & Hossein Derakhshan, 2017. Source: https://firstdraftnews.org/wp-content/uploads/2018/07/Agent-Message-Interpreter.png

The examples of Figures 2, 5 and 6 can now be examined using the elements agent and message, see table 1. The effect on the interpreter cannot be deduced from the examples.

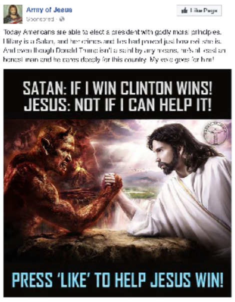
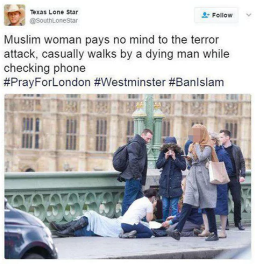
	 Army of Jesus Sponsored Like Page Today Americans are able to elect a president with godly moral principles. Hillary is a Satan, and her crimes and lies had proved just how evil she is. And even though Donald Trump isn't a saint by any means, he's at least an honest man and he cares deeply for this country. My vote goes for him! SATAN: IF I WIN CLINTON WINS! JESUS: NOT IF I CAN HELP IT! PRESS 'LIKE' TO HELP JESUS WIN!	 We are giving £2 billion to Macedonia Serbia Albania Montenegro and Turkey to join the EU. Seriously. Vote Leave, take back control FACT	 Texas Lone Star @SouthLoneStar Follow Muslim woman pays no mind to the terror attack, casually walks by a dying man while checking phone #PrayForLondon #Westminster #BanIslam
Example	US 2016 Presidential Election	UK 2016 Brexit referendum	2017 Westminster Attack
Agent			
Actor	Russian	Leave.EU team	Russian
Motivation	Help Trump win	Winning referendum	Discredit Islam
Harmful?	Yes	Yes	Yes
Message			
Accuracy	Fabricated	Misleading	Out of context
Target	US voters	UK voters	UK citizens

Table 1: Elements of Information Transfer in the Examples of Disinformation

Phases in the life of information

Next to these elements Wardle and Derakhshan consider the three phases in the life cycle of information, see figure 9: (Wardle & Derakhshan, 2017)

- **Creation**. Information starts with the creation of a message, e.g. a text, a photo, or a video.
- **Production**. A message has to be turned into a media product that can be published, e.g. a newspaper article, a Facebook message or a YouTube video.
- **Distribution**. A message needs to be transmitted to the receivers, e.g. the physical or digital distribution of newspapers to subscribers, the broadcast of a television show, or the display of a message in a Facebook timeline.

After the message is distributed, the receiver can process the information and produce a new message, e.g. by referencing to the original message. On Facebook and Twitter, the reproduction of information is easily done for example by sharing or retweeting.

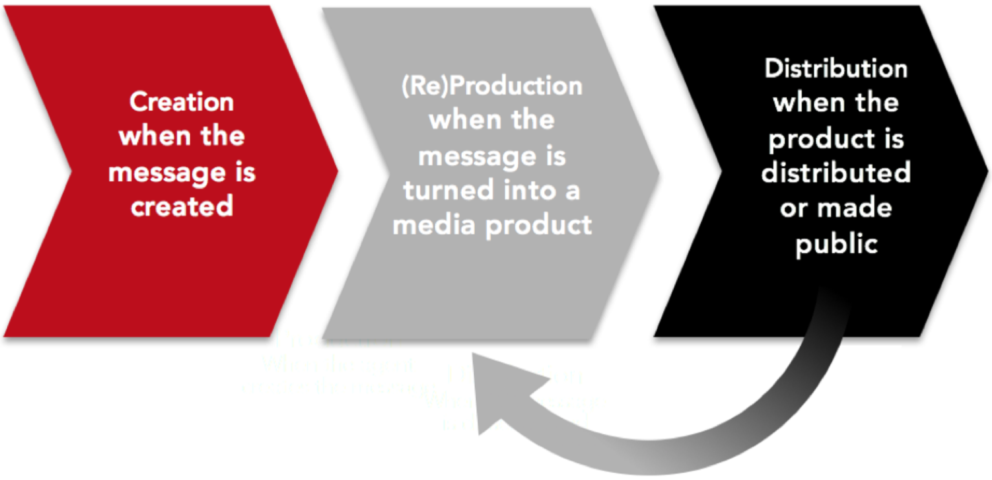

Creation when the message is created
(Re)Production when the message is turned into a media product
Distribution when the product is distributed or made public

Figure 9: Three Phases of Information Disorder. Credit: Claire Wardle & Hossein Derakshan, 2017. Source: https://firstdraftnews.org/wp-content/uploads/2018/07/Creation-Reproduction-Distribution.png

2.4 Conclusions

Information is powerful and it can be used for good or for bad. Three types of information disorder can be distinguished: misinformation, disinformation and malinformation. Our research focuses on disinformation that can be spread through social media platforms. We define disinformation as false information that has been distributed that can be wittingly or unwittingly harmful.

When examining the spread of information, we can look at the elements of information transfer and the phases in the information life cycle. The elements in information transfer are agent, message and interpreter. The phases in the life cycle of information are creation, production and distribution.

Disinformation can cause people to believe false information, such as conspiracy theories. It can be especially harmful when ideas get hold on large masses. This is where modern technology, like social media, play a big role in spreading disinformation. We now explain how social media facilitates the distribution of information on a massive scale.

3. Spreading (Dis)information through Social Media

> *Social networks are technologies of entertainment and diffusion. The social reality they create is real, but as a technology of immediacy you can't get no satisfaction. We initially love them for their distraction from the torture of now-time." … "Social networks register a 'refusal of work'. But our net-time, after all, is another kind of labour. Herein lies the perversity of social networks: however radical they may be, they will always be data-mined. They are designed to be exploited. Refusal of work becomes just another form of making a buck that you never see*
>
> ***Ippolita, The Digital Given: 10 Web 2.0 Theses (Lovink & Rossiter, 2009)***

Before the rise of the internet, newspapers, radio and television used to be our main source of news. They were the gatekeepers who decided which stories were published or broadcast to us. They had the public task of providing us with relevant information, made by professional journalists. There was little room for disinformation, which was filtered out by thoroughly checking facts before publishing a story.

The information landscape has changed profoundly. Traditional media brands are still popular as a source of news, but they increasingly face competition from online news sites, blogs, forums and social media (Newman et al., 2018). Social media platforms let people connect and share stories with each other. Whereas previously one talked about the news only with close friends and family, on social media ideas are discussed with anonymous strangers (Swart, 2018).

With social media overtaking the role as the bringers of information it became possible for anyone to tell their story to the world. While this was good for the freedom of expression, it also meant that barriers to spread information had disappeared and a form of quality control was lost. Disinformation could be distributed in the same way as reliable information.

This chapter examines how information spreads through social media by looking at the characteristics of social media platforms. First the technological change is discussed we now call 'Web 2.0'. This change enabled the origination of social media. Next the workings of social media platforms are discussed, as well as the phenomena that resulted from it, such as filter bubbles and echo chambers, which have a great effect on the spread of disinformation.

3.1 Web 2.0

The internet has a long history. In December 1990 Tim Berners-Lee invented the World Wide Web, later simply called the web, and the first internet browser (Berners-Lee, 2010). People with little technical skills could now create and publish web pages. In the early days the world wide web consisted of mainly static web pages with fixed content. Websites had to be maintained and updated manually.

New programming languages paved the way for applications that could dynamically control the content of a website. In the early 2000's new types of web services emerged enabling users to actively create content, whereas previously there were only a few content creators and many passive users. Static web pages were gradually being replaced by websites with dynamically generated content, which could be based on large datasets. This marks the start of web 2.0 (Reilly, 2007).

The power of data became evident, companies were distinguishing themselves by providing added value to data. Fixed banner advertisements could be replaced by targeted advertisements, personal web pages were replaced by blogs. Instead of static websites with fixed hyperlinks the search engines emerged, that indexed all existing web pages. What used to be static data could now be enhanced using user data, as Amazon did with reviews, and Google Maps became a laboratory for competition between application vendors and data suppliers (Reilly, 2007). The benefits of this new development was clear. User experience was enhanced and search engines made it effortless to find information on the vast internet.

But there were also disadvantages. The algorithms of search engines could be abused for spreading disinformation. For terms with few search outcomes the query result could be manipulated, for example by creating new content and promote it by search engine optimization. Microsoft researchers call it a data void (Golebiewski & boyd, 2018). Few search queries had ever been done on 'Sutherland Springs' when a shooting happened in a baptist church in this small village. Because little content on Sutherland Springs was present on the internet the data void was easily filled with messages on Twitter and Reddit associating the shooting in Sutherland Springs with the term 'Antifa', falsely implying the massacre was the work of a left anti-fascist movement, a connection that had soon been taken over even by news media (Golebiewski & boyd, 2018).

3.2 Social Media

With the emergence of web 2.0 the social media platforms originated. Built around user generated content and user connections, they started to capture user data into profiles. With the user profile as the backbone, social media platforms were able to handle interactions between users (Obar & Wildman, 2015).

Social media platforms have empowered online users. Platforms used slogans like 'Making the web more social' (Facebook) and 'Share your pictures, watch the world' (Flickr-Yahoo), argued to be interpreted as coded social behavior (Dijck, 2013). Social media encouraged users to create and share content. By doing so they attracted not only ordinary end users, but companies as well, discovering new marketing possibilities. Algorithms decided what content to show to users, so the social media platforms became more like traditional media encountering the same dilemmas about freedom of speech (Gillespie, 2010). It became apparent that social media platforms were more than just facilitators of interaction. Through their code and algorithms they were shaping user behavior and user experience, which made them mediators rather than intermediaries (Dijck, 2013). The big platforms were able to capture behavioral data effectively into user profiles and were able to monetize it by selling advertisements.

Social media platforms Facebook and Twitter started with the possibility for users to create content and share it with friends or followers. At first people had to copy content from others when they wanted to share it with their friends and followers. This changed with the retweet-button of Twitter and the share-button on Facebook, which made redistributing as easy as just a single click. The effect was that people thought less and shared information impulsively, which had a detrimental effect that was later described by the developer of the retweet-button: “We handed a loaded weapon to 4-year olds” (Kantrowitz, 2019).

With the retweet and share functionality information can spread easily through social media because information that has been posted by one user, can be shared by the friends of that user, and again be shared by their friends, etc. The term ‘viral’ is used for information spreading fast, referring to how viruses can spread from person to person. Marketeers started using tactics for their marketing messages to go viral, but it could also happen unexpectedly to simple messages. A single tweet that was meant to tease some friends had the effect of losing a woman her job and ruining her life when it went viral within hours and became the number one worldwide trend on Twitter (Ronson, 2015).

3.3 Echo Chambers and Filter Bubbles

When people have “a preference for information that is consistent with a hypothesis rather than information which opposes it,” psychologists speak of a confirmation bias (Plous, 1993, p. 233). The tendency can get stronger with emotional involvement. People are looking for information that is in accordance with and confirms their viewpoints. They attract like-minded people and find their ideas being ‘echoed’ by other people. When people are exposed primarily to corresponding information only, their beliefs are being reinforced and we speak of ‘echo chambers’, a term introduced by Cass Sunstein (Sunstein, 2001). The term is not confined to social media, but applies to media in general. Social media platforms, however, have increased the existence and effects of echo chambers because they made it easy to connect with individuals with shared beliefs.

There is another related phenomenon. Web 2.0 companies have started to determine what content to show to each user based on their user profile data. In this way information is personalized and is different for each user. Showing personally relevant information is the platform’s formula to engage users, so they can offer the best products or sell the most advertisements. This strategy was adopted by search engines like Google, by commercial platforms like Amazon, by online streaming services like Netflix and Spotify, and, of course, by social media platforms. Algorithms were now deciding what information or product to show to which user and, even more important for understanding the effect, what to filter out. This personalized information filter has been called the filter bubble by Eli Pariser. According to Pariser filter bubbles are invisible, because all information is still available (Pariser, 2011). Some information just doesn’t show up in a search result or timeline for one user, while it does for others. It is the result of algorithms trying to capture the preference of users without them being aware. “The filter bubble is a centrifugal force, pulling us apart”, he said (Pariser, 2011, p. 10).

The magnitude of the effect of echo chambers and filter bubbles in social media platforms is disputed. One study of US Twitter data confirmed the effect only for political topics, not for other subjects (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015). Because people use combinations of various media, research shows no empirical support for the danger of audiences becoming fragmented, and there is no indication for online audiences to be more fragmented than offline audiences (Fletcher & Nielsen, 2017). Another research shows that people who either are interested in politics or who use several media sources are less likely to be caught in an echo chamber, because they actively search for confirmation of ideas (Dubois & Blank, 2018). A study in Germany also found that people get news from different sources. Filter bubbles were not observed around political influences, only slight reinforcing effects around the far right AfD. The research concluded that the spread of radical ideas cannot be blamed on filter bubbles, it is a social problem and is not to be solved by algorithms or laws (Meineck, 2018).

Axel Bruns summarizes it nicely: “In a hyperconnected yet deeply polarised world, the most important filter remains in our heads, not in our networks: it is the cognitive filter that makes us reject some ideas out of hand, even despite the evidence that supports them, while we cling to others that have long since been disproven and discredited. Not unlike the now thoroughly untenable idea of the ‘filter bubble’ itself, perhaps.” (Bruns, 2019, p. 121)

3.4 Conclusions

News used to be distributed by newspaper, radio, and television. But nowadays news increasingly finds its way to us through social media platforms.

Social media platforms are the result of the emergence of web 2.0, which was a development driven by technology that made web services possible with dynamic content. They were built around user generated content and provided us with the means to easily share content with each other. But they also made it possible for information to go viral by means of the like, share, retweet, and comment functions.

When social media started to collect our data into user profiles, they were able to give us a personalized experience, unique to each of us, based on algorithms deciding which content was to be shown to whom. Because the algorithms might show us only information that matches our preferences and because information might be shown to specific users only, there is a risk of echo chambers and filter bubbles. The results of echo chambers and filter bubbles are still under discussion, but the effects have been observed.

The spread of disinformation can go unnoticed because of these characteristics. Information could circulate among only a limited number of us. We might not be aware of being presented with disinformation. Social media platforms themselves have developed techniques for the detection of undesirable content, such as hate speech or disinformation. After detection they can intervene by removing content, or by preventing content from appearing to other users. There are also measures being proposed by governments. The proposed measures and the efforts social media companies already have come up with are explored in chapter 5.

First we show a conspiracy theory that has flourished on the internet in recent years. This is examined as an example of disinformation on social media.

4. Case Study of Conspiracy Disinformation: Flat Earth

An idea that has prevailed for quite some time and gained enormous popularity in recent years, which defies centuries of scientific evidence, is the theory of flat earth.

The concept of a spherical earth dates back to ancient Greece. The first documented determination of the earth's circumference is by Eratosthenes, who in 240 BC came at a result of 250,000 stades (Evans, 1998). He observed that at noon on summer solstice the sun shone right into a vertical well, while in Alexandria there were small shadows. From the angle of these shadows and the distance between the two cities he was able to deduce the circumference of the earth, see figure 10. This is a classical example of the scientific method, over two thousand years ago, which makes it poignant to see that the idea of flat earth is so popular today.

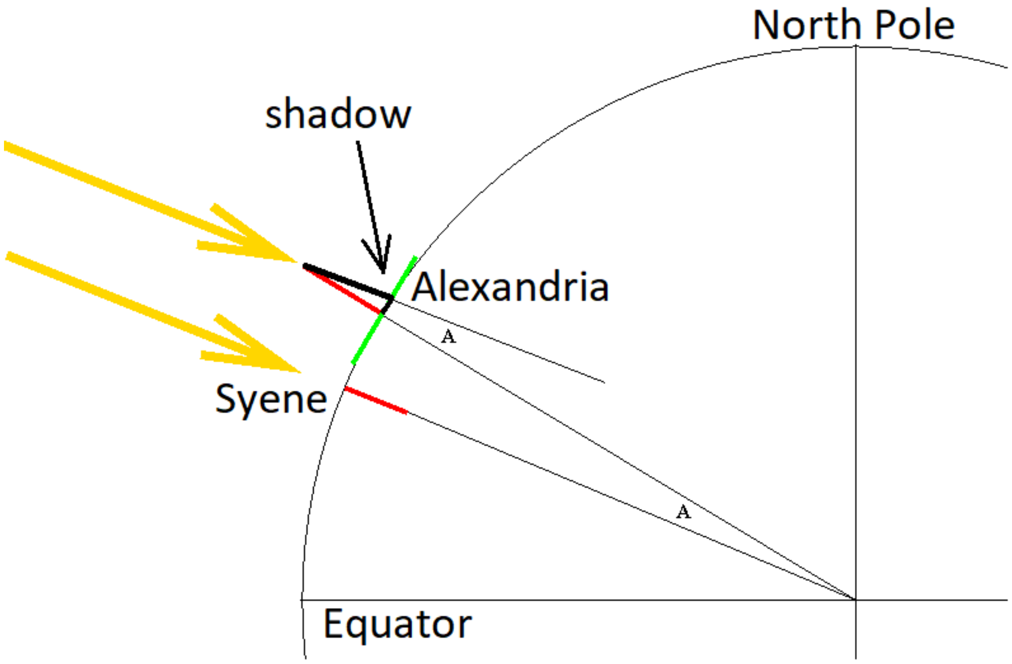


Figure 10: Method of Eratosthenes to Estimate the Earth's Circumference

The modern belief in a flat earth started when Samuel Rowbotham wrote a pamphlet in 1849 and later on published a book in 1865 with several experiments to prove the earth is flat (Rowbotham, 1865). He founded the Zetetic Society to promote his ideas, which was later renamed as the International Flat Earth Research Society, better known as the Flat Earth Society. It increased to 3,500 members in the 1990's, but declined after a fire destroyed the members administration (Martin, 2001), The organization was revived in 2004 when a website was launched containing a forum and a wiki ("the Flat Earth Society," n.d.).

After the origination of social media platforms, the idea of a flat earth gained popularity. It is estimated from polls that millions of people nowadays belief in a flat earth (Ingold, 2018). Flat earthers, as the believers are called, have grown into a strong community organizing several conferences around the world with the annual Flat Earth International Conference attracting hundreds of people (Picheta, 2019).

The flat earthers are conspiracy thinkers who are convinced that all maps and pictures from space are fake and that science, the education system, the government and the financial system are all part of one big conspiracy, a belief that causes a polarization of society into people who trust science and people who are moving away and start questioning everything (Ingold, 2018). This demonstrates that the idea of flat earth is not merely misinformation, but actually harms society.

We place the conspiracy theory of flat earth into the category of disinformation. It contains information that is false and contradicted by scientific evidence. It’s also harmful because of the polarizing effect to society. Some people genuinely believing in flat earth are actively expressing the theory on social media and unconsciously contribute to the distribution of disinformation.

A recent research has shed light on the growth of the flat earth phenomenon. Most flat earthers had been converted by videos with ‘flat earth proofs’ recommended by YouTube after having watched other conspiracy videos (Sample, 2019). “YouTube’s algorithm is spreading information to people who are most susceptible to accepting it.” says Asheley Landrum, who led the research (Hvistendahl, 2019). This demonstrates the presence of the problem of filter bubbles and echo chambers on the platform.

With these facts we can discern the elements of the information transfer. A few people from the flat earth movement have acted as agents by creating videos to prove their viewpoint and spread the word about the earth being flat. Not intended to be harmful or misleading they spread their ideas to reach a large audience. The messages were created as long, engaging videos with convincing arguments and proofs. They were targeted by YouTube to users that were already in to conspiracy theories, so they were misleading messages. The number of flat earthers nowadays proves that many of the interpreters of these videos have been won over by the flat earth viewpoint.

The phases in the life cycle of information are clear. A few people have collected evidence for the earth being flat and created compelling narratives. They produced videos with this material and published them on YouTube. The video platform distributed these messages to people that were already susceptible to conspiracy thinking. In this way the growth of the flat earth movement was fueled and people started to redistribute the word of flat earth.

YouTube has been confirmed to prioritize growth and engagement over the prevention of disinformation, in fact having been turned into an addictive platform favoring disturbing or sensational content, irrespective of its trustworthiness (Bergen, 2019). In January 2019 YouTube announced a change in its recommendation algorithm. Misleading content, such as conspiracy videos, would still be available, but less recommended (YouTube, 2019). Former YouTube developer Guillaume Chaslot confirmed that the original algorithm promoted flat earth videos because of high user engagement and he considered this change of YouTube as a victory (Chaslot, 2019). Indeed the number of search queries on Flat Earth has declined since March 2019, see figure 11.

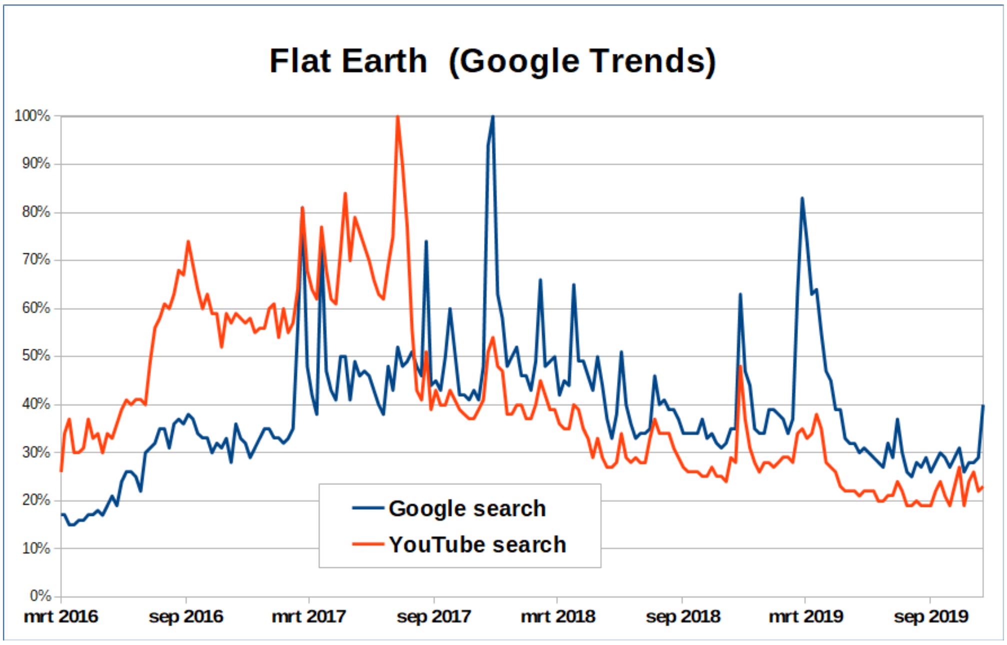


Figure 11: Search Queries for Flat Earth. Source: Google Trends

While this may mean that very few new people will be drawn towards believing the earth not to be a sphere, a lot of people have already been converted into flat earthers. They have turned into a community of like-minded people who find themselves losing their friends and being laughed at by the scientific world (Picheta, 2019).

The decrease in number of search queries for flat-earth proves that intervention by social media platforms can have an effect in reducing the spread of disinformation. We now look into the measures that have been proposed by governments to social media companies as well as measures that have already been taken by social media platforms themselves.

5. Fighting Disinformation by Social Media

We have seen how easily and fast information can spread through social media platforms and reach a large audience. Echo chambers and filter bubbles have been explored as phenomena on social media. We have explained the effect of spreading disinformation to certain users in an almost invisible way. We have also demonstrated how people can be influenced by encountering disinformation repeatedly. In this chapter we examine what governments have proposed to deal with disinformation and how social media companies themselves are already combating disinformation on their platforms. Disinformation is one of the headaches social media platforms face with respect to the behavior of their users and the information they place on their platforms. Detection of disinformation requires fact-checking and actual understanding of content. After detecting disinformation the challenge is to decide if and how to intervene.

Social media platforms have acknowledged their moral obligation for intervention in case of harmful content. Twitter's general counsel Vijaya Gadde admitted in 2015 that Twitter had done too little to protect users from abusive behavior: "Freedom of expression means little as our underlying philosophy if we continue to allow voices to be silenced because they are afraid to speak up. We need to do a better job combating abuse without chilling or silencing speech." (Levy, 2018)

Social media corporations have defined policies to describe what is allowed on the platform. Defining policies is a first step, which is only helpful if a form of enforcement is in place. To detect content violating their policies social media platforms have to use artificial intelligence (AI) systems, because the amount of new content continuously being added simply is too much to check manually. AI can be used for scanning text, photos and videos. For the detection of disinformation the actual understanding of the meaning of content is required in any language. Facebook said in May 2018 that it would take three years of development to be able to deploy effective detection of disinformation by AI (Simonite, 2018).

How hard it is to intervene at the right moment, especially with live videos, became clear when in March 2019 the shootings in Christchurch were streamed live on Facebook by the shooter. The video went viral and copies appeared on Facebook, Reddit, Twitter and YouTube, and it took hours before every copy was removed (Lapowsky, 2019).

Governments have started to propose measures to the problem of disinformation on social media platforms. The European Union has done research on the topic of disinformation and has come up with a Code of Practice for tech platform companies like social media. The set of measures in this Code is the most extensive we've come across and it is very promising. We describe this Code in order to understand the proposed measures. Next to governmental suggestions, social media companies are already taking measures themselves. This chapter examines the efforts of Facebook and Twitter.

5.1 Political Approach to Fighting Disinformation

After 2016 the subject of fake news has started to appear on the political agenda. Governments have become aware of the problem of disinformation on social media influencing public opinion. They are struggling to define their strategy to regulate the tech sector. Many governments "agree that the large tech companies that have come to dominate the online realm, such as Google, Twitter, and Facebook, should be regulated, but caution over-regulation in forms that would curtail expression and press freedoms." (Rogers & Niederer, 2019)

In the European Union the worries on disinformation that could influence the elections of the European Parliament of May 2019 have resulted in research on the subject. In January 2018 the European Commission set up the HLEG, the High Level Expert Group on Fake News and Online Disinformation. The HLEG included members from academia, journalism, online platforms and fact-checking organizations. In March 2018 they published their recommendations, resting on five pillars: (*A Multi-Dimensional Approach to Disinformation. Report of the Independent High Level Group on Fake News and Online Disinformation*, 2018)

- Enhancing transparency of online news
- Promoting media and information literacy
- Empowering users and journalists to tackle disinformation
- Safeguard the diversity and sustainability of the European news media ecosystem
- Promoting continued research on disinformation in Europe

The three organizations Access Now, Civil Liberties Union for Europe, and European Digital Rights have evaluated the HLEG report. In their feedback report they urge policy recommendations to be based on evidence of negative impact of disinformation. Tech companies should be held accountable and only impose restrictions that are necessary and proportionate, and not succumb to governmental or public pressure. They warn that the use of artificial intelligence and other emerging technology must be human-oriented and respect fundamental rights (*Informing the "Disinformation" Debate*, 2018).

In September 2018 the European Commission presented the Code of Practice on Disinformation. Signatories of the Code commit to the verification of advertisers, to publicly disclose political advertisements, and to have policies in place on the misuse of automated bots. Furthermore to improve digital media literacy, users should be empowered by prioritized relevant information and by being presented diverse perspectives. Finally the signatories should encourage research on disinformation and fact-checking (*EU Code of Practice on Disinformation*, 2018). The Code has been signed by Facebook, Google, Mozilla, Twitter in October 2018 and by Microsoft in May 2019. Between January 2019 and May 2019 The European Commission actively monitored the implementation of the Code at Facebook, Google and Twitter with regard to the integrity of the European Parliament elections of May 2019 ("Code of Practice on Disinformation," 2018).

Because social media are an important means of communication for politicians of today, they have drawn attention of politics. Politicians in the US are watching for changes that personally affect them. When a change in the Twitter algorithm in July 2018 negatively affected the names of some Republican politicians in search results, Twitter was wrongly accused of political bias by shadow banning Republicans. This was an improper use of the term shadow banning, a technique used to prevent posts from appearing to other users, effectively silencing a user without them knowing (Stack, 2018). Although Twitter reversed the change within a day, CEO Jack Dorsey later had to testify in Congress that the change had only affected users who were followed by a large number of suspicious accounts (Romano, 2018). In July 2019, after collecting examples of people who suspected social media platforms of suspending their accounts because of a political motive, the White House organized a ‘Social Media Summit’. What was intended as an open discussion on the power of social media, ended, again, in social media platforms being accused of political bias (Rogers, 2019)

But US politics is not only targeting social media companies because of personal discord. After the revelations about Russian meddling in the 2016 US presidential elections the tech giants Facebook, Google and Twitter had to testify in Congress about disinformation on their platforms (McCarthy, 2017). In October 2017 the Honest Ads Act was proposed in the US Senate to get more transparency in political ads on social media, trying to impose existing political ad regulation of TV and radio also on social media companies, but the bill has not yet been passed (Lecher, 2017). In January 2019 a company that had created fake social media posts and comments settled with the state of New York, which demonstrates that distributing disinformation on social media can be prosecuted as an illegal activity (Jones, 2019).

5.2 Facebook

The biggest social media platform shows a transformation towards more transparency and partnering with external parties to deal with disinformation. Events of the past two years illustrate this development.

For some time Facebook had shown users a warning in red to content that was suspected to be fake. In December 2017 Facebook replaced the red flag with ‘Related Articles’ to give users extra context to a suspicious news story, which helped users rethink before sharing and has shown to reduce the number of shares and likes of false news stories (Lyons, 2017). But in the run up to the 2020 US presidential Elections Facebook announced in October 2019 to use fact-checking labels again (Rosen, Harbath, Gleicher, & Leathern, 2019).

In May 2018 Facebook explained the strategy of stopping misinformation, consisting of three parts (Lyons, 2018):

- Remove accounts and content that violate our Community Standards or ad policies
- Reduce the distribution of false news and inauthentic content like clickbait
- Inform people by giving them more context on the posts they see

In September 2018 Facebook explained the actions to prevent election interference on their platform. Next to the automated removal of fake accounts and special effort to detect networks of fake accounts involved in information operations, Facebook is working together with governments, other companies and external fact-checkers, to uncover disinformation before going viral (Zuckerberg, 2018).

In 2019 Facebook changed its strategy into becoming a privacy-focused company. In March 2019 the development of interoperability between Facebook Messenger, Instagram Direct, WhatsApp and SMS with complete encryption was announced with the promise that data would not be stored anymore “in countries with weak records on human rights like privacy and freedom of expression.” (Newton, 2019) In April 2019 Facebook announced a redesign of its app to prioritize groups and events, which are not always public, over public posts in a news feed (Statt, 2019). The focus on privacy shows a concern for the user’s safety on the platform, which is a positive development. The announcement of end-to-end encryption, however, raised concern from law enforcement officials, fearing that it would “prevent law enforcement agencies from finding illegal activity conducted through Facebook, including child sexual exploitation, terrorism, and election meddling.” (Mac & Bernstein, 2019)

Next to their own efforts Facebook is asking help from academia. Because they are the owner of large datasets that are not publicly available for external researchers, Facebook is teaming together with academics. In April 2018 a new initiative was launched to enable independent research about the role of social media in elections and democracy in general. Researchers would be able to use privacy-protected datasets from Facebook (Schrage & Ginsberg, 2018). In September 2019 the Deep Fake Detection Challenge was announced, in which Facebook would be making datasets with both genuine footage by real actors and deepfake videos derived from this footage, in order to enable researchers to develop better detection methods of deepfake videos (Schroepfer, 2019).

To become transparent, Facebook started in May 2018 with the quarterly publication of the Community Standards Enforcement Report. It is reported that billions of fake accounts are removed each quarter, most of them within minutes of registration. Over 99% is detected by Facebook itself, the rest is reported by users. Action is being taken against billions pieces of content per quarter, around 99,8% detected by Facebook itself. In the period January – March 2019 there were appeals for 20,8 million pieces of content that had been removed and 44,3 million pieces of content were restored (“Community Standards Enforcement Report,” 2019).

The Community Standards Enforcement Reports are reviewed by the independent Data Transparancy Advisory Group (DTAG), consisting of academic experts in metrics. In their report of April 2019 they conclude that “Facebook’s system for enforcing the Community Standards, and its methods of auditing the accuracy of that system, seem well designed.” (Bredford et al., 2019) Several recommendations are made to Facebook to become even more transparent on their methods, which is a step in the desirable direction.

5.3 Twitter

Twitter, once started as a microblogging service with little messages called tweets, later on enabled the use of photo and video in tweets, so it became a major tool for spreading all kinds of information. Twitter is open to anybody and popular to “many politicians, celebrities, journalists, tech types, conference goers, and experts working on fast-moving topics.” (Rid, 2017) There is no limit to the amount of accounts a user can create, which makes it easy to abuse the service. Bad actors can create accounts automatically, called bots, and use them to spread content. Bots generally post more content in tweets than humans do, place more URLs, and retweet more often than humans do (Gilani, Farahbakhsh, Tyson, & Crowcroft, 2019), so they can easily be used to spread information.

Twitter, too, has to deal with the issue of disinformation and is showing efforts of becoming more transparent. Twitter tried to remain as neutral as possible for a long time. In 2011 the company stated: “There are Tweets that we do remove, such as illegal Tweets and spam. However, we make efforts to keep these exceptions narrow so they may serve to prove a broader and more important rule—we strive not to remove Tweets on the basis of their content.” (Stone, 2011) Twitter even published how the company had been requested by the FBI for information on accounts related to Wikileaks (Cohen, 2011). However, there were exceptions to their neutrality, such as filtering out abusive language directed at US president Obama (Warzel, 2016). And of course, being the president of the US has its privileges. When president Trump expressed a threat to North Korea in a tweet, it was not removed. It was not considered a policy violation, instead it was labeled as newsworthy by the platform (Romano, 2017).

In June 2017 Twitter announced to expand resources to detect “spammy behaviours at source, such as the mass distribution of Tweets or attempts to manipulate trending topics.” (Crowel, 2017) Little is said about the methods used for this detection. The only explicit statement is that details of detecting abuse of the Twitter API could not be disclosed because it could help people circumventing the measures (Crowel, 2017).

In March 2018 Twitter introduced new measures to improve what it called healthy conversation. Twitter said that troll-like behavior would result in posts being less visible to other users, but admitted that mistakes would be made and promised to be open and honest about it (Harvey & Gasca, 2018). Twitter invited other companies to submit proposals for measuring the health of conversation on the platform (“Twitter Health Metrics Proposal Submission,” 2018).

In October 2018 Twitter took a new step and released all accounts and corresponding tweets that were believed to be state information operations. Because the accounts had already been deleted after detection, the content was not publicly available anymore. With this move Twitter intended to accomplish a level of transparency and “enable independent academic research and investigation.” (Gadde & Roth, 2018) The datasets Twitter published consisted of accounts attributed to the Russian Internet Research Agency (IRA) and accounts from Iran. It was supplemented in January 2019 and again in June 2019 with newly discovered material of accounts from Russia, Iran, Venezuela, Bangladesh and Catalonia (“Elections Integrity,” 2018). Twitter states that it only discloses “datasets associated with coordinated malicious activity” that they are “able to reliably associate with state-affiliated actors”. (Roth, 2019)

The data set of IRA related accounts helped researchers get an idea of the tactics the IRA had used and understand “how IRA operatives used accounts mimicking communication within differing ideological networks, and likely differing echo chambers.” (Linvill, Boatwright, Grant, & Warren, 2019, p. 298) While this might be invaluable, Twitter hasn’t revealed how suspicious accounts are detected, how an account is labeled as state-sponsored, and which countries are focused upon, so it’s difficult to judge how effective Twitter is at removing disinformation (Harrison, 2019). To detect information campaigns Twitter partners with governments, law enforcement and peer companies and bases analysis on thousands of signals and behaviors (Roth, 2019).

In August 2019 Twitter tightened the advertising policies and did no longer accept advertisements from state-controlled news media entities to ensure media freedom without political pressure (“Updating our Advertising Policies on State Media,” 2019). In October 2019 Twitter banned political advertising altogether to prevent people from buying influence (Dorsey, 2019), a rather drastic move which goes a long way further than just providing transparency on political ads.

In 2018 the Platform Manipulation report became one of the components of the Transparency Report, that had first been published in 2012. It discloses the number of spam reports per month and the number of accounts challenged per month. By the challenge of an account Twitter means the verification of the authenticity (“Platform Manipulation,” 2018).

Although we see the efforts of Twitter paying off, there is no information about partnering with academia, nor is there any validation of the transparency report. We therefore conclude that there are serious steps yet to be taken by Twitter.

5.4 Conclusions

Social media companies have attracted attention from politics. Not only because politicians themselves are frequent users of social media, but also for political reasons. Due to revelations about Russian interference with the 2016 US presidential elections governments worldwide are now proposing actions to be taken to combat disinformation.

Meanwhile social media platforms have defined their own policies to determine what user behavior is allowed. Facebook and Twitter have taken steps to fight disinformation. Combinations of human activity and automated measures have been taken. Reports and data are published to give a level of transparency. Both platforms have their own approach and employ their own strategy. They are partnering with governments and private companies to detect false information. While the first steps have been taken towards more transparency, there is still a long way to go to make sure the world really understands what social media are actually doing in their fight against disinformation.

The question is now whether governmental proposals and the approach of social media companies themselves have been given enough consideration and are not applied blindly. This is examined in the next chapter.

6. Considerations When Fighting Disinformation

> *Falsehood flies, and truth comes limping after it, so that when men come to be undeceived, it is too late; the jest is over, and the tale hath had its effect: like a man, who hath thought of a good repartee when the discourse is changed, or the company parted; or like a physician, who hath found out an infallible medicine, after the patient is dead.*
>
> ***Jonathan Swift****, The Examiner No. XIV*

We have looked into the measures proposed by governments and we have seen the efforts social media companies have taken already of fighting disinformation. We examine whether all measures have been given enough consideration. This is done by looking at the limitations of the measures and techniques.

The main conclusion about the efforts of Facebook and Twitter of last chapter was they are becoming more transparent. We now ask ourselves to what degree transparency really helps understand the effort social media companies undertake and the effectiveness of these efforts.

For combating disinformation it is necessary to be confident about the truthfulness of information. Social media companies are partnering with academia and fact-checking organizations. Because the results of these fact-checking activities determine whether information should be labeled as fake, the risk for affecting the right to freedom of speech is significant when information is wrongly labeled as being fake, but also if disinformation goes undetected. For this reason it is important to know if fact-checking comes with limitations.

All methods of intervention rely on a high degree of automation due to the sheer volume of content being produced on social media platforms. As we have seen in chapter 3, the algorithms of social media technology make the decisions on what content to show to us, and they can be blamed, to a certain level, for the origination of filter bubbles and echo chambers. Because interventions are mostly automated, it is important to take a look at potential limitations of algorithms.

Finally we dive into the motivations of the users as interpreters of information. The harm of disinformation depends on the effect at the receiving end. The conscious user, who is aware of the truthfulness of information, would think twice before redistributing disinformation. Therefore one of the proposed measures is to increase the media and information literacy of users. The effect of this literacy is only helpful when the user is fully active in a critical thinking mode. We explore the limitations of the user’s media and information literacy.

6.1 Limitations of Transparency

The rationale behind transparency is that if the inner working of a system is known, the system and its makers can be held accountable. Transparency gives a sense of control and security by merely observing. There are limitations, however. It could be harmful, or it could reduce trust when corporations give away their secrets. Complete transparency might not be possible, and making things visible doesn't necessarily make them understood (Ananny & Crawford, 2018). Helen Nissenbaum calls it the transparency paradox. Being entirely transparent about an algorithm does not mean that ordinary users would be able to grasp the workings and be able to make meaningful choices, whereas just summarizing the essence of the algorithm would leave out too much detail to be helpful (Nissenbaum, 2011).

In order to understand how ethical standards can be applied to algorithms Mike Ananny defines a networked information algorithm as the combination of computational code, human practices and normative logic, that creates socio-technical relationships among people and data (Ananny, 2016). According to Ananny it only makes sense to judge an entire network of people and data when approaching from three angles together: the technical inner workings, the effects and consequences, and the underlying values and ethics, which means that real insight goes beyond just the transparency of code (Ananny, 2016). To be able to judge accountability of a socio-technical system we should ask "what is being looked at, what good comes from seeing it, and what are we *not* able to see?" (Ananny & Crawford, 2018, p. 985)

6.2 Limitations of Fact-Checking

Around the world fact-checking organizations and websites are checking news stories. The Reporter's Lab, a center for journalism research at Duke University maintains a list of fact-checking organizations ("Fact-Checking - Duke Reporter's Lab," n.d.). In March 2017 the International Fact-Checking Network (IFCN) introduced a code of principles. Organizations with the IFCN badge have been verified to adhere to a minimum standard for fact-checkers, consisting of commitment to non-partisanship and fairness, transparency of sources, funding, organization, and methodology, and commitment to open and honest corrections. Facebook chooses fact-checking partners that are signatories of the IFCN code of principles ("IFCN Code of Principles," n.d.).

Fact-checking organizations are independent from political or commercial interest and try to be as transparent as possible. They regard themselves as an extension to traditional journalism as well as the correcting factor of it (Singer, 2019). But fact-checking does not prevent disinformation from appearing or from spreading. Creators of fake stories will launch new websites or create new social media accounts if their existing sites or accounts are removed. Debunking disinformation by fact-checking isn't effective for people who have lost their trust in traditional media and view them as the bearers of fake news (Borel, 2017a). The fight against disinformation needs more than telling the truth.

When the Dutch newspaper De Telegraaf published an interview with a scientist with a critical attitude towards climate change, they received an official warning from Facebook for the distribution of disinformation. Apparently the viewpoint was too controversial for the fact-checkers (Jansen, 2019), which shows the power of fact-checking in deciding what information is allowed.

Whereas it was believed that technology was bridging differences between people, we now see new technology being used in ways that actually sow discord. But it's not only technology that is to blame, we need to acknowledge the flaws of the larger system which it is part of. We need social, political, economic and technical structures to understand each other and be able to bridge different viewpoints. The discussion about disinformation forces us to understand how society is constructed through communication of ideas, it is a socio-technical problem that cannot be solved by technology alone (boyd, 2017b).

Countering disinformation on social media with arguments that support the opposite view isn't enough. It fuels a fight and not a dialogue. As Brooke Borel explains, we need to start healthy conversations with our friends and family. Especially on social media, which is public space, we should treat other people as friends or family. People share disinformation because it resonates with their identity. Debunking the information they share means they have to change their identity, which is threatening. People's ideas and habits are hard to change, but that should not withhold us from trying (Borel, 2017b).

Focusing on debunking false information isn't always enough, disinformation campaigns might not spread just untrue stories. As Facebook has discovered, most content of manipulation campaigns isn't provably false, and would be acceptable if shared by authentic sources. The problem is that the identity of the creators is concealed and their activity appears to be trustworthy or popular. To deal with new tactics, the focus need to be on inauthentic behavior, not only the content of messages (Gleicher, 2019).

6.3 Limitations of Algorithms

The sheer volume of content makes it an enormous challenge to enforce policy rules. Detecting and judging whether a post violates platform policies cannot be done just by human moderators alone. Social media platforms need to bring automatic detection into the equation. The advantage of automatic systems appears to be a fair judging system without human bias. The reality, however, is that automated tools and algorithms are designed by humans and trained on data that has been previously labeled by humans. This training data could have limited applicability or could contain bias. Computers detect patterns, but they aren't able to understand a message in its context. Social media platforms still need human involvement for the understanding of hidden meanings and for refining the detection algorithms (Lohr, 2013).

Another caveat of algorithms is that it only detects known patterns, such as certain words in a text or elements of an image. But words can have several meanings, as do images. The context needs to be taken into account to determine whether the use of a word is inappropriate. Words for intimate body parts can be used in an insulting way, but also in education or science. Furthermore the algorithm is trained on existing data, so it cannot detect new phenomena, like new words, new forms of expression, or new threats (Gillespie, 2018).

The problem of bias in an algorithm can be persistent. For years the LGBTQ community has had complaints about their videos being demonetized (which means not being allowed advertisements) and videos being recommended less. YouTube CEO Susan Wojcicki has had to apologize several times emphasizing the fairness of their algorithms (Alexander, 2019a). A recent investigation hints at the YouTube algorithms judging videos as inappropriate for monetization based on the words 'gay' and 'lesbian', a claim that is denied by the platform itself, declaring there is no list of words, only the machine learning systems (Alexander, 2019b).

Facebook announced in June 2018 that it would use machine learning to detect duplicates of stories that had already been debunked. Machine learning would also be used to detect repeat offenders of disinformation. Facebook said the effort would never be finished. Facebook didn't reveal how fake stories were being detected, what algorithms were being used, on what data it had been trained, and whether training data might contain biases, so the effectiveness of Facebook's efforts isn't clear (Shaban, 2018).

New sophisticated methods will arise to evade the ever-improving detection mechanisms of social media platforms. Peter Singer and Emerson Brooking concluded in their book LikeWar: "Within a decade, Facebook, Google, Twitter, and every other internet company of scale will use neural networks to police their platforms. Dirty pictures, state-sponsored botnets, terrorist propaganda, and sophisticated disinformation campaigns will be hunted by machine intelligences that dwarf any now in existence. But they will be battled by other machine intelligences that seek to obfuscate and evade, disorient and mislead. And caught in the middle will be us—all of us—part of a conflict that we definitely started but whose dynamics we will soon scarcely understand." (Singer & Brooking, 2018, p. 259)

Lastly, a recommendation is to be considered on the use of automation. Because social media platforms encode all of their choices into their algorithms, it is important to be able to check the algorithms. An 'algorithm review board' has already been proposed to oversee the big tech companies (Grind, Schechner, McMillan, & West, 2019).

6.4 Limitations of Media and Information Literacy

One of the recommendations is to educate young people as well as adults, at schools and in society in general. For media literacy to be effective, however, we should also look at the motivations behind user behavior. With the use of social media platforms there are forces at work that influence which content users are actively consuming, and that have an effect on their ideas and on their sharing, retweeting, liking, and commenting behavior.

Investigations have found that social media platforms can have a radicalizing effect. When viewing political videos on YouTube, the recommendation algorithm starts recommending more extreme content and before long, the user is presented with extreme-right or extreme-left content (Tufekci, 2018). The effect is observable not only with political content. Start with vegetarianism videos and you will get videos about veganism. Jogging videos lead to videos about running ultra-marathons. The recommendation algorithm seems to be biased towards extreme or provocative content. The explanation is that people are inclined to click more on sensational titles in search of uncovering new secrets and ‘deeper truths’ (Tufekci, 2018). The Wall Street Journal found that “YouTube's recommendations often lead users to channels that feature conspiracy theories, partisan viewpoints and misleading videos, even when those users haven't shown interest in such content. When users show a political bias in what they choose to view, YouTube typically recommends videos that echo those biases, often with more-extreme viewpoints.” (Nicas, 2018)

The effect popular people at social media platforms have on other users can be explained by the theory of parasocial relationships (Chung & Cho, 2017). When a user is encountering someone on social media repeatedly and starts to follow that person, an emotional connection is perceived by the user. Unlike the relation with media celebrities on radio or television, within social media the user has the ability to interact directly. When the social media person reacts on comments or tweets from a user, which is easily done on social media, the user perceives it as genuine interaction and starts to feel personal emotional connections. Engagement encourages intimate and affectionate relationships with users, turning them into loyal followers (Kang, 2014). Especially with celebrities the possibility of getting a direct message, even though the chances are very slim, creates a sense of intimacy and reciprocity, so social media platforms foster parasocial relationships between users and celebrities (Chung & Cho, 2017).

Not only for the viewers there is something to gain from parasocial relations. Creators are forming their digital selves through their videos and the confidence in their digital self is reinforced by the number of views of their videos and the reactions from viewers. They start to feel being famous because of the attention, and want to cultivate their digital identities and present themselves and their ideas (Chen, 2016). Parasocial relations cause people to feel connected to the media creators and trust them. This builds authoritative figures from popular creators, due to their large following, causing the algorithms of social media platforms to favor their content.

Disinformation on social media is hardly distinguishable from reliable content. Posts from our friends are alternated with news messages, advertisements and content the algorithms think we like. We lose the power to think critically and perceive everything equally. But not everything gets equal attention. Content that provokes a strong emotional reaction, has a greater chance of being shared. The design flaw in the business model of social media is that the content and users with the most clicks, likes, or views are most rewarded. Because online discussions are less personal, people are easily attracted to gossip and sensation if they are concerned about their digital identity and the approval of their audience (Haidt & Rose-Stockwell, 2019). Extreme, sensational or polarizing content provokes the strongest emotional reactions in us. Together with the algorithms that serve content that we find most engaging, social media platforms are the perfect environment in which internet hoaxes, conspiracy theories and other disinformation thrive exuberantly.

Trust is undermined and calm, productive conversations between diverse groups are becoming less common (Vaidhyanathan, 2018).

There is research on how users are able to discern fake from true on social media. Encountering disinformation may lead to justify prior beliefs. People may be convincing themselves fake information is accurate when it corresponds with their ideas. However, analytical thinking has be shown to be associated to the rejection of fake stories. The problem of people believing and sharing disinformation appears to be credited to a failure to think at all, social media lead to a tendency to rely on intuition rather than on reasoning (Pennycook & Rand, 2019b). People don’t evaluate new information based on rationality, they fit it in their existing worldviews. If a story feels like a coherent narrative, it will be accepted. This has nothing to do with accuracy of facts. To convince people and steer them away from disinformation they need to hear a different coherent story they can comprehend (Martineau, 2019).

People are passive consumers of information. When they become active, their determination of media truth turns out to be rather good. Experiments have shown that trust ratings of headlines by laypeople match those of professional fact-checkers closely, so incorporating ratings by ordinary users in the ranking algorithms of social media platforms could lead to less distribution of disinformation (Pennycook & Rand, 2019a). This kind of experiments are promising, new ideas that could help turning users into critical thinkers.

The problem is our attitude to truth. To effectively counter disinformation not only rationality about verifiable facts is needed in the argument, but also emotional engagement. To tell the truth we must reach head as well as heart, charismatic scientists and celebrities can help get attention and bring the argument back. (D’Ancona, 2017). Our relationship with information is based on emotion, so disinformation has a huge impact when it taps into our deepest fears. When it does so, simple narratives, conspiracies and hateful speech become far more effective (Wardle, 2019).

6.5 Conclusions

There is consensus among policymakers, academics and tech companies on the need to combat disinformation. But intervention poses a risk to the freedom of expression when people are deprived of the opportunity of speaking out divergent opinions.

Governments worldwide, as well as platform owners, are struggling with this dilemma. The measures they have come up with should be applied with consideration of limitations. In this chapter four measures and techniques have been considered and limitations have been indicated.

- Transparency is a useful measure insofar the public understands what is done. Explanations can be either too general or too technical.
- Fact-checking is a useful technique to expose false information, but it isn’t enough to convince everyone. Furthermore there are disinformation campaigns that don’t

use false information and won’t be debunked by fact-checking. It also provides us with the dilemma of deciding which stories are to be checked.

- Fighting disinformation cannot be done by humans alone because of the sheer amount of information. Automatic detection however, has its flaws. Algorithms may contain hidden bias and can do unintended harm. Besides adversaries will evolve and use their own algorithms to find ways of evading detection mechanisms.
- Finally the improvement of media and information literacy helps users debunk disinformation, but it requires an active role for users. We need to reckon with the behavior of users in terms of trust, and take into account the effects of parasocial relations. There is a need for users to acquire a more active way of consuming information, thereby triggering critical thinking and common sense.

With all the actions being taken against disinformation we should be wary of having implemented a new kind of censorship. Before being able to return to the main question we first explain the concept of internet filtering.

7. Internet Filtering

> *On the one hand information wants to be expensive, because it's so valuable. The right information in the right place just changes your life. On the other hand, information wants to be free, because the cost of getting it out is getting lower and lower all the time.*
>
> ***Stewart Brand, Hackers' Conference 1984***

The internet has been designed without central governance authority. With its architecture of interconnected decentralized networks the internet has been a voluntary collaboration project that could not be governed centrally for a long time. According to the pioneers of the internet information being spread on the internet should not be controlled by governments. "Information wants to be free," Stewart Brand said (Brand, 1984), meaning that technology should be liberating instead of oppressive. Free access to information for all internet users worldwide became a major force for freedom, facilitating dissidents and international movements to speak out. The internet was thought to be uncontrollable.

However, several countries have started to implement new regulations, legally as well as technically, to control access of their citizens to information. This is described by the term internet filtering. This chapter shows how countries use various techniques to prevent people from getting their ideas out to other people. It can be done by preventing access to information that is out there, but there are other, more subtle, forms of internet filtering.

The sheer volume of information available to today's users creates a limited attention, as economist Herbert Simon already anticipated in the 1970s that an overabundance of information is creating a scarcity of attention (Simon, 1971). This can be applied as a censoring technique. Overwhelm the user with information from a certain viewpoint and the dissident's voice is not heard, effectively silencing anyone with opposing viewpoints.

Even more subtle, and perhaps more disturbing because of its opaque nature, is the control by social media platforms by means of their algorithms, that decide what information each individual user gets to see and what information gets filtered out. In their combat against disinformation social media platforms are exercising their power to filter information. In doing so there is a danger of limiting freedom of expression. This freedom is not only at risk by removing content or preventing content from showing up to users. By not intervening and letting disinformation or offensive content exist, the risk is just as real that someone's opinions are not heard. Preserving real freedom of speech requires finding the balance between letting people express themselves and intervening effectively when there is a negative impact for users.

7.1 Definition of Filtering and Blocking

The technical measures used for censoring and controlling access to information are called blocking and content filtering. Blocking refers to the denying of access to certain domains, IP addresses or port numbers, content filtering refers to the blocking of access to information based on its content, for example certain keywords (Deibert & Villeneuve, 2005).

Internet filtering is a common practice in many countries. Motives for filtering can be political, social or security-related. Websites and content containing information that fall into affected categories are blocked, but governments can also block the tools used to share information about websites, e.g. blogs, anonymizers or social media. (Faris & Villeneuve, 2007). In order to have control over the internet some countries have started to make sure data is processed locally, a process called the ‘balkanization’ of the internet. When this has been done there are several options to take control. ISPs can be taken over to make sure certain content is not accessible, or to monitor user activity and track down citizens who publish undesirable content.

Another, rather harsh way of control is the use of blackouts. During the revolts in Syria the main ISP was forced to switch off the internet on Fridays, the day people gathered to go to mosques and protest, and in Algeria the internet was brought down for three days during high school exams (Singer & Brooking, 2018). A nationwide blackout is costly and damaging to the economy, so governments are trying to intervene more efficiently, for example Bahrain set up an internet curfew only in a few rioting villages (Singer & Brooking, 2018).

Another method is to launch a counterattack with an information campaign. Authoritarian governments have been developing cyberspace dominance through network attacks as well as psychological information operations with the intent “to silence information that is strategically threatening and sow confusion and doubt among opponents dependent on cyberspace for information and organization.” (Deibert & Rohozinski, 2010)

People in authoritarian countries have experienced the same absence of freedom of expression online as offline. A Russian woman was sentenced 320 hours of hard labor after posting negatively about the invasion of Ukraine, and the first country to sentence someone to death for online speech was Pakistan (Singer & Brooking, 2018). When governments are actively tracking down dissidents, because they follow any social media post from their citizens, a form of filtering called ‘self-censorship’ starts to appear, sometimes labeled as the ‘spiral of silence’ (Singer & Brooking, 2018). In that case people don’t dare to express anything that could provoke the ire of authority.

Next to national filtering for political motives several institutions are exercising another form of filtering, for example public libraries, internet cafes and organizations like businesses and schools. This can be attributed to organizational policies.

The activity of internet filtering can have a reverse effect. When the Chinese government blocked Instagram in 2014, other sites like Twitter, Facebook and Wikipedia, gained users, who became more politically engaged as an unintended consequence (Hobbs & Roberts, 2018).

A technical system used for blocking or filtering content has its flaws. There is the risk of blocking either too much or too less, called overblocking and underblocking respectively. Furthermore users with technical skills try to circumvent the controls. As a reaction the filtering systems need to become more sophisticated (Zittrain & Palfrey, 2007). People are very ingenious when it comes to circumventing any filtering control. When AOL and other blog services blocked the references to “anorexia” people started referring cryptically to their friend “Ana” to be able to write about anorexia without triggering the filters (boyd, 2017b).

The problem with technology for filtering is that it can easily be used or abused for means other than originally intended. When an internet censorship law was passed in Germany in 2009 to protect children, concerns were raised about the lack of public oversight to the police maintaining the list of blocked websites. When suggestions were proposed by politicians to add video game sites, gambling sites and Islamic sites to the blocking list, the debate on the possible abuse of this technology eventually resulted in overturning the law in 2011 (MacKinnon, 2012).

The various appearances of internet filtering do have some common features. First, internet filtering is automatic and self-enforcing, without human intervention. This means there is no feedback loop, normally associated with regulation and enforcement. Second, it is often opaque. Users and website owners might not be aware that a page isn’t loading because of filtering. The reasons of blocking and the entire list of blacklisted content may not be known to the users. Thirdly filtering is mostly done by intermediaries who supply technology used for filtering. They might not be transparent on their techniques and their effectiveness (McIntyre & Scott, 2008).

The OpenNet Initiative distinguishes different generations of control, which shows the ongoing development of internet filtering. First-generation controls focus on blocking access to certain domains, keywords, servers or IP addresses. Second-generation controls create legal and technical possibilities to take down or filter websites by labeling content as illegal, or slander or untrue. Content can also be temporarily blocked during specific times, such as periods of political unrest or election periods. In third-generation controls the focus is not on filtering or blocking access to information, but to interfering. Instead of denying access, a psosible tactic is to compete through counter-information campaigns. (Deibert & Rohozinski, 2010).

7.2 Internet Censorship through Social Media

With the emergence of new technologies the techniques for internet filtering are being developed further and are becoming more sophisticated. While the Arab Spring showed the world in 2011 how the internet and social media had provided citizens the power to assemble and successfully demand change, a mere four years later the affected regimes had adapted to the new technology and now used it to track down dissidents. Authoritarian regimes “have developed an arsenal that extends from technical measures, laws, policies, and regulations, to more covert and offensive techniques such as targeted malware attacks and campaigns to co-opt social media.” (Deibert, 2015)

The big tech companies are adopting to local legislation. After being excluded from China Google was building a new search engine for the Chinese market meeting the strict Chinese censorship laws and restricting access to content not allowed by the Chinese (Gallagher, 2018). After the revelation of the secret project several employees were indignant about Google secretly supporting a regime violating human rights. In July 2019 Google confirmed to the US Senate that the project had been abandoned (Alba, 2019). Rumors about Facebook exploring the technological possibilities of complying with Chinese laws have not been confirmed (Lee, 2016).

People rely on technology companies to take decisions that are in their user’s interest. It can mean that the safety of citizens is at the mercy of social media platforms and their policies. When Facebook changed its privacy settings in 2009, disabling the possibility to hide one’s list of friends, it resulted in Iranian users deleting their Facebook account, afraid their list of friends might link them to protesters. Globally there was so much indignation on the way Facebook had changed its privacy settings overnight without warning, that Facebook had to change its privacy settings to meet user demands (MacKinnon, 2012). It was perceived as a violation of trust by many people.

Social media were heralded as the bringers of freedom helping to fight against totalitarian regimes, but nowadays they are being used for totalitarian practices. Because social media platforms are built on trust, intimacy and sharing, they are an easy tool for spying on citizens and also for spreading disinformation. Users are overwhelmed with information, including conspiracy theories, leaks and other misinformation and cannot determine objective truth anymore and start to question the integrity of all media (Deibert, 2019).

7.3 Internet Censorship by Social Media

As platforms in which people are able to express themselves to a large audience of friends and followers, social media can’t escape from imposing some rules. It’s necessary to protect users from each other and intervene by removing offensive or illegal content, for example harassment, pornography or copyrighted material. This protection is in the interest of users, advertisers, and society in general because most content is public. This forces social media into the role of arbiter by defining norms, interpreting laws, and judging content according to their standards. Social media platforms have become the custodians of the internet (Gillespie, 2018).

Platforms have long stressed their neutral position, declaring that they are merely intermediaries in distributing content. Popular content likely ends up high in the ranking by algorithms, regardless of a possible offensive nature of the content. By this declaration they were avoiding obligations to their customers and liability for the content that was hosted on their platforms. In the design of their algorithms however, a bias might exist, and certain values might have been implicitly encoded. Therefore they need to acknowledge their role as moderator of public forums and respect user rights into their technology (Pasquale, 2016).

Censorship can be exercised by social media platforms at different levels. Content can be removed or a user account can be suspended, which is a harsh all-or-nothing form of filtering. Content can also be preserved, but hidden from users who might be offended, for example underage users. This practice is no different from how adult content used to be dealt with, putting the adult movies at the back of the video store, or scheduling the adult content on TV after children’s bedtime. In this case the censorship is known and accepted by the user. Social media platforms can also censor in a more subtle way by preventing content from showing up in search queries or by only appearing to some users, not based on their preferences, but on their user profile. If users aren’t aware of the choices made by the algorithms, they don’t know what information they are missing, which means the social media platforms are the ones to decide how important each bit of information is to each individual user (Gillespie, 2018). This form of internet filtering through the use of algorithms by social media has also been called social filtering (Hanani, Shapira, & Shoval, 2001).

The question is who is setting the criteria for defining what content is allowed. Social media companies are acting from their own values, assuming they are able to decide for all of their users. But the workforce of these companies is no cross section of society. Employees are mostly white, male, young, and technically skilled (Wiener, 2016). How can they be able to appreciate the values of minorities or cultures in other countries?

7.4 Conclusions

From the early days of the internet countries have started to deny access to certain categories of content that they deemed undesirable. The techniques were very rough at first, blocking entire IP-addresses or websites, or even shutting down the entire internet. Techniques for internet filtering are becoming more sophisticated and subtle. The long arm of state regulation has already appeared on social media too.

Next to state regulation institutional internet filtering is applied by institutions, such as public libraries, schools, and corporations. On social media too, some form of filtering has been introduced. Because of their social nature social media companies have started to define their own policies of behavior on their platforms. In this way they are their own regulators and decide for their users what content is allowed. Through their algorithms information gets filtered out on an individual level, which makes it a very subtle and opaque form of internet filtering.

8. Conclusion & Discussion

Social media platforms nowadays play a major role in the way we communicate and spread ideas. They have provided the masses with the means to easily express ideas to the world. The success of these platforms is shown by the number of active users, which run in hundreds of millions to even billions for the biggest platforms. Providing access to this amount of users also means that people with bad intentions are attracted, who now have the perfect tool for spreading disinformation on a massive scale.

After the discovery of the Russian disinformation campaign during the 2016 US presidential elections the need for intervention has been expressed by politicians. Social media platforms have started efforts to counter disinformation. But with the power of intervention comes the responsibility of respecting the right to freedom of expression and right to information. When action is taken blindly, countering disinformation is becoming a new form of censorship. The aim of this research was to make clear when fighting disinformation can turn into a form of internet filtering.

This study has looked at the definition of disinformation and at the mechanisms of the distribution of information through social media. It has considered the efforts undertaken by social media platforms to combat disinformation. Current practices and proposals to fight disinformation have been considered. The limitations and challenges of these recommendations have been indicated. Finally the definition and practices of internet filtering in general and by social media platforms in particular have been examined.

In this chapter the results of the research are summarized and the main research question is answered. Limitations of the research are considered as well as future research suggestions.

8.1 Main Findings

This research started with the question:

> **When does fighting disinformation by social media platforms become a form of internet filtering?**

To answer this question five subquestions have been asked. The main findings of each of the subquestions are summarized.

What is disinformation?

Several definitions have been examined and criticized. For this research we define disinformation as false information that has been distributed that can be wittingly or unwittingly harmful. To understand the workings of disinformation we have adopted the framework of Wardle and Derakhshan, describing the elements of communication and the phases in the life of information. Information is transferred by communication, in which the three elements agent, message and interpreter play a role. The phases in the life cycle of information that can be distinguished are creation, production and distribution.

How does (dis)information spread via social media?

Several characteristics of social media technology play a role in the distribution of information. Social media emanate from web 2.0, a technology-driven change in the nature of web services in the early 2000s. Static web pages were gradually replaced by services with dynamic, personalized content. Social media have been built upon user behavior data, captured into profiles. Content on social media is personalized to each individual user based on their user profile. By sharing, liking, retweeting, and commenting, content can be redistributed to friends, who can subsequently redistribute it again. This can cause information to go viral to massive amount of users.

By use of their algorithms social media platforms have great power over what each user will see. Each individual user is being presented content chosen by the algorithms and is getting a unique experience, invisible to other users. This phenomenon is called the filter bubble. When a user is repeatedly confronted with information that confirms a certain bias, we speak of an echo chamber. The effect of echo chambers is reinforced by social media algorithms, that are showing content to the user in accordance with their user profile. This means people can be drawn into believing disinformation and conspiracy thinking.

The effect of filter bubbles and echo chambers is disputed. Research has demonstrated the effect for extreme political ideas. People generally are not confined to a limited viewpoint or a single source of information and get to see content with different viewpoints, which softens the effect. The number of people believing in conspiracy theories, on the other hand, shows how real the problem of filter bubbles and echo chambers is.

How are social media platforms combating disinformation?

After the Russian interference with the 2016 US presidential elections was revealed, there has been a call from politics to combat disinformation. Measures have been proposed by governments. Social media companies themselves have also started to take measures.

Both Twitter and Facebook have made a start with becoming more transparent by publishing transparency reports and blogs explaining their efforts. Twitter has released entire datasets of accounts that have been disabled because they were suspected of political interference. Facebook is teaming with academia by providing researchers access to nonpublic datasets. While the information published by social media companies gives some insight in the efforts to combat disinformation and their effects, their detection modes are still not entirely clear and there is no way to determine the effectiveness of the interventions. There is still a long way to go before reaching a sufficient and understandable level of transparency.

What are considerations when balancing the right to freedom of expression and the fight against disinformation?

All measures proposed by regulators, governments, and tech companies are valuable. But the problem of disinformation being spread through social media is very complex and intervention can be a risk to the freedom of expression. Measures cannot be applied without consideration. The limitations of transparency, fact-checking, algorithms, and media and information literacy have been considered.

- Transparency is useful if people can comprehend it. If explanations are too general or too detailed, there is no actual understanding.
- Debunking false information can be done with fact-checking, but this method only exposes a limited amount of disinformation and is not enough to convince anyone.
- Automatic detection of disinformation isn’t perfect. Algorithms may contain hidden bias and can do unintended harm. Adversaries will be looking for new methods to evade existing detection algorithms.
- Media and information literacy is important to recognize disinformation, but it requires an active role of the user. We need to first understand the behavior of the user and create the setting in which critical thinking and common sense is activated.

What is internet filtering?

Whereas the internet was once seen as liberating, a lot of nation states have started to impose restrictions on the internet use of their citizens, and are using internet technology to spy on citizens or to launch information campaigns. Internet filtering started with blocking certain IP-addresses or websites, but the methods have become more sophisticated and the legal frameworks are in place to enable governments to apply filtering to the internet.

Next to governments imposing regulations on social media platforms a new form of internet filtering is emerging. By means of their algorithms social media companies themselves are deciding what users see and what they don’t see. They have defined their own policies to describe the allowed user behavior. The decision whether information is harmful and should be removed, or distributed less often, means that social media platforms are effectively censoring content. They do this based on their own assessments and are acting from their own values. Because the effects can only be observed at the level of individual users, this form of internet censorship is very opaque and can go unnoticed. It therefore is a very dangerous threat to the freedom of expression.

When does fighting disinformation by social media platforms become a form of internet filtering?

To get back to the main research question we conclude that there is a tension between doing too little and doing too much when it comes to fighting disinformation by social media platforms. Do too little and we let disinformation campaigns, hoaxes, and conspiracy theories prosper and drown out other sources of information. Do too much and we have limited people's possibility to speak out. Both affect our right to freedom of expression and information.

But the task ahead is broader than just finding the right balance between these two. The problem is much more complex. Due to the sheer amount of information being produced on social media we rely on automation. We need to understand the limitations of algorithms, e.g. the presence of bias, to determine their effectiveness. For the foreseeable future we will still have a need for human moderators to decide whether content is allowed. We need fact-checkers to debunk the false information, but a choice has to be made which stories are to be checked.

When intervention is the best option, it has to be done in a non-obtrusive way. By removing false or misleading content we might limit freedom of expression and are practicing a form of internet filtering. Instead an atmosphere needs to be created in which people feel safe to express their opinions, so a healthy discussion can arise. Whether fighting disinformation is perceived as internet filtering also depends on how well people understand why and how information is moderated. Social media platforms need to strive for a transparency level that is comprehensible and has enough detail.

8.2 Limitations

This is a qualitative study that has brought together different aspects to shed light on the challenges in fighting disinformation. The goal was to clarify definitions and measures in order to understand to what extent the battle against disinformation by social media platforms can limit our right to freedom of expression and information. While we have given some indications of how to find the right balance between intervention and letting information flow freely through social media platforms, the aim was not to give actual solutions. The results of this research are not to be taken as a conclusive set of actual recommendations.

The focus of this research has been on Twitter and Facebook, with occasional references to the efforts taken by YouTube. Other social media platforms have similar challenges, but there are differences in functionality, in user policies, in the underlying business model, and of course, in size and reach. Phenomena like filter bubbles and echo chambers might not be present on other platforms and distribution of information might be facilitated in another way.

To really grasp the problem of disinformation being spread, it is insufficient to examine the issue within one of the social media platforms. Information can be distributed from one platform to another. And social media is only a subset of the entire ecosystem of systems and platforms that are contributing to the distribution of information on the internet. The subjects of search engines, wikis, blogs or forums haven't been touched upon. Search engines have a significant impact on what information gets filtered out by means of the order of search results. To be able to effectively fight disinformation it isn't enough for some platforms to intervene, because distributors of information will just move to another platform. The big tech platforms can be imagined working together to reach the common understanding how intervention can best be done in a coordinated way. But this possibility clearly hasn't been examined in this research.

8.3 Suggestions for Further Research

This study has touched upon various subjects that can be further explored. To start with user behavior there is more knowledge to be gained on the drives of users causing them to like, share, or retweet content. The effect on users not being able to like or share with just a single click can be researched, as well as the effect of warnings being shown when people are intending to share content.

The way people are discussing delicate or controversial subjects on social media can be examined, as well as the way online discussions can be shaped into open, healthy conversations.

Next to further research on the effects of filter bubbles and echo chambers within one platform, the effects across several several platforms can be explored, as well as the distribution of information on the entire internet ecosystem.

Interventions by social media platforms and other platforms are largely encoded in their algorithms. Independent research is needed to be able to measure the effectiveness of interventions, and to be able to judge whether algorithm changes have the desired effect.

One of the biggest challenges in the fight against disinformation and unwanted behavior on social media platforms is the difference in regulations across countries. The big platforms are offering their services in almost all countries around the world, but they are confronted with a different regulation regime in each country. Information being legal in one country could be prohibited in another country. If we don't take care of this issue, we will end up with social media platforms choosing the safe way and block any information globally when it is illegal in just one country. Research can be done on international cooperation to determine measures, both technical as well as legal, against disinformation and regulations to social media platforms. We need to acknowledge the influence of the big tech companies in shaping our societies and if we don't unite, it's impossible to impose regulation that is effective against disinformation.

Finally the changes in algorithms of big tech companies can have a profound impact on existing businesses or on the way we spread our ideas. Therefore it is recommended to research how existing algorithms and algorithm change proposals can best be reviewed and by whom.

8.4 Concluding Remarks

This study has shed light on the pivotal role of social media platforms in today's quest for truth. To preserve our fundamental right to freedom of speech and right to information we have imposed an immense burden on their shoulders. We need their cooperation because they are anchoring themselves deeply in our society and are changing how we handle information, whether it concerns the communication with our friends and relatives, or the way we gather ideas to arrive at our beliefs and opinions. The problem cannot be solved by social media platforms alone. We have to view the broader picture of how people process information and how they connect with ideas emotionally. We need to understand the effects this has on the structure of our society.

Neil Postman warned us in 1985 that elements described in Aldous Huxley's novel *A Brave New World* had already become reality. As explained in the preface the people in this 1932 novel are occupied with entertainment and don't care about truth anymore. Postman argues that "when a population becomes distracted by trivia, when cultural life is redefined as a perpetual round of entertainments, when serious public conversation becomes a form of baby-talk, when, in short, a people become an audience and their public business a vaudeville act, then a nation finds itself at risk; culture-death is a clear possibility." (Postman, 1985, pp. 155–156)

Whereas Postman was referring to television as the culprit of a passive attitude towards information, we can conclude that the situation has not improved in the age of social media. We now have far more distractions. The interactive character of social media makes us complicit in the distribution of disinformation, while giving us the impression of being in control. The reality, however, is that we are non-thinking participants, following algorithm-based structures we aren't aware of. Like the people in *A Brave New World* we amuse ourselves with entertainment and communicate without deliberation because we have let our minds run away with whatever information captivates us. To regain control and effectively oppose the spread of disinformation we have to become conscious of our attitude.

The problem of disinformation being spread massively isn't a pure technical problem caused by the complex algorithms of social media, so it cannot be solved by technical interventions alone. Technology has certainly contributed to the severity of the problem, for example by making it too easy to share and like content that is provoking an emotional reaction. The role of technology is to reduce the opportunity for us to react purely from emotion and to force us to rethink before we share information. But the decision about what content should be allowed on our platforms and in what form needs an extensive public discussion between citizens, governments, academics, and tech companies. Together we can understand how social media has impacted the way we communicate with each other and the way our ideas are shaped. Together we can figure out how to deal with this profound change and to regain control. Let's build a world enriched with modern technology in which we trust social media again because we know how to debunk disinformation and find truth.

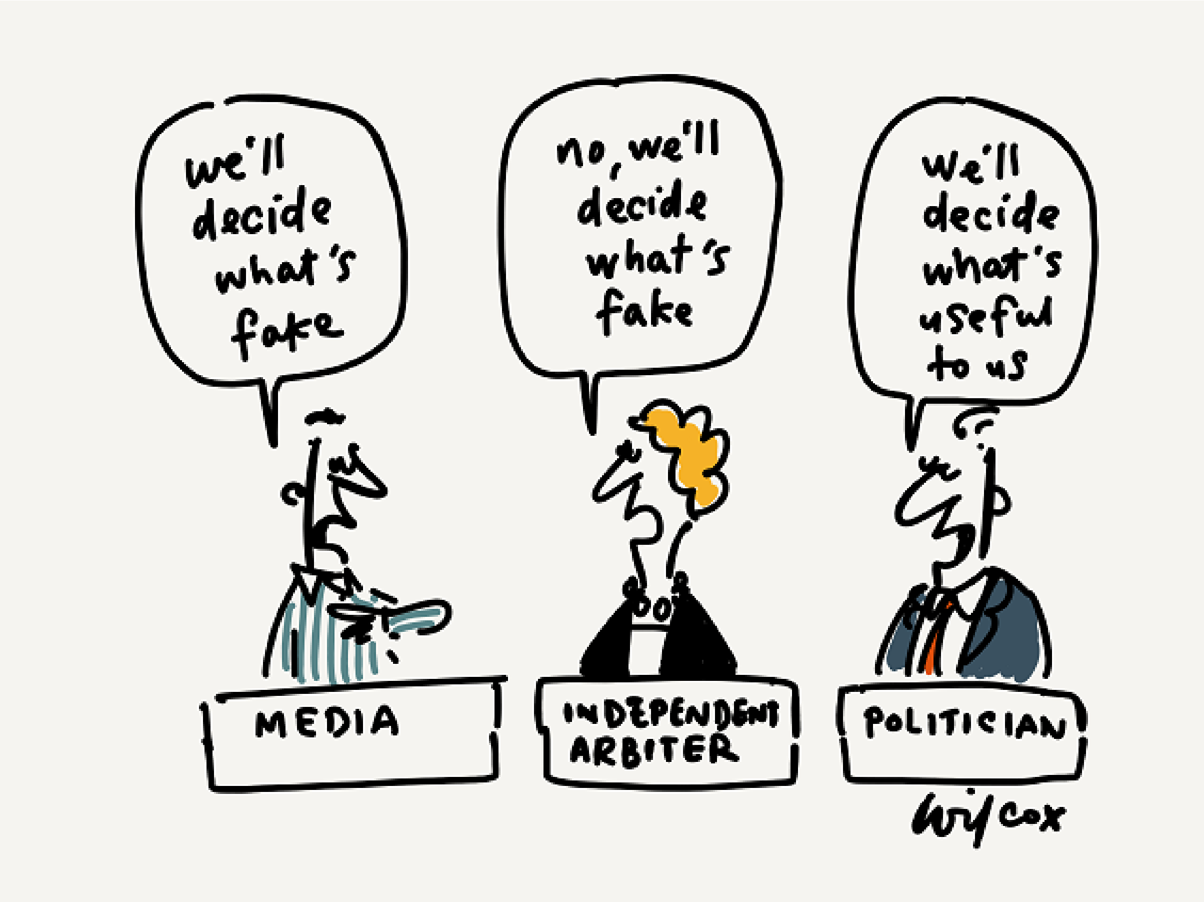


Figure 12: Cartoon by Cathy Wilcox, drawn for UNESCO for World Press Freedom Day 2017. Source: https://en.unesco.org/world-press-freedom-day-2017/drawing-wpfd-2017

References

A Multi-Dimensional Approach to Disinformation. Report of the Independent High Level Group on Fake News and Online Disinformation. (2018). *European Commission*. Retrieved from http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271

Alba, D. (2019). A Google VP Told The US Senate the Company Has “Terminated” the Chinese Search App Dragonfly. Retrieved October 4, 2019, from https://www.buzzfeednews.com/article/daveyalba/google-project-dragonfly-terminated-senate-hearing

Alexander, J. (2019a, August 5). YouTube CEO Addresses LGBTQ Community’s Ongoing Demonetization Concerns. *The Verge*. Retrieved from https://www.theverge.com/2019/8/5/20755315/youtube-lgbtq-creators-demonetization-age-gate-recommendation-susan-wojcicki

Alexander, J. (2019b, September 30). YouTube Moderation Bots Punish Videos Tagged as ‘Gay’ or ‘Lesbian,’ Study Finds. *The Verge*. Retrieved from https://www.theverge.com/2019/9/30/20887614/youtube-moderation-lgbtq-demonetization-terms-words-nerd-city-investigation

Ananny, M. (2016). Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology & Human Values*, *41*(1), 93–117. https://doi.org/10.1177/0162243915606523

Ananny, M., & Crawford, K. (2018). Seeing without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Bergen, M. (2019, April 2). YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant. *Bloomberg*. Retrieved from https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant

Berners-Lee, T. (2010). Long Live the Web. *Scientific American*, *303*(6).

Boghardt, T. (2009). Soviet Bloc Intelligence and Its AIDS Disinformation Campaign. *Studies in Intelligence*, *53*(4), 12–17.

Borel, B. (2017a, January 4). Fact-Checking Won’t Save Us from Fake News. *FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/fact-checking-wont-save-us-from-fake-news/

Borel, B. (2017b, February 21). How to Talk to Your Facebook Friends about Fake News. *The Open Notebook*. Retrieved from https://www.theopennotebook.com/2017/02/21/how-to-talk-to-your-facebook-friends-about-fake-news/

boyd, d. (2017a, January 27). The Information War Has Begun. Retrieved September 13, 2019, from http://www.zephoria.org/thoughts/archives/2017/01/27/the-information-war-has-begun.html

boyd, d. (2017b, March 27). Google and Facebook Can’t Just Make Fake News Disappear. *Data & Society: Points*. Retrieved from https://points.datasociety.net/google-and-facebook-cant-just-make-fake-news-disappear-48f4b4e5fbe8

boyd, d. (2019, April 24). The Fragmentation of Truth. *Data & Society: Points*. Retrieved from https://points.datasociety.net/the-fragmentation-of-truth-3c766ebb74cf

Boyd, W. (2014, April 28). Hitler’s Amazing Map that Turned America against the Nazis: A Leading Novelist’s Brilliant Account of How British Spies in the US Staged a Coup That Helped Drag Roosevelt to War. *Dailymail*. Retrieved from https://www.dailymail.co.uk/news/article-2673298/Hitlers-amazing-map-turned-America-against-Nazis-A-leading-novelists-brilliant-account-British-spies-US-staged-coup-helped-drag-Roosevelt-war.html

Brand, S. (1984). “Keep Designing” Hackers’Conference 1984. Retrieved October 14, 2019, from https://tech-insider.org/personal-computers/research/acrobat/8505-a.pdf

Bredford, B., Grisel, F., Meares, T. L., Owens, E., Pineda, B. L., Shapiro, J. N., Tyler, T. R., & Evans Peterman, D. (2019). Report Of The Facebook Data Transparency Advisory Group. Retrieved September 22, 2019, from https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf

Bruns, A. (2019). *Are Filter Bubbles Real?* Polity.

Charter of Fundamental Rights of the European Union. (2012). *European Convention*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN

Chaslot, G. (2019, February 9). YouTube Announced They Will Stop Recommending Some Conspiracy Theories such as Flat Earth. *Twitter*. Retrieved from https://twitter.com/gchaslot/status/1094359564559044610

Chen, C. (2016). Forming Digital Self and Parasocial Relationships on YouTube. *Journal of Consumer Culture*, *16*(1), 232–254. https://doi.org/10.1177/1469540514521081

Chung, S., & Cho, H. (2017). Fostering Parasocial Relationships with Celebrities on Social Media: Implications for Celebrity. *Psychology & Marketing*, *34*(4), 481–495. https://doi.org/10.1002/mar.21001

Code of Practice on Disinformation. (2018, September 26). Retrieved October 18, 2019, from https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

Cohen, N. (2011, January 9). Twitter Shines a Spotlight on Secret F.B.I. Subpoenas. *The New York Times*. Retrieved from https://www.nytimes.com/2011/01/10/business/media/10link.html

Community Standards Enforcement Report. (2019). Retrieved September 22, 2019, from https://transparency.facebook.com/community-standards-enforcement

Convention for the Protection of Human Rights and Fundamental Freedoms. (1950). *Council of Europe*. Retrieved from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680063765

Crowel, C. (2017, June 14). Our Approach to Bots and Misinformation. *Twitter*. Retrieved from https://blog.twitter.com/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html

D'Ancona, M. (2017). *Post-Truth: the New War on Truth and How to Fight Back.* Random House.

Deibert, R. (2015). Authoritarianism Goes Global : Cyberspace under Siege. *Journal of Democracy*, *26*(3), 64–78.

Deibert, R. (2019). The Road to Digital Unfreedom: Three Painful Truths about Social Media. *Journal of Democracy*, *30*(1), 25–39.

Deibert, R., & Rohozinski, R. (2010). Control and Subversion in Russian Cyberspace. In *Access Controlled: The Shaping of Power, Rights and Rule in Cyberspace* (pp. 15–34). The MIT Press.

Deibert, R., & Villeneuve, N. (2005). Firewalls and Power: An Overview of Global State Censorship of the Internet. In *Human Rights in the Digital Age* (pp. 125–138). Routledge-Cavendish.

Dijck, J. van. (2013). *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press.

Dorsey, J. (2019, October 30). We've Made the Decision to Stop All Political Advertising on Twitter Globally. Retrieved November 9, 2019, from https://twitter.com/jack/status/1189634360472829952

Dubois, E., & Blank, G. (2018). The Echo Chamber is Overstated: the Moderating Effect of Political Interest and Diverse Media. *Information, Communication & Society*, *21*(5), 729–745. https://doi.org/10.1080/1369118X.2018.1428656

Elections Integrity. (2018). Retrieved September 15, 2019, from https://about.twitter.com/en_us/values/elections-integrity.html#data

EU Code of Practice on Disinformation. (2018). *European Commission*. Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54454

Evans, J. (1998). *The History and Practice of Ancient Astronomy*. Oxford University Press.

Fact-Checking - Duke Reporter’s Lab. (n.d.). Retrieved October 12, 2019, from https://reporterslab.org/fact-checking/

Fallis, D. (2015). Exploring Philosophies of Information. *Trends*, *63*(3), 401–426.

Faris, R., & Villeneuve, N. (2007). Measuring Global Internet Filtering. In *Access Denied: The Practice and Policy of Global Internet Filtering* (pp. 5–27). The MIT Press.

Fetzer, J. H. (2004). Disinformation: The Use of False Information. *Minds and Machines*, *14*(2), 231–240. https://doi.org/10.1023/B:MIND.0000021683.28604.5b

Fletcher, R., & Nielsen, R. K. (2017). Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication. *Journal of Communication*, *67*(4), 476–498. https://doi.org/10.1111/jcom.12315

Floridi, L. (2011). *The Philosophy of Information*. Oxford University Press.

Gadde, V., & Roth, Y. (2018, October 17). Enabling Further Research of Information Operations on Twitter. *Twitter*. Retrieved from https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html

Gallagher, R. (2018, August 1). Google Plans to Launch Censored Search Engine in China, Leaked Documents Reveal. *The Intercept*. Retrieved from https://theintercept.com/2018/08/01/google-china-search-engine-censorship/

Gilani, Z., Farahbakhsh, R., Tyson, G., & Crowcroft, J. (2019). A Large-Scale Behavioural Analysis of Bots. *ACM Transactions on the Web (TWEB)*, *13*(1), 1–23.

Gillespie, T. (2010). The Politics of ‘Platforms.’ *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Gioe, D. V., Goodman, M. S., & Wanless, A. (2019). Rebalancing Cybersecurity Imperatives: Patching the Social Layer. *Journal of Cyber Policy*. https://doi.org/10.1080/23738871.2019.1604780

Gleicher, N. (2019, October 21). How We Respond to Inauthentic Behavior on Our Platforms: Policy Update. *Facebook*. Retrieved from https://newsroom.fb.com/news/2019/10/inauthentic-behavior-policy-update/

Golebiewski, M., & boyd, d. (2018, May). Data Voids: Where Missing Data Can Easily Be Exploited. *Data & Society*. Retrieved from https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

Grind, K., Schechner, S., McMillan, R., & West, J. (2019, November 15). How Google Interferes With Its Search Algorithms and Changes Your Results. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753

Haidt, J., & Rose-Stockwell, T. (2019, December). The Dark Psychology of Social Networks: Why it Feels Like Everything is Going Haywire. *The Atlantic*. Retrieved from https://www.theatlantic.com/magazine/archive/2019/12/social-media-democracy/600763/

Hanani, U., Shapira, B., & Shoval, P. (2001). Information Filtering: Overview of Issues, Research and Systems. *User Modeling and User-Adapted Interaction*, *11*(3), 203–259. https://doi.org/10.1023/A:1011196000674

Harrison, S. (2019, July 7). Twitter's Disinformation Data Dumps Are Helpful — to a Point. *Wired*. Retrieved from https://www.wired.com/story/twitters-disinformation-data-dumps-helpful/

Harvey, D., & Gasca, D. (2018, May 15). Serving Healthy Conversation. *Twitter*. Retrieved from https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html

Hearing Before the United States Senate Select Committee on Intelligence. (2017). *115th Cong. 13 (11/1/17) (Testimony of Colin Stretch, General Counsel of Facebook)*. Retrieved from https://www.intelligence.senate.gov/sites/default/files/documents/os-cstretch-110117.pdf

Hobbs, W. R., & Roberts, M. E. (2018). How Sudden Censorship Can Increase Access to Information. *American Political Science Review*, *112*(3), 621–636. https://doi.org/10.1017/S0003055418000084

Hvistendahl, M. (2019). Citizens of the World's Edge. *Popular Science*, *291*(3), 74–123.

IFCN Code of Principles. (n.d.). Retrieved October 12, 2019, from https://www.ifcncodeofprinciples.poynter.org/

Informing the "Disinformation" Debate. (2018). Retrieved from https://edri.org/files/online_disinformation.pdf

Ingold, J. (2018, November 20). We Went to a Flat-Earth Convention and Found a Lesson about the Future of Post-Truth Life. *The Colorado Sun*. Retrieved from https://coloradosun.com/2018/11/20/flat-earth-convention-denver-post-truth/

Jansen, P. (2019, September 7). Als middel erger blijkt dan kwaal. Retrieved September 9, 2019, from https://www.telegraaf.nl/watuzegt/407789292/als-middel-erger-blijkt-dan-kwaal

Johnson, J. R. (2013). The Authenticity and Validity of Antony's Will. *L'antiquité Classique*, *47*(2), 494–503. https://doi.org/10.3406/antiq.1978.1908

Jones, A. (2019, January 30). First on CNN: NY Attorney General Targets Fake Social Media Activity. *CNN*. Retrieved from https://edition.cnn.com/2019/01/30/tech/new-york-attorney-general-social-media/

Jowett, G. S., & O'Donnell, V. (1999). *Propaganda & Persuasion*. SAGE Publications.

Kang, M. (2014). Understanding Public Engagement: Conceptualizing and Measuring its Influence on Supportive Behavioral Intentions. *Journal of Public Relations Research*, *26*(5), 399–416. https://doi.org/10.1080/1062726X.2014.956107

Kantrowitz, A. (2019, July 23). The Man Who Built The Retweet: "We Handed A Loaded Weapon To 4-Year-Olds." Retrieved from https://www.buzzfeednews.com/article/alexkantrowitz/how-the-retweet-ruined-the-internet

Kreitner, R. (2016, November 30). Post-Truth and Its Consequences: What a 25-Year-Old Essay Tells Us About the Current Moment. *The Nation*. Retrieved from https://www.thenation.com/article/post-truth-and-its-consequences-what-a-25-year-old-essay-tells-us-about-the-current-moment/

Lapowsky, I. (2019, March 15). Why Tech Didn't Stop the New Zealand Attack From Going Viral. *Wired*. Retrieved from https://www.wired.com/story/new-zealand-shooting-video-social-media/

Lecher, C. (2017, October 19). Senators Announce New Bill that Would Regulate Online Political Aads. *The Verge*. Retrieved from https://www.theverge.com/2017/10/19/16502946/facebook-twitter-russia-honest-ads-act

Lee, D. (2016, November 23). Facebook "Made China Censorship Tool." *BBC.Com*. Retrieved from https://www.bbc.com/news/technology-38073949

Levy, S. (2018, May 1). Mark Zuckerberg Says It Will Take 3 Years to Fix Facebook. *Wired*. Retrieved from https://www.wired.com/story/mark-zuckerberg-says-it-will-take-3-years-to-fix-facebook/

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). "THE RUSSIANS ARE HACKING MY BRAIN!" Investigating Russia's Internet Research Agency Twitter Tactics during the 2016 United States Presidential Campaign. *Computers in Human Behavior*, *99*(February), 292–300. https://doi.org/10.1016/j.chb.2019.05.027

Lohr, S. (2013, March 10). Algorithms Get a Human Hand in Steering Web. *The New York Times*. Retrieved from https://www.nytimes.com/2013/03/11/technology/computer-algorithms-rely-increasingly-on-human-helpers.html

Lovink, G., & Rossiter, N. (2009). The Digital Given: 10 Web 2.0 Theses. Retrieved October 12, 2019, from http://fourteen.fibreculturejournal.org/fcj-096-the-digital-given-10-web-2-0-theses/

Lyons, T. (2017, December 20). Replacing Disputed Flags with Related Articles. *Facebook*. Retrieved from https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/

Lyons, T. (2018, May 23). Hard Questions: What's Facebook's Strategy for Stopping False News? *Facebook*. Retrieved from https://newsroom.fb.com/news/2018/05/hard-questions-false-news/

Mac, R., & Bernstein, J. (2019, October 3). Attorney General Bill Barr Will Ask Zuckerberg to Halt Plans for End-to-End Encryption across Facebook's Apps. *BuzzFeedNews*. Retrieved from https://www.buzzfeednews.com/article/ryanmac/bill-barr-facebook-letter-halt-encryption

MacKinnon, R. (2012). *Consent of the Networked: The Worldwide Struggle for Internet Freedom*. Basic books.

Martin, D. (2001, March 25). Charles Johnson, 76, Proponent of Flat Earth. *The New York Times*. Retrieved from https://www.nytimes.com/2001/03/25/us/charles-johnson-76-proponent-of-flat-earth.html

Martineau, P. (2019, August 22). Why People Keep Falling for Viral Hoaxes. *Wired*. Retrieved from https://www.wired.com/story/why-people-keep-falling-viral-hoaxes/

McCarthy, T. (2017, October 31). Facebook, Google and Twitter Grilled by Congress over Russian Meddling – as it Happened. *The Guardian*. Retrieved from https://www.theguardian.com/technology/live/2017/oct/31/facebook-google-twitter-congress-russian-election-meddling-live

McIntyre, T. J., & Scott, C. (2008). Internet Filtering: Rhetoric, Legitimacy, Accountability and Responsibility. In *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes* (pp. 109–124). Hart Publishing.

Meineck, S. (2018, March 9). Deshalb ist "Filterblase" die blödeste Metapher des Internets. *Vice.Com*. Retrieved from https://www.vice.com/de/article/pam5nz/deshalb-ist-filterblase-die-blodeste-metapher-des-internets

Miller, J. M., Saunders, K. L., & Farhart, C. E. (2016). Conspiracy Endorsement as Motivated Reasoning: The Moderating Roles of Political Knowledge and Trust. *American Journal of Politicial Science*, *60*(4), 824–844. https://doi.org/10.1111/ajps.12234

Mosseri, A. (2016, June 29). Building a Better News Feed for You. Retrieved from https://newsroom.fb.com/news/2016/06/building-a-better-news-feed-for-you/

Mueller, R. S. (2019). *Report on the Investigation into Russian Interference in the 2016 Presidential Election* (Vol. I and II).

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2018). Reuters Institute Digital News Report 2018. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). Reuters Institute Digital News Report 2019. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_1.pdf

Newton, C. (2019, March 6). Mark Zuckerberg Says Facebook Will Shift to Emphasize Encrypted Ephemeral messages. *The Verge*. Retrieved from https://www.theverge.com/2019/3/6/18253458/mark-zuckerberg-facebook-privacy-encrypted-messaging-whatsapp-messenger-instagram

Nicas, J. (2018, February 7). How YouTube Drives People to the Internet's Darkest Corners. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478

Nielsen, R. K., & Graves, L. (2017). "News You Don't Believe": Audience Perspectives on Fake News, (October), 1–8. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-10/Nielsen&Graves_factsheet_1710v3_FINAL_download.pdf

Nimmo, B. (2019). UK Trade Leaks. Retrieved December 7, 2019, from https://graphika.com/uploads/Graphika Report - UK Trade Leaks - 12.19.pdf

Nissenbaum, H. (2011). A Contextual Approach to Privacy Online. *Daedalus*, *140*(4), 32–48.

Obar, J. A., & Wildman, S. S. (2015). Social Media Definition and the Governance Challenge: an Introduction to the Special Issue. *Telecommunications Policy*, *39*(9), 745–750. https://doi.org/10.1016/j.telpol.2015.07.014

Pariser, E. (2011). *The Filter Bubble: What the Internet is Hiding from You.* Penguin UK.

Pasquale, F. (2016). Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power. *Theoretical Inquiries in Law*, *17*(2), 487–513.

Pennycook, G., & Rand, D. G. (2019a). Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

Pennycook, G., & Rand, D. G. (2019b). Lazy, not Biased: Susceptibility to Partisan Fake News is Better Explained by Lack of Reasoning Than by Motivated Reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Picheta, R. (2019, November 18). The Flat-Earth Conspiracy is Spreading around the Globe. Does it Hide a Darker Core? *CNN*. Retrieved from https://edition.cnn.com/2019/11/16/us/flat-earth-conference-conspiracy-theories-scli-intl/index.html

Platform Manipulation. (2018). Retrieved September 17, 2019, from https://transparency.twitter.com/en/platform-manipulation.html

Plous, S. (1993). *The Psychology of Judgment and Decision Making*. Mcgraw-Hill Book Company.

Postman, N. (1985). *Amusing Ourselves to Death: Public Discourse in the Age of Show Busines*. Viking Penguin.

Reilly, T. O. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, *1*(4580).

Rid, T. (2017, November 1). Why Twitter is the Best Social Media Platform for Disinformation. *Vice.Com*. Retrieved from https://www.vice.com/en_us/article/bj7vam/why-twitter-is-the-best-social-media-platform-for-disinformation

Rogers, K. (2019, July 11). White House Hosts Conservative Internet Activists at a ‘Social Media Summit.’ *The New York Times*. Retrieved from https://www.nytimes.com/2019/07/11/us/politics/white-house-social-media-summit.html

Rogers, R., & Niederer, S. (2019, October 18). The Politics of Social Media Manipulation. *Digital Methods Initiative, Media Studies, University of Amsterdam*. Retrieved from https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2019/10/18/rapport-politiek-en-sociale-media-manipulatie/rapport-politiek-en-sociale-media-manipulatie.pdf

Romano, A. (2017, September 26). Twitter Says Trump’s North Korea Tweets Don’t Violate its Policy against Violent Threats. *Vox.Com*. Retrieved from https://www.vox.com/culture/2017/9/26/16367510/twitter-trump-threats-not-policy-violations

Romano, A. (2018, September 6). How Hysteria over Twitter Shadow-Banning Led to a Bizarre Congressional Hearing. *Vox.Com*. Retrieved from https://www.vox.com/2018/9/6/17824652/twitter-dorsey-energy-and-commerce-hearing-shadow-banning

Ronson, J. (2015, February 15). How One Stupid Tweet Blew Up Justine Sacco’s Life. Retrieved from https://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html

Rosen, G., Harbath, K., Gleicher, N., & Leathern, R. (2019, October 21). Helping to Protect the 2020 US Elections. Retrieved December 7, 2019, from https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/

Roth, Y. (2019, June 13). Information Operations on Twitter: Principles, Process, and Disclosure. *Twitter*. Retrieved from https://blog.twitter.com/en_us/topics/company/2019/information-ops-on-twitter.html

Rowbotham, S. B. (1865). *Zetetic Astronomy: Earth not a Globe*.

Salim, S. (2019, January 4). How Much Time Do You Spend on Social Media? Research Says 142 Minutes Per Day. Retrieved from https://www.digitalinformationworld.com/2019/01/how-much-time-do-people-spend-social-media-infographic.html

Sample, I. (2019, February 17). Study Blames YouTube for Rise in Number of Flat Earthers. *The Guardian*. Retrieved from https://www.theguardian.com/science/2019/feb/17/study-blames-youtube-for-rise-in-number-of-flat-earthers

Schrage, E., & Ginsberg, D. (2018, April 9). Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections. *Facebook*. Retrieved from https://newsroom.fb.com/news/2018/04/new-elections-initiative/

Schroepfer, M. (2019, September 5). Creating a Data Set and a Challenge for Deepfakes. *Facebook Artificial Intelligence*. Retrieved from https://ai.facebook.com/blog/deepfake-detection-challenge

Seidel, J. (2017, November 15). US Congress Told How Russia 'Weaponised' This Photo of a 'Muslim Woman.' *News.Com.Au*. Retrieved from https://www.news.com.au/technology/online/social/us-congress-told-how-russia-weaponised-this-photo/news-story/8135ca050976761ab476c04c32d44fd2

Shaban, H. (2018, June 21). Facebook Expands its Fact-Checking Tools But Says its Work 'Will Never Be Finished.' Retrieved October 21, 2019, from https://www.washingtonpost.com/news/the-switch/wp/2018/06/21/facebook-expands-its-fact-checking-tools-but-says-its-work-will-never-be-finished/

Silverman, C. (2016, November 16). This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. *BuzzFeedNews*. Retrieved from https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Simon, H. A. (1971). Designing Organizations for an Information-Rich World. In *Computers, Communication, and the Public Interest* (pp. 37–72). The Johns Hopkins Press.

Simonite, T. (2018, May 3). How Artificial Intelligence Can—and Can't—Fix Facebook. *Wired*. Retrieved from https://www.wired.com/story/how-artificial-intelligence-canand-cantfix-facebook/

Singer, J. B. (2019). Fact-Checkers as Entrepreneurs. *Journalism Practice*, *13*(8), 976–981.

Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The Weaponization of Social Media*. Eamon Dolan Books.

Smith, K. (2019a, June 1). 53 Incredible Facebook Statistics and Facts. Retrieved from https://www.brandwatch.com/blog/facebook-statistics/

Smith, K. (2019b, June 13). 126 Amazing Social Media Statistics and Facts. Retrieved from https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/

Stack, L. (2018, July 26). What is a ‘Shadow Ban,’ and is Twitter Doing It to Republican Accounts? *The New York Times*. Retrieved from https://www.nytimes.com/2018/07/26/us/politics/twitter-shadowbanning.html

Statt, N. (2019, April 30). Facebook Is Redesigning its Core App around the Two Parts People Actually Like to Use. *The Verge*. Retrieved from https://www.theverge.com/2019/4/30/18523265/facebook-events-groups-redesign-news-feed-features-f8-2019

Stone, B. (2011, January 28). The Tweets Must Flow. *Twitter*. Retrieved from https://blog.twitter.com/en_us/a/2011/the-tweets-must-flow.html

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Swart, J. A. C. (2018). *Haven’t You Heard? Connecting Through News and Journalism in Everyday Life*. Rijksuniversiteit Groningen.

Tackling Online Disinformation: a European Approach. (2018). *European Commission* (Vol. COM(2018)2). Retrieved from https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-236-F1-EN-MAIN-PART-1.PDF

the Flat Earth Society. (n.d.). Retrieved August 30, 2019, from https://www.theflatearthsociety.org/home/

Tufekci, Z. (2018, March 10). YouTube, the Great Radicalizer. *The New York Times*. Retrieved from https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Twitter Health Metrics Proposal Submission. (2018, March 1). *Twitter*. Retrieved from https://blog.twitter.com/official/en_us/topics/company/2018/twitter-health-metrics-proposal-submission.html

Universal Declaration of Human Rights. (1948). *United Nations*. Retrieved from https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf

Update on Twitter’s Review of the 2016 US Election. (2018, January 19). *Twitter*. Retrieved from https://blog.twitter.com/en_us/topics/company/2018/2016-election-update.html

Updating our Advertising Policies on State Media. (2019, August 19). *Twitter*. Retrieved from https://blog.twitter.com/en_us/topics/company/2019/advertising_policies_on_state_media.html

Vaidhyanathan, S. (2018). *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford University Press.

Wardle, C. (2018). The Need for Smarter Definitions and Practical, Timely Empirical Research on Information Disorder. *Digital Journalism*, *6*(8), 951–963. https://doi.org/10.1080/21670811.2018.1502047

Wardle, C. (2019). How You Can Help Transform the Internet into a Place of Trust. Retrieved November 6, 2019, from https://www.ted.com/talks/claire_wardle_how_you_can_help_transform_the_internet_into_a_place_of_trust

Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. *Council of Europe*.

Warzel, C. (2016, August 11). Sources: Twitter CEO Dick Costolo Secretly Censored Abusive Responses to President Obama. *BuzzFeedNews*. Retrieved from https://www.buzzfeednews.com/article/charliewarzel/sources-twitter-ceo-dick-costolo-secretly-censored-abusive-r

Wiener, A. (2016, November 26). Why Can’t Silicon Valley Solve Its Diversity Problem? *The New Yorker*. Retrieved from https://www.newyorker.com/business/currency/why-cant-silicon-valley-solve-its-diversity-problem

Word of the Year 2016 is... (2016). Retrieved November 3, 2019, from https://languages.oup.com/word-of-the-year/word-of-the-year-2016

YouTube. (2019, January 25). Continuing Our Work to Improve Recommendations on YouTube. Retrieved November 24, 2019, from https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html

Zittrain, J., & Palfrey, J. (2007). Internet Filtering: The Politics and Mechanisms of Control. In *Access Denied: The Practice and Policy of Global Internet Filtering* (pp. 29–56). The MIT Press.

Zuckerberg, M. (2018, September 12). Preparing for Elections. Retrieved September 22, 2019, from https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/10156300047606634/