Rathenau Instituut

Digital threats to democracy

On new technology and disinformation

Report

Authors
Pieter van Boheemen, Geert Munnichs and Elma Dujso

Cover photo
Shutterstock

Preferred citation:
Rathenau Instituut (2020). *Digital threats to democracy – On new technology and disinformation*. The Hague (authors: Boheemen, P. van, G. Munnichs & E. Dujso)

Foreword

Until recently, the Netherlands could nurture the belief that disinformation has not had a major impact on society in recent years. However, the flood of misleading messages that have circulated about the coronavirus pandemic has now shown that Dutch society is also not impervious to it. At the same time, it is still too soon to form a definitive opinion about their significance for the resilience of Dutch society against disinformation.

However, the rapid pace of technological developments in the field of IT could change the situation within the foreseeable future. This report, which has been written at the request of the Ministry of the Interior and Kingdom Relations and falls under the theme of Digital Society in our Working Programme, gives a general overview of technological developments that could play a role in the production and dissemination of disinformation in the coming years. The survey is far from reassuring. Technologies such as text synthesis, voice cloning, deepfakes, micro-targeting and chatbots provide producers and disseminators of disinformation significant and diverse possibilities to mislead internet users.

It therefore appears that more measures will have to be taken to counter the threats to public debate and the democratic process from technology-driven disinformation. Measures such as better detection of deepfakes, monitoring by platform companies of abuse of the potential of micro-targeting by advertisers and better facilitation of fact-checkers.

Primary responsibility for many of the measures suggested in this report lie with the platform companies. An important question in the years ahead will therefore be how much public responsibility they are willing to accept in combatting disinformation. If they fail to do enough, the government could decide that the public interest demands stricter regulation of platform companies in relation to preventing the harmful effects of disinformation. This report offers suggestions for both eventualities.

Melanie Peters
Director, Rathenau Instituut

Summary

Introduction
The Ministry of the Interior and Kingdom Relations asked the Rathenau Instituut to conduct research into the impact of technological developments on the production and dissemination of disinformation and measures that could be taken to mitigate their potential negative effects. The report focuses mainly on disinformation aimed at disrupting public debate and the democratic process. The study reflects the action lines that the minister, Kajsa Ollongren, announced in her letter to the House of Representatives on 18 October 2019 as part of the government's strategy to protect Dutch society against disinformation.

This study focuses on the following questions:
- What is the impact of technological developments on the production and dissemination of disinformation?
- What measures have already been taken to contain the threats that disinformation poses for public debate and the democratic process?
- What new measures can be taken to counter those threats, taking account of freedom of speech and press freedom?
- Who are the relevant actors in that respect?

Approach
This study is based on desk research, interviews with experts and two case studies. The case studies relate to important technological developments connected with the production and dissemination of disinformation: deepfakes and psychographing. The interim results of the study were discussed during an online meeting with experts. This report describes the results of the research.

All of the technological developments that were investigated have a digital component. The developments discussed are already underway or are expected to occur within the next five years. None of the technologies described in this study can be regarded as 'entirely new'. However, we show how technological innovations that are already in development or which are starting to emerge could evolve and what impact those innovations could have on the production and dissemination of disinformation.

Disinformation
In this study we adopt the definition of disinformation used by the Minister of the Interior and Kingdom Relations in the aforementioned letter to the House of Representatives: 'the conscious, usually covert, dissemination of misleading

information with the aim of causing damage to the public debate, democratic processes, the open economy or national security'. We make the reservation that this study focuses primarily on disinformation that undermines or disrupts public debate and the democratic process, for example by stirring up social divisions or feeding distrust in political institutions.

Previous research has shown that there are no visible signs that disinformation is having a major impact on society at present. Most of the examples of disinformation in this study are therefore taken from other countries, but they also illustrate what the Netherlands might come to face in terms of disinformation in the coming years.

The study consists of three parts, each with its own distinct character: a quick scan with a survey of technological developments; case studies that explore two specific technologies in more depth; and a preview of new measures that could be taken.

Part I: Quick scan

The quick scan provides an overview of technological developments that could play a role in the production and dissemination of disinformation in the coming years. It also presents a concise survey of measures that have already been taken to combat the negative effects of disinformation. In the quick scan we make a distinction between general technologies, production technologies and dissemination technologies.

General technologies

- Database technology: the large-scale collection and analysis of (personal) data;
- Artificial intelligence: self-learning algorithms and systems.

Technologies with which disinformation can be produced

- Text synthesis: algorithms that generate readable and logical news reports and messages;
- Voice cloning: manipulation of voice messages using artificial intelligence;
- Image synthesis and deepfakes: generation and modification of videos using artificial intelligence;
- Augmented and virtual reality and avatars: presentation of information in a virtual environment;
- Memes: images designed to be widely shared on social media.

Technologies with which disinformation can be disseminated

- Social media platforms: online platforms such as Facebook, Twitter and TikTok, which use recommendation algorithms to select messages;

- Micro-targeting: reaching specific target groups with a message geared to them (using campaign software, dynamic prospecting, programmatic advertising, psychographing and influencer marketing);
- Chat apps: sharing (encrypted) messages, one-to-one or in small groups;
- Bots: (partially) automated accounts on social media;
- Search engines: platforms that enable the internet to be searched;
- Virtual assistants: voice-operated devices which can be used to consult search engines, among other things;
- Distributed autonomous applications: online platforms with no central control;
- Games: online games;
- Cross-media storytelling: reaching a specific person or target group via various channels and devices.

Part II: Case studies

Building on the quick scan, two case studies were elaborated to provide a more coherent picture of how technological developments in the area of disinformation could evolve in the coming years and what impact they could have on public debate and the democratic process. The case studies concern deepfakes and psychographing.

Deepfakes

Artificial intelligence can be used to manipulate audiovisual material. This can make it difficult for people to distinguish manipulated videos – deepfakes – from the real thing. For example, the face in an image can be changed with ‘face swaps’ or an artificial head or body can be generated with ‘digital puppetry’. Deepfakes can be used, for example, to create the impression that a certain person made a particular statement, which can impair public debate.

It is likely that further technological innovation will make deepfakes increasingly difficult to distinguish from authentic, non-manipulated images. In addition, increasingly advanced deepfake technologies will come onto the market in easy-to-use apps and gadgets. Accordingly, the use of deepfakes will become increasingly common. Given the growing importance of video on internet, that could undermine the credibility of visual material published by established news media.

In response to the increasing ability of platform companies to detect deepfakes, producers and disseminators of deepfakes could switch to closed channels without moderators.

Psychographing

Psychographing is an advanced form of micro-targeting. It is an advertising technology that can be used to gear messages in an automated way to the

personality traits of a target group. The idea behind the method is that people can be influenced by feeding them information that is tailored to their psychological profile. Large numbers of internet users could be misled or manipulated in this way.

The case study sketches a scenario in which a group sets out to influence public debate with the help of psychographing. By involving itself in sensitive social issues, the group endeavours to stir up social divisions and undermine public confidence in established institutions. To cause maximum unease, the messages could be disseminated via non-public channels, such as private groups on Facebook or Telegram, and since there is little chance of the messages being contradicted on those channels, the disinformation campaign would have an even greater impact.

Part III: Outlook

In the outlook we describe new measures that could be taken to combat the most important technology-driven threats to public debate and the democratic process.

Measures against deepfakes

Investment in detection of deepfakes

Platform companies could invest in the active detection of deepfakes in order to combat them. They will need to if they are to compete in the possible race with the producers and disseminators of increasingly advanced deepfakes.

Establishment of a hotline for malicious image manipulation

Companies like SnapChat, Instagram and TikTok, on whose platforms deepfakes are increasingly common, could create a hotline where users can report suspicions of malicious image manipulation.

Authentication of visual material and other messages

The digital authentication of visual material and other messages would enable internet users to verify whether material is from a source they regard as reliable. That calls for a reliable system of registering digital hallmarks. The government and large technology companies could take the lead in this.

Restricting possibilities for micro-targeting

Monitoring the use of advertising technology

Platform companies could build a monitoring function into their services to combat abuse of the advertising technology they provide.

Technical possibilities for limiting advertising technology

Platform companies could impose restrictions on advertisers with respect to their selection of target groups and monitor the responsible use of the advertising technology they provide by the advertisers.

Providing transparency for internet users
Platform companies could provide internet users with better information about the use that advertisers make of advertising profiles.

Measures against the harmful effects of recommendation algorithms
A built-in pause for reflection in platform services
Recommendation algorithms of platform companies frequently reinforce the social and political preferences of users and – by extension – social divisions. To combat the harmful effects of this, platform companies could build a pause for reflection into the use of their services. In this way, users would be less likely to share information (and disinformation) impulsively.

Providing transparency about recommendation algorithms
To combat the harmful effects of recommendation algorithms, platform companies could be transparent about how the algorithms work. To start with, they could provide scientific researchers with access to them.

Warning system for closed and encrypted channels
One way of combating the dissemination of disinformation on closed and encrypted channels would be to establish an independent warning system that identifies and issues a warning about disinformation campaigns on sensitive social issues. The government and platform companies could facilitate this warning system.

Critical analysis of the revenue model of platform companies
Measures such as limiting the use of advertising technology and providing transparency about how recommendation algorithms work could conflict with the business model of platform companies. They might therefore be disinclined to take those measures. In that case, the government could go further, for example by compelling greater transparency about the use of recommendation algorithms or critically analysing the platform companies' revenue model.

Investment in fact-checking remains important
Because fact-checking is important to provide certainty for internet users looking for reliable information, the government and platform companies could invest, or continue to invest, in facilities for fact-checkers.

Investment in media literacy remains important
The production and dissemination of disinformation could be reduced with technological measures and stricter regulation of platform companies. But there will always be safe havens on the internet, and internet users will therefore continue to be confronted with disinformation. The government must therefore continue to invest in media literacy.

Conclusion: platform companies are primarily responsible, but the government can intervene

With many of the above measures to combat disinformation, responsibility lies primarily with the platform companies. But given the public interest in preventing the harmful effects of disinformation, the government could decide to act if platform companies do not fully meet that responsibility. For example, the government could urge the platform companies to adopt an active policy on the detection and prevention of deepfakes or to monitor irresponsible use by advertisers of the possibilities of micro-targeting.

And if urging the companies doesn't help, measures could be compelled. Those measures could also be at the expense of the platform companies' earnings model. Whether the government should take this step will depend in part on the seriousness of the threats to public debate and the democratic process arising from the polarising effect of recommendation algorithms or disinformation campaigns by advertisers facilitated by platform companies. To carry sufficient weight, compulsory measures should logically be taken at EU level.

Contents

Introduction

1.1 Background

Political influence campaigns by Russian trolls, the interference by the political consultancy firm Cambridge Analytica in the Brexit referendum and confusing reports about the source of and strategy for combatting the coronavirus have led to growing concerns about the political and social effects of disinformation. Political propaganda and misrepresentation are of course nothing new. For example, the distribution of pamphlets with insinuations against political opponents for political gain was not uncommon in Dutch politics in the seventeenth century.[1] But with the rise of the information society, the manner in which untrue or misleading information is developed and disseminated has assumed new and far more extensive forms, which we are also confronted with on a daily basis. That raises new questions.[2]

This study explores the nature and spread of disinformation in the present era, how new technologies influence them and what measures can be taken to combat their negative effects. The Rathenau Instituut has carried out this research at the request of the Dutch Ministry of the Interior and Kingdom Relations. The study ensues from the action lines announced by the minister, Kajsa Ollongren, in her letter to the House of Representatives on 18 October 2019 as part of the government's strategy to protect Dutch society against disinformation.[3] With that strategy, the government has demonstrated its awareness of the importance of technological developments for the way in which various forms of disinformation can be produced and disseminated, including the question of whether Dutch society is adequately equipped to cope with them.

1.2 Goal and research question

The goal of this study is to investigate the potential impact of technological developments on the production and dissemination of disinformation and the

[1] Haverkate, J.M.M. (2019). Spindoctors van de Gouden Eeuw: De eerste pamfletoorlog van Overijssel (1654-1675). https://research.vu.nl/en/publications/spindoctors-van-de-gouden-eeuw-de-eerste-pamfletoorlog-van-overij.
[2] Balaban, D. (2018). News Sharing on Social Media Platforms. Theoretical Approaches. *Communication. Strategic Perspectives*. www.academia.edu/38666829/News_Sharing_on_Social_Media_Platforms._Theoretical_Approaches
[3] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2019). *Kamerbrief over beleidsinzet bescherming democratie tegen desinformatie*. www.rijksoverheid.nl/documenten/kamerstukken/2019/10/18/kamerbrief-over-beleidsinzet-bescherming-democratie-tegen-desinformatie

measures that can be taken to mitigate their potential negative effects. The main focus is on disinformation aimed at disrupting public debate and the democratic process.

The following questions are addressed in the study:

- What is the impact of technological developments on the production and dissemination of disinformation?
- What measures have already been taken to contain the threats that disinformation poses for public debate and the democratic process?
- What new measures can be taken to counter those threats, taking account of freedom of speech and press freedom?
- Who are the relevant actors in that respect?

1.3 Approach

The study is based on desk research and interviews with experts, as well as two case studies that were carried out to explore the subject matter in more depth on the basis of the results of the first two steps. The findings from the desk research, the interviews and the case studies were then discussed in an online meeting with experts. This final report presents the results of these activities. The table below contains a list of the research activities.

Research activity	Function
Desk research	To produce an overview of relevant technological developments and their significance for disinformation and measures that could be taken to counter the threats it poses for public debate and the democratic process.
Interviews	To supplement the desk research with new insights into technological developments, their significance for disinformation and possible measures.
Quick scan	Synthesis of the results of the desk research and the interviews. The quick scan constituted an interim report on the research.
Case studies	To gain a deeper understanding of some important technological developments and related measures.
Expert meeting	To explore possible measures in more depth, on the basis of the quick scan and the case studies.
Final report	To provide a synthesis of the findings from the various research activities.

Table 1 List of research activities

The following section gives a further explanation of the approach, including the scope of the study, the desk research and the interviews.

1.3.1 Scope of the study

The two central concepts in this study, ‘technological developments’ and ‘disinformation’, are both broad topics. To create a workable structure for the study, we defined the scope of the study as follows.

Technological developments that could influence the production and dissemination of disinformation include both current developments and developments that are likely to occur in the coming years. Because of the rapid pace of technological developments in this field, a time horizon of roughly five years was adopted. In our view, in such a rapidly evolving domain there is little point in looking more than five years ahead in making suggestions for measures that could be taken.

All of the technological developments that were investigated in the context of disinformation have a digital component, which can relate to the production of disinformation, its dissemination, or both. They include the possibilities that artificial intelligence and online platform companies provide for producing and disseminating disinformation.

None of the technologies described in this study can be regarded as ‘entirely new’, in the sense of as yet unknown. Any description of such technologies would be science fiction. What we do show is how technological innovations that are already being developed or are starting to emerge could further evolve and what impact they could have on the production and dissemination of disinformation. These innovations are embodied in steadily improving applications – as in the case of deepfakes – or in the possibilities of methods such as psychographing or influencer marketing to facilitate more advanced forms of micro-targeting.

The definition of disinformation that we have used is discussed in more detail in Chapter 2.

1.3.2 Desk research

Desk research yielded an overview of relevant technological developments and their significance for disinformation, as well as an initial list of possible policy

options, for which we used primary and secondary scientific sources. Relevant references in these sources were consulted in online databases.

The scientific databases Scopus, ISI Web of Science, Google Scholar, IEEE Explore and SSRN were searched for recently published scientific literature using the following search terms: ‘disinformation’ (402 results), ‘strategic communications’ (53 results), ‘micro-targeting’ (28 results), ‘deepfake’ (16 results), ‘post-truth’ (711 results) and ‘online-harm’ (4 results). The articles that at first glance provided most insight into the phenomenon of disinformation, related technological developments and measures relating to them were then selected.

Also consulted were relevant publications from national organisations (AIVD, NCTV, Ministry of the Interior and Kingdom Relations, ROB, CPB), European institutions (EC, STOA), media research firms (Reuters, PEW Research), online platform companies (Facebook, Twitter, Google) and relevant scientific organisations (Bits of Freedom, AlgorithmWatch, The Intercept, Electronic Frontier Foundation).

Finally, news sources that are followed by professionals in the field (iBestuur, Emerce.nl, Security.nl, Reddit and MrDeepfakes forum) were also consulted.

1.3.3 Interviews

The findings from the desk research were supplemented with the results of interviews with experts in the field of disinformation and associated areas of expertise. The interviews focused mainly on what the interviewees regarded as the most relevant developments in the field of disinformation and the most important measures that can be taken against it. The interviews were conducted on the basis of a questionnaire sent to the interviewees in advance (see Appendix 1) and were semi-structured. Appendix 2 contains a list of the interviewees.

1.3.4 Quick scan, case studies and expert meeting

The results of the desk research and the interviews were incorporated into the quick scan and the two case studies. The case studies concern two important and impactful technological developments relating to the production and dissemination of disinformation: deepfakes and psychographing.

The quick scan and the case studies were discussed during an online expert meeting. The emphasis in that meeting was on measures that could be adopted to

contain threats to public debate and the democratic process from technology-driven disinformation. A list of the participants at the expert meeting can be found in Appendix 3.

The selection of the participants in the interviews and the expert meeting and the choice of case studies were discussed with the Ministry of the Interior and Kingdom Relations. Draft versions of the quick scan and the final report were discussed with an interdepartmental focus group established by the ministry for the purpose. As an independent research institute, the Rathenau Instituut used the outcomes of that discussion as it saw fit. The Rathenau Instituut is therefore entirely responsible for the content of this report.

1.4 Reader's guide

Chapter 2 discusses what is meant by disinformation in this study and to what extent disinformation occurs in the Netherlands.

The other chapters describe the findings from the research and are divided into three parts, each of which is different in nature.

Part I contains the results of the quick scan. It provides a broad overview of technological developments that could play a role in the production and dissemination of disinformation in the coming years (Chapters 3 to 5). The quick scan also provides a concise overview of measures that are already being taken to counter the threats from disinformation for public debate and the democratic process (Chapter 6).

Part II describes two case studies that provide a more coherent impression of how technology relating to disinformation could develop in the coming years and what impact those developments could have on public debate and the democratic process. The case studies concern deepfakes and psychographing (Chapter 7).

Part III looks ahead with a description of new measures that could be taken to combat harm to public debate and the democratic process as a result of technology-driven disinformation, and what actors are responsible for taking those measures (Chapter 8). It also contains a concluding chapter with a summary of the main findings from the study (Chapter 9).

2 Disinformation

To describe the impact of technological developments on the production and dissemination of disinformation, it is first necessary to explain what we mean by disinformation in this study. After all, opinions differ on that. We also describe what is known about the extent to which disinformation occurs in the Netherlands and how that relates to the situation in other countries.

2.1 Definition of disinformation

There are numerous definitions of disinformation. The word usually refers to the spread of 'untrue', 'inaccurate' or 'misleading' information, but those terms are themselves often difficult to interpret. For example, in practice it is often difficult to make a distinction between providing poor-quality information and spreading lies.[4] [5] It is clear that the context in which information is disseminated and the purpose for which it is done partly determine whether that information can be regarded as disinformation.

Over the years there have been shifts in the definition of disinformation in the scientific literature. Schultz and Godson, for example, defined disinformation as 'false, incomplete or misleading information that is passed, fed, or confirmed to a targeted individual, group, or country'.[6] In other words, their focus was on the content of the information and the envisaged target group.

More recent definitions of disinformation have also focused on the intention of the producer or disseminator of the disinformation. It can only be regarded as disinformation if it involves a malicious party acting consciously.[7] For example, the Netherlands Bureau for Economic Policy Analysis (CPB) defines disinformation as 'knowingly creating and disseminating false, inaccurate or misleading information'.[8]

---

[4] Wardle, C., & Derakhshan, H. (2017). *INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making*. https://rm.coe.int/information-disorder-report-november-2017/1680764666
[5] RAND (2019). *What's Being Done to Fight Disinformation Online*. www.rand.org/research/projects/truth-decay/fighting-disinformation.html
[6] Shultz, R. H., & Godson, R. (1984). *Dezinformatsia: Active Measures in Soviet Strategy* (1st edition). University of Nebraska Press.
[7] Gelfert, A. (2018). Fake News: A Definition. *Informal Logic*, *38*(1), 84–117. https://doi.org/10.22329/il.v38i1.5068
en Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317
[8] CPB (2019). *Cyber Security Risk Assessment for the Economy 2019*. CPB. www.cpb.nl/sites/default/files/omnidownload/cpb-notitie-risicorapportage-cyberveiligheid-2019.pdf

The problem with this definition is that it is not always clear in practice what the intention of the individual who produces or disseminates disinformation is or whether that person is acting knowingly.

Humprecht therefore suggests inserting a more specific objective into the definition of disinformation: the person who produces or disseminates disinformation must have the intention of causing harm or generating profit or social influence with it.[9]

In her letter to the House of Representatives on 18 October 2019, Minister Ollongren appeared to endorse Humprecht's proposed definition, while limiting the intention required of a person who produces or dissemination information to causing harm. She described disinformation as 'the conscious, usually covert, dissemination of misleading information, with the aim of causing harm to public debate, democratic processes, the open economy or national security'.[10] This definition is very much in line with those adopted by the European Commission and the British House of Commons.[11] [12]

In a more recent letter to the House of Representatives on 13 May 2020, the minister clarified what she meant by conscious dissemination, remarking that misleading information can also be disseminated by people without any conscious desire to cause harm.[13] This unconscious dissemination of misleading information could be described as misinformation.[14]

At the other end of the spectrum of disinformation are utterances that are prohibited by law, such as defamation, hate speech or incitement of violence. In contrast to misinformation, such utterances can be prosecuted.

This also means that the production and dissemination of disinformation is not as such prohibited. The government can therefore not remove misleading information

---

[9] Humprecht, E. (2018). Where 'fake news' flourishes : a comparison across four Western democracies. *Information, Communication and Society*, *21*, 1–16. https://doi.org/10.1080/1369118X.2018.1474241
[10] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2019). *Kamerbrief over beleidsinzet bescherming democratie tegen desinformatie*. www.rijksoverheid.nl/documenten/kamerstukken/2019/10/18/kamerbrief-over-beleidsinzet-bescherming-democratie-tegen-desinformatie
[11] EC DG CONNECT HLEG on Fake News (2018). *A multi-dimensional approach to disinformation*
[12] DCMSC of the House of Commons (2019). *Disinformation and 'fake news' Report* www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/fake-news-report-published-17-19/
[13] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2020). *Kamerbrief ontwikkelingen beleidsinzet bescherming democratie tegen desinformatie.* www.rijksoverheid.nl/documenten/kamerstukken/2020/05/13/kamerbrief-ontwikkelingen-beleidsinzet-bescherming-democratie-tegen-desinformatie
[14] Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to Online Disinformation: A Framework for Cross-National Comparative Research. The International Journal of Press/Politics, 1940161219900126. https://doi.org/10.1177/1940161219900126

without further reason, since that would be in conflict with freedom of speech.[15] [16] Platform companies, on the other hand, can remove misleading information if disseminating disinformation is in breach of their terms of service.

Harm to public debate and the democratic process

In this study we adopt the definition of disinformation formulated by Minister Ollongren, but with the reservation that it focuses mainly on the production and dissemination of disinformation that causes harm or disrupts public debate and the democratic process. Examples of such harm or disruption might be accentuating polarisation in society, feeding mistrust of political institutions or covertly influencing political opinion-forming by citizens. For example, the State Commission on the Parliamentary System in the Netherlands (the Remkes Commission) argued that covertly influencing political opinion-forming conflicts with the basic principles of free and fair elections.[17]

We would add that we regard public debate as a key element of the democratic legal order. The purpose of public debate is to enable public expression and contradiction of political preferences and views. It is also intended to enable voters to form their political preferences and views, and if necessary adjust or revise them, also in light of arguments put forward by others in the debate.[18] Encroachments on this key democratic function of public debate are in our view a disruption of it.

Against this background, it can also be clearly shown why covert influencing of political opinion-forming by citizens with the help of micro-targeting – something that Cambridge Analytica was accused of during the Brexit campaign in the United Kingdom – is problematic. Because the firm targeted specific groups with tailored political messages that were not revealed to others, those political messages could not be refuted by other voters or political groups. Consequently, it was also impossible to investigate to what extent the various political messages were mutually compatible.

---

[15] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2019). *Actielijnen tegengaan desinformatie.* www.rijksoverheid.nl/documenten/kamerstukken/2019/10/18/actielijnen-tegengaan-desinformatie

[16] IViR (2019). *Het juridisch kader voor de verspreiding van desinformatie via internetdiensten en de regulering van politieke advertenties*. Universiteit van Amsterdam. www.ivir.nl/publicaties/download/Rapport_desinformatie_december2019.pdf

[17] Staatscommissie parlementair stelsel (2018). *Lage drempels, hoge dijken: Democratie en rechtsstaat in balans.* www.staatscommissieparlementairstelsel.nl/documenten/rapporten/samenvattingen/12/13/eindrapport

[18] Munnichs, G.M. (2000). *Publiek ongenoegen en politieke geloofwaardigheid: democratische legitimiteit in een ontzuilde samenleving.* https://www.rug.nl/research/portal/nl/publications/publiek-ongenoegen-en-politieke-geloofwaardigheid(fffc2cec-2962-4114-b655-9118961af83c).html

2.2 Relevant groups

A variety of actors can be involved in the production and dissemination of disinformation. Frequently mentioned groups include:[19]

- State actors and affiliated groups;
- Extremist groups;
- Economically motivated actors, such as the young Macedonians who were active during the US presidential election in 2016;[20]
- Professional marketing organisations, such as the political consultancy firm Cambridge Analytica;
- Social media platforms.

The motives of the relevant groups in disseminating disinformation can vary greatly. State actors and allied groups frequently intend to stoke public unrest by causing confusion, creating social divisions and/or casting doubt on the reporting of established institutions – often without any clear political agenda. Disinformation spread by extreme-right groups, for example, does often have a clear political agenda, but not necessarily. Disinformation can also be spread for purely economic reasons, as was the case with the young Macedonians. By posting messages designed to attract maximum attention on social media, the disseminators can earn money – whereby the content of the message is purely incidental.

The actions of the groups concerned are often opportunistic. They exploit vulnerabilities in society, take advantage of discussions being conducted in the media, and use the technical resources that will have the greatest effect.

It is often impossible to discover who is responsible for producing or spreading disinformation, or why they are doing so. The disseminator might, for instance, have an interest in not being recognised and adopt a false identity. For example, Facebook and Twitter recently revealed that Russian actors are involved in organisations in Ghana and Nigeria that portray themselves as American and join in debates on politically sensitive issues.[21]

---

[19] Benkler, Y., Faris, R., & Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press. www.oxfordscholarship.com/view/10.1093/oso/9780190923624.001.0001/oso-9780190923624

[20] Subramanian, S. (2017). *Meet the Macedonian Teens Who Mastered Fake News and Corrupted the US Election.* Wired www.wired.com/2017/02/veles-macedonia-fake-news/

[21] Culliford, E. (2020). Facebook, Twitter remove Russia-linked accounts in Ghana targeting U.S. *Reuters*. www.reuters.com/article/us-facebook-content-idUSKBN20Z3LW

2.3 Key elements

In the following chapters we employ a number of key elements to help in describing the impact of technological developments on the production and dissemination of disinformation. These elements are listed in the table below.

Element	Explanation
Sender	One or more actors who are responsible for the production and/or dissemination of disinformation.
Intention	The sender's motives for producing or disseminating disinformation.
Content	The message that is conveyed, which is intended to persuade the recipients to change their minds or pursue a particular course of action.
Form	The form in which disinformation is presented, for example in audio or video.
Medium	The method by which the disinformation is transmitted and reaches the recipient, for example via a platform.
Recipient	The person or group who receives the disinformation.
Effect	The (envisaged) change in the thoughts or behaviour of the recipient.

Table 2 Key elements

2.4 Disinformation in the Netherlands

According to the available literature, there is little disinformation in the Netherlands at the moment. Previous research by the Rathenau Instituut (2018), showed that disinformation presently has little visible impact on society in the Netherlands.[22] Disinformation disseminated in the Netherlands was found to come mainly from economically motivated actors, who often use 'pulp news' or 'click-bait' to draw people to advertising sites. Only a small proportion of it is of a political nature or concerns socially sensitive issues. Research by the Oxford Internet Institute confirms that impression.[23]

[22] Van Keulen, I., Korthagen, I., Diederen, P., & Van Boheemen, P. (2018). *Digitalisering van het nieuws*. Rathenau Instituut.
[23] Blood, D. (2017). *Is social media empowering Dutch populism?* Oxford Internet Institute. www.ft.com/content/b1830ac2-07f4-11e7-97d1-5e720a26771b

In its annual report for 2019 the Dutch General Intelligence and Security Service (AIVD) observed that Russian groups were endeavouring to spread disinformation, but that its impact appeared to be limited at the moment.[24] This means that they generated little online interaction (likes and shares) and that the narratives they spread attracted scarcely any attention. Recent research by the University of Amsterdam further showed that disinformation played no significant role during the provincial and European elections in 2019.[25]

A lot of the literature on disinformation relates to the situation in other countries, such as the United States and the United Kingdom. But because English-language messages can also reach the public in the Netherlands, English-language disinformation campaigns can indirectly have an effect on public debate in the Netherlands.

Dutch citizens are in any case concerned about disinformation. According to a survey by the Dutch national newspaper de Volkskrant, 82% of the Dutch population regard disinformation as a threat to the functioning of democracy and the rule of law.[26] [27] According to the CPB, concerns about the consequences of disinformation for public opinion have increased in recent years. In that context, the CPB refers to examples of disinformation in other countries and to the activities of Russian trolls on Twitter after the shooting down of Flight MH17.[28]

Disinformation during the corona pandemic

A possible exception to the situation described above has occurred recently and concerns the large number, by Dutch standards, of misleading messages and conspiracy theories surrounding the coronavirus outbreak. That is in fact a worldwide phenomenon. The World Health Organization (WHO) also drew attention to the threat of disinformation when it declared the pandemic. Tedros Adhanom Ghebreyesus, the WHO's director-general, spoke of an 'infodemic': 'We're not just fighting an epidemic; we're fighting an infodemic'.[29]

---

[24] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2020). *AIVD-jaarverslag 2019*. www.aivd.nl/documenten/jaarverslagen/2020/04/29/jaarverslag-2019

[25] Rogers, R., & Niederer, S. (2019). *Politiek en Sociale Media Manipulatie*. University of Amsterdam. www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2019/10/18/rapport-politiek-en-sociale-media-manipulatie/rapport-politiek-en-sociale-media-manipulatie.pdf

[26] Kranenberg, A. (2017). *Nederlanders bezorgd over 'nepnieuws' - een op drie weet vaak niet meer wat waar is en wat onwaar.* Volkskrant www.volkskrant.nl/nieuws-achtergrond/nederlanders-bezorgd-over-nepnieuws-een-op-drie-weet-vaak-niet-meer-wat-waar-is-en-wat-onwaar~b6914596/

[27] Kranenberg, A. (2017). *Wie weet nog wat er waar is?* Volkskrant www.volkskrant.nl/kijkverder/2017/desinformatie/

[28] CPB (2019). *Risicorapportage cyberveiligheid economie 2019*. CPB. www.cpb.nl/sites/default/files/omnidownload/cpb-notitie-risicorapportage-cyberveiligheid-2019.pdf

[29] Adhanom, T. (2020). *Munich Security Conference*. www.who.int/dg/speeches/detail/munich-security-conference

The corona crisis has unleashed a wide range of disinformation and false rumours in the Netherlands, from falsification of reports published by the Dutch public broadcaster NOS and the government to wild conspiracy theories.[30] For example, misleading messages about medicines to treat corona and warnings against particular medicines have circulated on social media and in chat groups.[31]

However, it is still too soon to make a conclusive judgement of the possible impact of the flood of corona-related disinformation on how the public and political debate is conducted in the Netherlands. For example, it is not always clear whether there is malicious intent, or whether it is misinformation rather than disinformation.

2.5 International developments

Because most examples of disinformation come from other countries, for this study we have also reviewed what is happening abroad. This has also given us a better insight into what the Netherlands might face in terms of disinformation in the coming years.

In this section, we describe a number of trends that can be observed in other countries. It has to be remembered that the nature and impact of disinformation can vary from one country to another. For example, researchers found that the themes that were the subject of disinformation were very different in the United Kingdom than in Germany.[32] This also means that the harm caused by disinformation in other countries will not necessarily occur in the same way or with the same effect in the Netherlands.

Disinformation is a growing international phenomenon
Research by the University of Oxford has shown that disinformation has occurred in a growing number of countries over the years. In 2018, evidence of organised disinformation campaigns was found in 48 countries, a sharp increase compared with the 28 countries in the previous year.[33]

[30] Vermanen, J., & Van Bree, T. (2020). *Flinke stijging van onbetrouwbaar nieuws over coronavirus op Twitter.* Pointer https://pointer.kro-ncrv.nl/node/280
[31] Kist, R., & Nieber, L. (2020). *Misinformatie over coronavirus gaat ook viraal.* NRC. www.nrc.nl/nieuws/2020/03/09/misinformatie-over-coronavirus-gaat-ook-viraal-a3993140
[32] Humprecht, E. (2018). Where 'fake news' flourishes : a comparison across four Western democracies. *Information, Communication and Society*, *21*, 1–16. https://doi.org/10.1080/1369118X.2018.1474241
[33] Bradshaw, S., & Howard, P. (2018). *Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation*. University of Oxford. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf

Disinformation is widespread

Examples from other countries clearly show that disinformation campaigns are widespread.[34] The best-known examples are the disinformation campaigns during the presidential election in the United States in 2016. Research by the American government into Russian interference in the election showed that thousands of accounts controlled by Russian groups produced and disseminated more than a million tweets, hundreds of thousands of Facebook messages and a thousand YouTube videos.[35] The tweets were viewed 288 million times and the Facebook messages 126 million times. [36] Because of these enormous numbers, the Russian Internet Research Agency (IRA), which is held responsible for them, acquired the nickname 'the troll factory'.[37]

The IRA's activities are not confined to Facebook, Twitter and YouTube, but have also been observed on Google+, Vine, Meetup, Pinterest, Tumblr, Gab, Medium and Reddit. Between 2014 and 2017, the IRA was able to instigate 187 million interactions on Instagram and more than 76 million on Facebook.[38] This further demonstrates that senders of disinformation do not confine themselves to the major platforms, but also take advantage of the possibilities afforded by smaller platforms.

As already mentioned, disinformation can also be disseminated for economic gain. Many producers or disseminators of disinformation have an economic motive. Their revenue model is based on displaying advertisements in or with disinformation. According to the Disinformation Index website, the producers and disseminators of disinformation represent a total global market value of 235 million dollars.[39] This substantial amount also helps to explain the capacity of disseminators of disinformation to employ innovative technologies.

Professional services

Another noteworthy international trend is the growing professionalism of the production and dissemination of disinformation. It is now possible to buy disinformation campaigns on websites (underground forums) on which professional

---

[34] Nemr, C., & Gangware, W. (2019). *WEAPONS OF MASS DISTRACTION: Foreign State-Sponsored Disinformation in the Digital Age*. Park Advisors. www.state.gov/wp-content/uploads/2019/05/Weapons-of-Mass-Distraction-Foreign-State-Sponsored-Disinformation-in-the-Digital-Age.pdf

[35] US Office of the Director of National Intelligence (2017). *'Background to "Assessing Russian Activities and Intentions in Recent US Elections," The Analytic Process and Cyber Incident Attribution'*. www.dni.gov/files/documents/ICA_2017_01.pdf.

[36] US House of Representatives (2018). *Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements* https://intelligence.house.gov/social-media-content

[37] Linvill, D. L., & Warren, P. L. (2020). Troll Factories: Manufacturing Specialized Disinformation on Twitter. *Political Communication*, *0*(0), 1–21. https://doi.org/10.1080/10584609.2020.1718257

[38] DiResta et al. (2018). *'The Tactics and Tropes of the Internet Research Agency,' en Philip N. Howard et al. (2018). 'The IRA, Social Media and Political Polarization in the United States, 2012-2018'*. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/IRA-Report-2018.pdf

[39] Disinformation Index (no date) https://disinformationindex.org/research/

operators offer their services. Recorded Future explored this market and found extensive price lists, including, for example, an offer to place an authentic-looking article in the Financial Times, with content chosen by buyer, for roughly 50,000 dollars.[40]

Accordingly, senders of disinformation do not need to possess special skills to make use of particular technological possibilities. Professional service providers make it easier to use new technologies.

[40] Insikt Group (2019). *The Price of Influence: Disinformation in the Private Sector.* www.recordedfuture.com/disinformation-service-campaigns/

Part I Quick scan

3 General technologies

This quick scan presents a general overview of technological developments that could play a role in the production and dissemination of disinformation in the coming years. It also provides a summary of measures that are already being taken to counter the threats to public debate and the democratic process from disinformation.

As the term quick scan implies, it is an exploratory survey that was carried out in a relatively short space of time. Accordingly, it only discusses the various developments and their interrelationship to a limited extent. The case studies in Part II provide a more detailed description of a few major technological developments and their possible impact on the production and dissemination of disinformation.

The quick scan contains four chapters. This chapter (Chapter 3) describes two general technologies which often form the basis of the technologies, which are discussed afterwards, that are used for production (chapter 4) and dissemination (chapter 5) of disinformation. Chapter 6 describes the measures that are already being taken to counter the harmful effects of disinformation. Appendix 4 contains a list of the technologies discussed in the quick scan.

As already mentioned in the introductory chapter, it has to be remembered that all of the technologies discussed here are still evolving. This means that their significance for the production and dissemination of disinformation in the coming years may diverge from what is forecast in this study.

The two general technologies discussed in this chapter are:

- Database technology
- Artificial intelligence.

3.1 Database technology

Technologies that are used to produce and disseminate disinformation increasingly make use of database technology. Database technology allows large volumes of information to be collected and analysed. The data in databases – data for short – are, as it were, the raw material of disinformation. Access to large volumes of data

and the technology to mine them for information are therefore becoming increasingly important.

More and more data are being collected all the time, a trend that will only escalate in the coming years. For example, online advertisers use cookie technology and tracking codes to constantly monitor the online behaviour – and hence the preferences – of internet users, from the websites they visit to the time they spend scrolling through a web page. Google scans e-mails and private chats to gather information that can be used to target personalised advertisements at internet users.

It is becoming ever easier to link data from different databases and sources, which will allow producers and disseminators of disinformation to make a better estimate of what message and message design best match the views and needs of the recipient. Such forms of micro-targeting (see Chapter 5) could then also be used to disseminate disinformation tailored to specific personal characteristics.

The data that are relevant for producing disinformation are not necessarily acquired legally. Data can come from hacked databases, accidental leaks from databases or publicly accessible data (open data). For example, American scientists have succeeded in estimating which party will be voted for in a particular neighbourhood on the basis of features from the public databases of Google Streetview. This information can be used to produce and disseminate disinformation targeted at political campaigns (see figure 1).[41]

[41] Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, *114*(50), 13108–13113. https://doi.org/10.1073/pnas.1700035114

Figure 1 Useful information for the production of disinformation can be derived from features in public databases, such as Google Streetview.

3.2 Artificial intelligence

Some of the aforementioned database technologies and many of the technologies discussed in chapters 4 and 5 use artificial intelligence. This section briefly explains what we mean by that.

Artificial intelligence (AI) refers to computer systems that display a certain degree of intelligence.[42] There are various technologies for building such a system.[43] A basic technology is what is known as rule-based AI. Computer systems built with this method are programmed with the help of 'if this, then that' instructions. For example, a computer might suggest to a user that updates should be installed as soon as they are available. This type of message has now become so common that they are often no longer regarded as intelligent behaviour.

A more advanced form of AI is machine learning. Systems that use machine learning have pre-programmed instructions, but are also capable of deriving instructions from data. The system analyses existing data and learns to discover patterns in them, which it then applies to new data. The pattern recognition can constantly improve.

Deep learning (DL) is a specific form of machine learning. The technology is based on so-called neural networks, in which different layers of information are combined. For example, a deep-learning system devoted to voice recognition consists of three

[42] Van Boheemen, P., Munnichs, G., Kool, L., Diercks, G., Hamer, J., & Vos, A. (2020). *Cyberweerbaar met nieuwe technologie*. Rathenau Instituut. www.rathenau.nl/nl/digitale-samenleving/cyberweerbaar-met-nieuwe-technologie
[43] European Commission. (2019). *A definition of Artificial Intelligence: main capabilities and scientific disciplines*. https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

layers. First, the sound frequency of a voice is analysed. This is then combined with a layer that analyses the tempo and cadence of speech. A third layer analyses the use of language. By combining these elements, voices can be recognised. The deep-learning technology is also used to manipulate audio and video material – in other words, to produce deepfakes (see section 4.3).

Data are an important raw material for AI systems because the systems learn and improve from them. The use of AI systems and database technology therefore often go hand in hand.

4 Production technologies

In this chapter we discuss some technological developments that are likely to be relevant for the production of disinformation in the coming years. Those technologies are:

- Text synthesis
- Voice cloning
- Image synthesis and deepfakes
- Memes
- Augmented and virtual reality and avatars

4.1 Text synthesis

Text synthesis can be used to generate new, easy to read and logical texts, with minimal, or even no human control.

An example of this technology is OpenAI GPT-2. This AI system was trained on the basis of eight million text documents and web pages. The system is designed to predict the next word in a sentence from the preceding words. In contrast to other AI language models, which are trained in a specific domain, this model is not domain-specific and is therefore widely applicable.[44] [45]

This form of AI is expected to become steadily better at producing texts that are difficult or impossible to distinguish from authentic texts. At the moment, producers of disinformation still have to devote a certain amount of time and creativity to writing (misleading) texts, but with text synthesis large quantities of text could be generated in no time. Large-scale use of AI-generated text could lead to fake news reports drowning out authentic reporting, thus leading to disruption of public debate.

Text synthesis could also be used to influence the ranking algorithms of search engines. One of the criteria used by these algorithms is the number of links to a particular article. The more often an article is cited, the higher its ranking. With computer systems that use text synthesis, it would be possible to generate large numbers of articles that refer to each other and thus influence the ranking algorithm's scoring system. This would enable malicious parties to draw public

[44] OpenAI (2019). *GPT-2: 1.5B Release* https://openai.com/blog/gpt-2-1-5b-release/
[45] Vincent, J. (2019). *OpenAI has published the text-generating AI it said was too dangerous to share*. The Verge www.theverge.com/2019/11/7/20953040/openai-text-generation-ai-gpt-2-full-model-release-1-5b-parameters

attention to misleading articles that appear authentic, and hence influence public opinion.

4.2 Voice cloning

Artificial intelligence increases the possibilities for manipulating audio messages. Algorithms that modify spoken messages are particularly important for the production of disinformation, since people can be misled if the voice of a person they trust is successfully simulated.

Software such as Lyrebird, Adobe Voco, CorentinJ/Real-time Voice cloning, iSpeech, Resemble, Tacotron 2 and CereVoice Me is already able to do this, albeit with mixed results. The software enables users to alter recorded conversations with the help of synthesised speech, which can also be mixed with the ambient sound.

Short sound fragments are required for an accurate simulation of a person's voice. With AI and developments in the area of text-to-speech synthesis, researchers are already able to make an almost perfect voice clone on the basis of a recording of a person's voice lasting just a few seconds.[46] Since more and more people share fragments of their speech, for example in videos on social media, a growing number of people could be the target of this form of deception.

Voice-cloning technology already causes (economic) harm. It is used by criminals for what is known as CEO fraud, where employees in financial departments are persuaded to transfer money with a simulation of a manager's voice.[47]

4.3 Image synthesis and deepfakes

Image manipulation is an existing phenomenon, as demonstrated by the editing of photos with programs like Photoshop. New technologies also make it possible and ever easier to edit and to generate videos. Artificial intelligence also plays a crucial role in these technologies.[48]

---

[46] FTC (2019). *You Don't Say: An FTC Workshop on Voice Cloning Technologies.* www.ftc.gov/news-events/events-calendar/you-dont-say-ftc-workshop-voice-cloning-technologies

[47] Malik, D. (2020). *AI Based Voice Cloning Is Giving Rise To Another Big Security Scam.* www.digitalinformationworld.com/2020/03/ai-based-voice-cloning-is-giving-rise-to-another-big-security-scam.html

[48] Khodabakhsh, A., Busch, C., & Ramachandra, R. (2018). A Taxonomy of Audiovisual Fake Multimedia Content Creation Technology. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 372–377). Miami, FL: IEEE. https://doi.org/10.1109/MIPR.2018.00082

Deepfakes are an example of AI image synthesis. A deepfake is a video fragment that looks genuine, but has been manipulated using deep-learning algorithms. The method uses an autoencoder, which can reconstruct input, and a generative adversarial network (GAN). A GAN is a computer system that combines two neural networks: one generates images (see figure 2) and the other evaluates their quality.[49] [50] Deepfakes are generated from existing visual material.

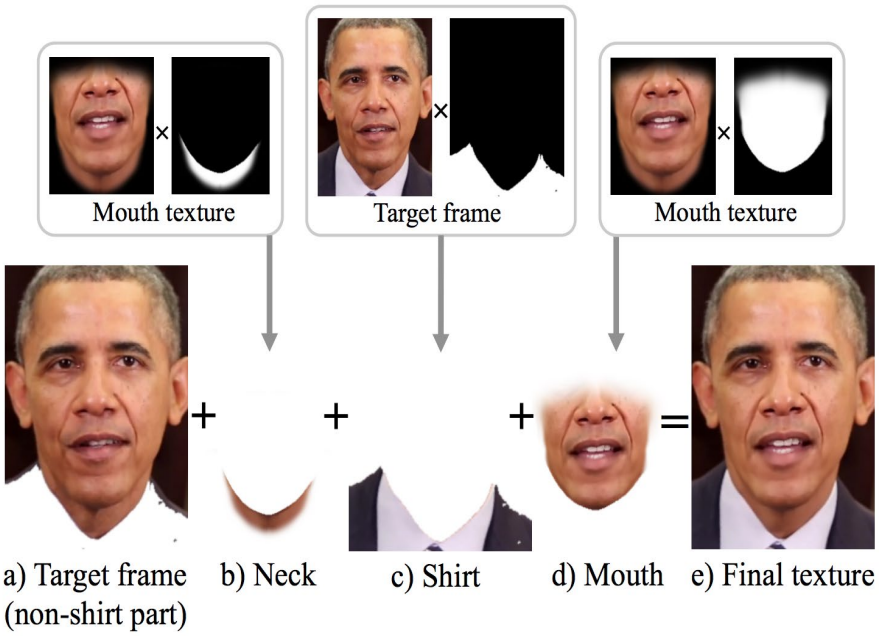

Figure 2 Illustration of the construction of a composite portrait.

Image synthesis can now also be applied to live video. For example, with the help of Face2Face and HeadOn, faces, facial expressions and movements can be replaced in live videos.[51] Given the rapid increase in the computing power of smartphones and the greater bandwidth of mobile telephone networks, real-time deepfakes are expected to become more common.

Deepfake technology is becoming increasingly accessible. Easy-to-use apps such as Doublicat already offer limited possibilities to swap faces.[52] The software used for more advanced deepfakes such as DeepFaceLab is also freely available (open source). However, considerable technical skill is required to use them. In more

[49] U.S. Government Accountability Office. (2020). *Science & Tech Spotlight: Deepfakes*, (GAO-20-379SP). www.gao.gov/products/gao-20-379sp
[50] Martineau, P. (2019). *Facebook Removes Accounts With AI-Generated Profile Photos. Wired.* www.wired.com/story/facebook-removes-accounts-ai-generated-photos/
[51] Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., & Nießner, M. (2018). HeadOn: Real-time Reenactment of Human Portrait Videos. *ACM Transactions on Graphics*, *37*(4), 1–13. http://arxiv.org/abs/1610.03151
[52] Neocortext, Inc. (2020) *REFACE*. Google Play https://play.google.com/store/apps/details?id=video.reface.app&hl=en

advanced versions of this software, the speech in the video can be converted into text and revised text can then be 'spoken' in the modified video.[53] The expectation is that easy-to-use apps with which many people can produce deepfakes will be available in the near future.

The technology behind deepfakes is still being developed. According to cyber security company Nisos, there are not yet any providers of advanced manipulation of videos as a service on the dark web. According to Nisos, that suggests that the technology is not yet sufficiently advanced, and is therefore not yet accurate enough. However, developments in this area are expected to progress rapidly.[54]

Because visual material is likely to play an increasingly important role on internet – see for example the popularity of YouTube, Instagram and TikTok – producers of disinformation are also expected to make frequent use of (manipulated) visual material.

Cheap fakes

As far as the manipulation of visual material is concerned, recent examples show that access to advanced technology is not in fact essential to produce disinformation. 'Cheap fakes' can already cause significant harm. During the last presidential election in Brazil, for example, considerable social unrest was caused with a low-tech resource, Photoshop (see figure 3).[55]

[53] Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D. B., Genova, K., Jin, Z., Theobalt, C., & Agrawala, M. (2019). Text-based Editing of Talking-head Video. *arXiv:1906.01524 [cs].* http://arxiv.org/abs/1906.01524

[54] Volkert, R. (2020). *Deep Fakes: Understanding the illicit economy for synthetic media.* NISOS https://cdn2.hubspot.net/hubfs/6068438/Resources/NISOS%20-%20Deep%20Fakes%20White%20Paper.pdf

[55] Panontin Scarabelli, A. (2018). *How did fake news run voters' opinions in the Brazilian elections.* Diggit Magazine www.diggitmagazine.com/articles/fake-news-brazilian-elections

Image manipulation leads to violence in Brazil

The picture shows the son of Brazilian presidential candidate Bolsonaro during the election campaign. On the left is the edited image, on the right the original. The following text has been superimposed on the photo: 'Moviment nodestinos [an ethnic group from the north-east of Brazil] go home. Rio [de Janeiro] is no place for donkeys'. The image was widely disseminated on social media, even after it was revealed that it had been Photoshopped. The image led to furious discussions, and even fights, between people from different regions of Brazil.

Figure 3 Image manipulation leads to violence in Brazil (copyright Aos Fatos Org).

4.4 Memes

Memes are images that, sometimes manipulated or with a caption added, try to convey an often humorous or satirical message. They are designed to be shared on social media.[56] In this study, we regard political memes as a marginal phenomenon. As well as for humorous or satirical effect, memes can also consciously be used to cause harm. But it is often difficult to say whether a meme was only intended for entertainment, or also to cause harm. During the presidential election in the United States in 2016, memes were also used to influence the voting behaviour of users of social media.

Memes can be consciously used to spread disinformation. After all, their simplified version of reality can easily distort or conflict with that reality. The combination of humour and visualisation can be a powerful instrument for influencing a person's view of the world.[57]

Memes are increasingly produced and disseminated on dedicated platforms, such as Giphy. Platforms for social media and chat apps could integrate these meme platforms into their services and thus further expand their reach.

[56] Rushkoff, D., Pescovitz, D., & Dunaga, J. (2018). *THE BIOLOGY OF DISINFORMATION: memes, media viruses, and cultural inoculation*. Institute for the Future. *www.iftf.org/fileadmin/user_upload/images/ourwork/digintel/IFTF_biology_of_disinformation_062718.pdf*
[57] Klein, O. (2018). *Manipulative Memes : How Internet Memes Can Distort the Truth – Connected Life Conference*. Oxford Internet Institute https://connectedlife.oii.ox.ac.uk/manipulative-memes-how-internet-memes-can-distort-the-truth/

4.5 Augmented and virtual reality and avatars

With applications of augmented and virtual reality (AR/VR), information is wholly or partially presented in all or part of the user's field of view. This usually involves the use of a special AR or VR headset, which contains a screen or a projector. For AR applications the image only covers part of the user's field of view, or the headset is also fitted with a camera so that part of the real world remains visible. With VR applications, the entire image is artificial.

AR and VR technology creates new possibilities for analysing the user's behaviour. For example, with AR/VR glasses the user's eye movements can be monitored and a user's response to images can be analysed on the basis of the pupillary reflex. On that basis, users can be presented with information tailored to their conduct and preferences. That can include disinformation geared to their personal characteristics, which could, for example, reinforce racial prejudices.[58]

VR technology has existed for decades, but has not yet been embraced by a large part of the Dutch population. However, the technology is becoming more affordable, large technology companies are developing more and more applications, and with new algorithms users can communicate with other people in a fairly natural manner while wearing a VR headset. An example of this is FaceVR, a technology that combines image manipulation and VR technology, whereby the face of the user of a VR headset is reconstructed. In this way, VR users can make video calls to each other while wearing a headset and still see each other's face (see figure 4).[59] Herein lies a risk of manipulation of the technology with real-time deepfake algorithms.

[58] Rose, J. (2016). *The Dark Side of VR.* The Intercept https://theintercept.com/2016/12/23/virtual-reality-allows-the-most-detailed-intimate-digital-surveillance-yet/
[59] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2018). FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. *ArXiv:1610.03151 [Cs]*. http://arxiv.org/abs/1610.03151

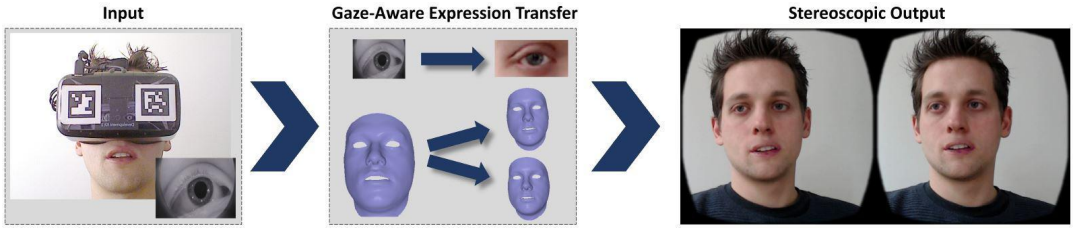

Figure 4 FaceVR: on the left the user wearing VR glasses, on the right the images the recipient sees.

There are a growing number of technological options for generating a three-dimensional lookalike avatar of a person in VR.[60] The emergence of this new generation of avatars is driven by the new possibilities afforded by the depth sensing camera in the iPhone X, the light radar technology in the latest iPad[61] and real-time face tracking, among others.[62]

Technology companies are investing heavily in the use of AR and VR. For example, Apple is expected to launch a VR system within a year.[63] And Facebook expects to launch its new VR environment Horizon in 2020.[64] Facebook suggests that in this virtual environment it will be possible to create an avatar from a 3D scan of your body (see figure 5).[65] Hacking a person's avatar, by copying or taking it over, could be used to produce and disseminate disinformation.

[60] Dempsey, M. (2018). *Avatar-First Products & Platforms*. Medium. https://medium.com/@mhdempsey/avatar-first-products-platforms-723fd637bd35
[61] Apple (2020). *Apple unveils new iPad Pro with LiDAR Scanner and trackpad support in iPadOS.* www.apple.com/newsroom/2020/03/apple-unveils-new-ipad-pro-with-lidar-scanner-and-trackpad-support-in-ipados/
[62] Gibbs, S. (2020). *Apple unveils iPad Pro with 3D scanner in major redesign.* The Guardian www.theguardian.com/technology/2020/mar/18/apple-unveils-ipad-pro-with-3d-scanner-in-major-redesign
[63] Gurman, M. (2019). *Apple Plans Standalone AR and VR Gaming Headset by 2022 and Glasses Later*. Bloomberg. www.bloomberg.com/news/articles/2019-11-11/apple-s-ar-push-will-start-with-ipad-and-culminate-with-glasses
[64] Oculus (2020). *Facebook Horizon* www.oculus.com/facebookhorizon/
[65] Facebook (2019). *Facebook is building the future of connection with lifelike avatars.* https://tech.fb.com/codec-avatars-facebook-reality-labs/

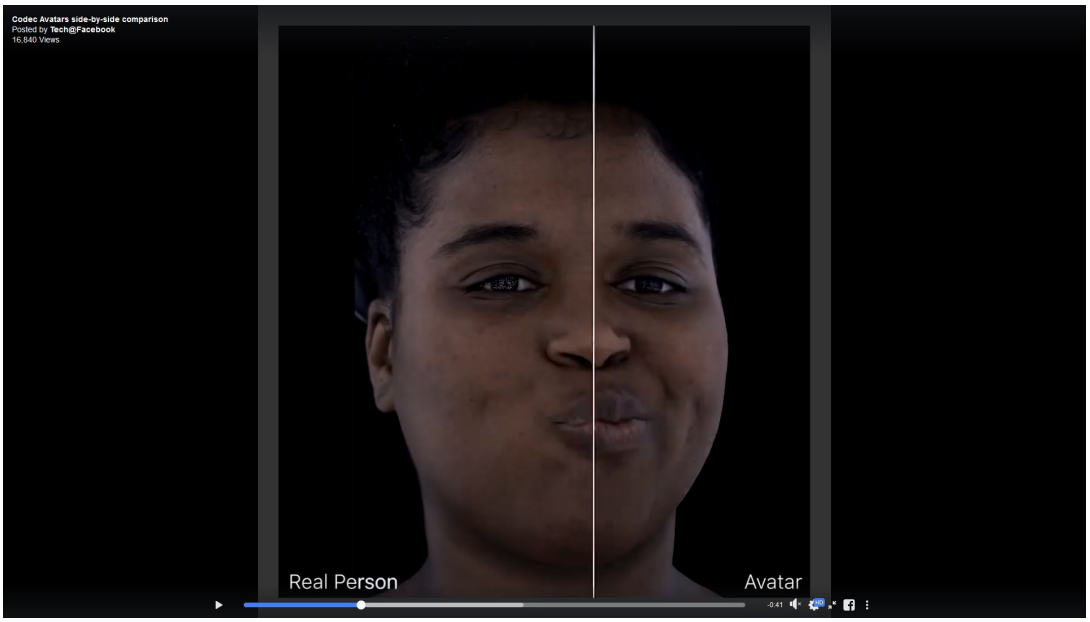

Figure 5 Facebook Codec Avatar. This technology uses very realistic 3D scans of users as an avatar in VR, so that you can be ‘yourself’ in VR. This image is a screenshot of a video which shows how closely the avatar (right section) resembles a real person (left section).

5 Dissemination technologies

In this chapter we describe technological developments that are likely to be relevant for the dissemination of disinformation in the coming years. The following technologies are discussed:

- Micro-targeting
- Chat apps
- Bots
- Search engines
- Digital assistants
- Distributed Autonomous Applications
- Games
- Cross-media storytelling

Because of the large part they play in the dissemination of disinformation, we also discuss the role of social media platforms.

5.1 Social media platforms

Social media platforms bring enormous numbers of people together and into contact with a wide range of information sources. A wide range of social media platforms are used in the Netherlands. They include popular platforms such as Facebook, YouTube, Instagram, Twitter, TikTok and LinkedIn, but also numerous smaller platforms that meet particular information needs or target specific groups, such as Tumblr, Reddit, Flickr, Medium, Spotify and Pinterest.

Users can share all sorts of information on these platforms. They frequently share written messages and photos, but increasingly also videos. The relatively new platform TikTok (formally Musicaly), which is devoted exclusively to the sharing of videos, is rapidly gaining in popularity. As soon as a particular type of information becomes more popular on a platform, other platforms often quickly adopt the formula. For example, SnapChat's popular Stories (messages that are only visible temporarily) can now also be found on Instagram, and they are also being tested on Twitter.[66] in other words, a popular new type of information is usually quickly replicated by other parties and so made available to a wider audience.

[66] Wilson, M. (2020). *Twitter is about to become an even bigger weapon of disinformation.* FastCompany www.fastcompany.com/90472066/twitters-new-self-destruct-feature-is-just-another-weapon-of-disinformation

5.1.1 Revenue model of platforms

Social media platforms map users' preferences so that they can provide them with targeted messages and advertisements. The success of the platforms depends on the data they have collected about their users and the quality of their recommendation algorithms. The sale of advertising space is generally their principal source of income. The more successful a platform is in bringing advertisers into contact with potential customers, the larger its turnover.

Technological developments such as micro-targeting (see 5.2) enable social media platforms to gather, aggregate and analyse even more user data with the aim of creating even better advertising profiles of users. The growing use of various smart internet-connected devices (the Internet of Things) further enables social media platforms to combine data derived from online surfing behaviour with data from offline behaviour that is monitored with sensors. It is therefore conceivable that the temperature setting in the thermostat in the home can influence the clothing shown in advertisements or that the presence of solar panels is used to estimate a user's political preference.[67]

5.1.2 Filter bubbles

The information presented to users of social media platforms is usually selected by recommendation algorithms. The information is geared to the user's preferences and analyses of other data known about users. It is not known precisely how the recommendation algorithms work because they are regarded as trade secrets by social media platforms.

The use of recommendation algorithms can create filter bubbles, or echo chambers. Because most of the information users receive is specifically tailored to their personal characteristics, they are presented with a limited view of reality – a view that often corresponds with their existing preferences and opinions. This can assume innocent forms, such as reporting geared to an interest the user has previously shown in news about sport, but can also lead to narrow, biased views on social and political issues. If this occurs on a large scale, it can lead to disruption of public debate because people are no longer confronted with alternative opinions and points of view.[68]

---

[67] TacticalTech. (2019). *Personal Data: Political Persuasion - The Guidebook and Visual Gallery*. https://ourdataourselves.tacticaltech.org/posts/inside-the-influence-industry
[68] Pariser, E. (2012). *The filter bubble: what the Internet is hiding from you*. London: Penguin Books

According to research carried out by the Institute for Information Law (IViR) for the Dutch Media Authority, there is no evidence of the existence of filter bubbles in the Netherlands at present.[69] But the institute did identify risk factors suggesting that this could change in the near future. The majority of the Dutch population still use a wide range of information sources, including television, newspapers, radio and internet. This varied news consumption is seen as a positive factor in the fight against disinformation, but there are now major differences between age groups in how they use media. Young people in particular use social media a lot, also for news. The question is what effect that will have on how they deal with misleading reporting.

5.1.3 Radicalisation and polarisation

In general, the recommendation algorithms of social media companies are aimed at retaining the user's attention for as long as possible.[70] This generally means that messages, photos or videos with more sensational content are assigned a higher ranking. It seems very likely that recommendation algorithms consequently draw more attention to messages with radicalising or polarising content than to other messages.[71]

For example, de Volkskrant and the Correspondent concluded on the basis of their own research that YouTube's recommendation algorithm encourages viewers to watch increasingly radical videos. They also found that extreme-right messages are relatively over-represented in the videos on YouTube.[72] [73] Research by the University of Amsterdam supports that finding.[74]

Malicious parties that wish to expand the reach of information intended to cause radicalisation or polarisation can use this (assumed) effect of recommendation algorithms. For example, the aforementioned research by the University of Amsterdam pointed to a 'growing number of tendentious and highly partisan news-

[69] Commissariaat voor de Media. (2019). *Filterbubbels in Nederland*. www.mediamonitor.nl/analyse-verdieping/filterbubbels-in-nederland-2019/.
[70] European Data Protection Supervisor (2018). *EDPS Opinion on online manipulation and personal data* https://edps.europa.eu/sites/edp/files/publication/18-03-19_online_manipulation_en.pdf
[71] Quinn, B., Blackall, M., & Dodd, V. (2020). *YouTube accused of being 'organ of radicalisation*.' The Guardian. www.theguardian.com/technology/2020/mar/02/youtube-accused-of-being-organ-of-radicalisation
[72] Bahara, H., Kranenberg, A., & Tokmetzis, D. (2019). *Hoe YouTube rechtse radicalisering in de hand werkt*. Volkskrant. www.volkskrant.nl/kijkverder/v/2019/hoe-youtube-rechtse-radicalisering-in-de-hand-werkt
[73] Tokmetzis, D., Bahara, H., & Kranenberg, A. (2019). *Aanbevolen voor jou op YouTube: racisme, vrouwenhaat en antisemitisme*. De Correspondent. https://decorrespondent.nl/9149/aanbevolen-voor-jou-op-youtube-racisme-vrouwenhaat-en-antisemitisme/445528853-0f710148
[74] Rogers, R., & Niederer, S. (2019). *Politiek en Sociale Media Manipulatie*. Universiteit van Amsterdam. www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2019/10/18/rapport-politiek-en-sociale-media-manipulatie/rapport-politiek-en-sociale-media-manipulatie.pdf

like organisations' in the Netherlands, within which trolls operate and artificial amplification is used.[75]

Also relevant in this context is the growing popularity of online streaming via YouTube, TikTok, SnapChat, InstagramTV and Twitch. At present, the most popular items are pre-produced videos, so-called vlogs, but it is conceivable that greater use will be made of live streams in future. Social media companies can still counter disinformation by filtering videos, for example by checking them before they are published. But the situation is different with livestreams, because the message then reaches the recipient immediately and filtering or fact-checking is difficult.

5.2 Micro-targeting

Micro-targeting enables senders of information to reach a precise target group with a message tailored to them. The technology required for this is often developed for commercial marketing purposes. As soon as the technology appears on the market, it can also be used to disseminate disinformation.[76]

Micro-targeting changes the possibilities for spreading disinformation in three ways. First, the collection of data and the compilation of advertising profiles are automated, so they can be applied on a far greater scale. Second, the technology is increasingly capable of determining what target group(s) a person belongs to. The system can then automatically select the best channels to use to reach that person. Third, with micro-targeting the content of messages can be automatically tailored to the recipient.[77]

Many of the examples of the use of micro-targeting for the dissemination of disinformation described in the literature relate to the United States. But the practices can also occur in Europe.[78] However, an important difference is the European legislation relating to data protection (the General Data Protection Regulation, GDPR), which imposes constraints on the collection of personal data. For example, data about political preferences may not be collected without the data

[75] Idem
[76] Kreling, T., & Modderkolk, H. (2020). *Hoe Spaanse software (onbedoeld) een gevaarlijk wapen werd voor online beïnvloeding*. Volkskrant. www.volkskrant.nl/nieuws-achtergrond/hoe-spaanse-software-onbedoeld-een-gevaarlijk-wapen-werd-voor-online-beinvloeding~b135b1bb/
[77] Crain & Nadler (2019). Political Manipulation and Internet Advertising Infrastructure. *Journal of Information Policy*, *9*, 370. https://doi.org/10.5325/jinfopoli.9.2019.0370
[78] Bennett, C. J. (2016). Voter databases, micro-targeting, and data protection law: can political parties campaign in Europe as they do in North America? *International Data Privacy Law*, *6*(4), 261–275. https://doi.org/10.1093/idpl/ipw021

subject's consent. In addition, campaigns using micro-targeting are currently expensive and are therefore not accessible to everyone.[79]

Various applications of micro-targeting are discussed below.

5.2.1 Campaign software

There is a wide range of dissemination technologies that can be used to reach a precisely defined audience, for example during political campaigns. Campaign software can be used to coordinate the use of various methods. It makes campaigns more efficient and more effective, for example by analysing multiple social media networks simultaneously to estimate what information will have the greatest impact on each particular target group. On that basis, decisions can be made about the use of offline campaign activities, such as public and media appearances, advertisements, flyers, door-to-door canvassing or recruiting members. Campaign software automates this process and uses artificial intelligence to estimate the effects of the various activities in advance. In this way, campaign software forms the cockpit for the dissemination of political campaign material.

Highly advanced specialist software is available for this process in the United States, such as CampaignGrid. This software supports large-scale data collection and analysis, and divides the country into virtual regions. The software is also used to register the results of campaign activities. As far as is known, this form of technology-driven campaigning does not yet occur in the Netherlands.

As with the technology developed for commercial marketing purposes, campaign software can also be used to spread disinformation (for political or other reasons).

5.2.2 AdTech

AdTech stands for advertising technology. The title encompasses a variety of micro-targeting technologies that can be used for advertising purposes. The dissemination of advertisements can be an effective way of spreading

[79] Dommett, K. (2019). Data-driven political campaigns in practice: understanding and regulating diverse data-driven campaigns. *Internet Policy Review*, *8*(4). https://policyreview.info/articles/analysis/data-driven-political-campaigns-practice-understanding-and-regulating-diverse-data

disinformation because the technology gives the sender control over the reach and the message.[80]

The worldwide advertising market is estimated to be worth 327 billion dollars.[81] According to PwC, the online advertising market in the Netherlands is worth over two billion euros, twice as much as the combined TV and radio advertising market. Facebook and Google dominate the online advertising market in the Netherlands.[82]

The following section describes two developments in the area of AdTech that could have a huge influence in the near future: dynamic prospecting and programmatic advertising.

Dynamic prospecting

Reaching the correct target group is one of the keys to an effective advertising campaign. Facebook helps marketeers to achieve this by giving them access to categorised user data. To this end, Facebook collects tens of thousands of characteristics of every user.[83] [84] These characteristics are derived from sources such as the messages the users post, their network of friends or facial recognition data from photos and videos. Since 2.5 billion people use Facebook every month, this is an enormous dataset.

Dynamic prospecting is used to select target groups automatically from this dataset, to estimate and analyse the effect of a particular advertisement and to revise target groups. Dynamic prospecting is used in political campaigns in the US.[85] Through the use of self-learning algorithms, this application of artificial intelligence is expected to improve even further in the future, for example to produce even more precisely defined target groups. In 2017, leaked documents revealed that Facebook bases its selection of target groups in part on the emotional state of teenagers, so that advertisements could be geared to a person's feelings, such as 'worthlessness', 'uncertainty' or 'anxiety'.[86]

---

[80] Kim, Y. M., Hsu, J., Neiman, D., Kou, C., Bankston, L., Kim, S. Y., Heinrich, R., Baragwanath, R., & Raskutti, G. (2018). The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on Facebook. *Political Communication*, *35*(4), 515–541. https://doi.org/10.1080/10584609.2018.1476425

[81] Crain & Nadler (2019). Political Manipulation and Internet Advertising Infrastructure. *Journal of Information Policy*, *9*, 370. https://doi.org/10.5325/jinfopoli.9.2019.0370

[82] Consultancy.nl (2019). *Reclame-inkomsten tv en radio steeds verder achterop bij internet*. www.consultancy.nl/nieuws/25963/reclame-inkomsten-tv-en-radio-steeds-verder-achterop-bij-internet

[83] Tobin, J. A., Madeleine Varner,Ariana. (2017). *Facebook Enabled Advertisers to Reach 'Jew Haters'*. ProPublica. www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters

[84] Dean, S. (2019). Facebook decided which users are interested in Nazis — and let advertisers target them directly. Los Angeles Times. www.latimes.com/business/technology/la-fi-tn-facebook-nazi-metal-ads-20190221-story.html

[85] Montgomery, K., & Chester, J. (2020). *The digital commercialisation of US politics — 2020 and beyond.* Center for Digital Democracy. www.democraticmedia.org/article/digital-commercialisation-us-politics-2020-and-beyond

[86] Reilly, M. (2017). *Is Facebook Targeting Ads at Sad Teens?* MIT Technology Review. www.technologyreview.com/2017/05/01/105987/is-facebook-targeting-ads-at-sad-teens/

The algorithms used for dynamic prospecting are not public because the information is regarded as competitively sensitive. Nor is there any independent supervision of their use. This lack of transparency makes it difficult to assess how legitimate their use is and whether, for example, they are used to disseminate disinformation.

In view of the potential of the use of dynamic prospecting to provide specific information to particular target groups and the lack of transparency about how precisely it is used, this technology lends itself to the dissemination of disinformation.

Programmatic advertising

Not only can the composition of target groups be revised automatically, the content of advertisements can also be tailored increasingly precisely to the recipient. There are various names for this technology: 'programmatic advertising', 'creative versioning', 'dynamic creative' and 'dynamic creative optimisation'.[87] The algorithms used for this method can modify the content of a message according to the recipient's response to it.

For example, Google offers advertisers the Directors Mix and Vogon applications, with which hundreds of versions of a video can be produced automatically.[88] Each video contains a unique combination of image and text, including variation in the font size and the positioning of the text.[89] Netflix, for example, can use this technology to promote the same television series as a special effects spectacle, a family drama or a love story – depending on the recipient of the advertisement.[90] Facebook has a similar dynamic creative environment.[91]

[87] Montgomery, K., & Chester, J. (2020). *The digital commercialisation of US politics — 2020 and beyond*. Center for Digital Democracy. www.democraticmedia.org/article/digital-commercialisation-us-politics-2020-and-beyond
[88] Jain, K., & Chetan, A. (2018). *What brands can learn from India on personalized storytelling*. Think with Google. www.thinkwithgoogle.com/intl/en-apac/ad-channel/video/what-brands-can-learn-india-personalized-storytelling/
[89] Google (2020). *Project Vogon* https://opensource.google/projects/vogon
[90] Rothwell, J. (2018). *Perspectives: Find your audience on digital and storytell with data*. Think with Google www.thinkwithgoogle.com/intl/en-apac/tools-resources/success-stories/perspectives-find-your-audience-digital-and-storytell-data/
[91] Facebook (z.d.). *Dynamic Creative www.facebook.com/business/m/facebook-dynamic-creative-ads*

Figure 6 In YouTube's Directors Mix, the image, sound and text in an advertisement can be modified according to the target group. The version on the left is for an American audience; the one on the right for a Dutch audience.[92]

This technology was used on a small scale during the American presidential election in 2016. For example, it was reported that 40,000 to 50,000 different versions of one advertisement were generated a day.[93] It is expected that the number that can be generated will be much larger in the near future.

Even if these advertisements were to be recorded in a public register of political advertisements, it would still probably be difficult, for journalists for example, to investigate how a political party is portraying itself to particular audiences because of the enormous number of advertisements.

A potential risk with this technology is that political parties will be able to target different groups with different political messages, making it difficult for the recipient to determine precisely what the party's actual position is.

5.2.3 Psychographing

Psychographs divide people into target groups on the basis of their character traits. The basic idea is that marketers can use them to interest consumers in a product by appealing to their personal values and desires. Cola, for example, is not promoted as a thirst quencher, but as a feel-good product, so that consumers drink it not only because they are thirsty, but also to cheer themselves up.

Naturally, this technology can also be used by producers and disseminators of disinformation to ensure that their message matches the feelings of the envisaged target group as closely as possible. Cambridge Analytica, for example, says it

[92] Newfangled (z.d.). *Google: Director Mix* www.newfangledstudios.com/projects/google-directormix/
[93] TacticalTech. (2019). *Personal Data: Political Persuasion - The Guidebook and Visual Gallery.* https://ourdataourselves.tacticaltech.org/posts/inside-the-influence-industry

creates its target groups on the basis of the five characteristics in their OCEAN model: Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism. Political messages can be tailored to an audience on the basis of the scores on each of these five personality traits.[94]

Figure 7 Former Cambridge Analytica director Alexander Nix demonstrates how the content of political advertisements is determined by the OCEAN model.[95]

5.2.4 Influencer marketing

Influencer marketing can be seen as a modern-day version of word-of-mouth advertising. Managers of social media accounts with a great many followers are paid to advertise a product, a service or a brand. It is not always clear to the recipients that what they are seeing is an advert, for example because the senders also post numerous non-commercial messages, usually with a view to attracting new followers.

Influencers are often well-known figures, like star footballer Cristiano Ronaldo who has more than 200 million followers on Instagram[96]. A message from Ronaldo usually generates around five million reactions. But not all influencers are international celebrities. There are also agencies that can arrange for influencers with just a few thousand followers to disseminate information. Influencers are also

[94] Concordia. (2016). Cambridge Analytica - The Power of Big Data and Psychographics. www.youtube.com/watch?v=n8Dd5aVXLCc
[95] Idem
[96] www.instagram.com/cristiano/

not necessarily persons. A brand can also build up a huge following – Nike, for example, has a hundred million followers on Instagram.[97]

Virtual and political influencers

A relatively new phenomenon is that of virtual influencers. These are accounts in which a virtual person promotes a product or service. Lil Miquela, with two million followers on Instagram, is one such virtual influencer.[98] She mainly promotes clothing and lifestyle brands on Instagram with photos in which she is often portrayed alongside actual artists.

Figure 8 Lil Miquela in Los Angeles.[99]

Influencers can also be politically active.[100] They are known as political influencers. Marketing firm Drawbridge (no longer active) provided a service called Political Influencer Identification, which classified influencers and their followers on the basis of political preferences with the aim of deploying them in campaigns. Twitter removed Drawbridge for violation of its terms of service.

Micro-influencer marketing

Micro-targeting techniques are increasingly used in influencer marketing. Advertisers want to use their resources as effectively as possible and avoid spending money to reach the wrong target group. To that end, the managers of social media accounts monitor the followers of their accounts ever more closely. Advertisers then instruct the managers of the accounts to disseminate specific information among those followers who match their target group most closely. The

[97] www.instagram.com/nike/
[98] www.instagram.com/lilmiquela/
[99] www.instagram.com/p/B82j0y-Hcia/
[100] Chester, J., & Montgomery, K. C. (2017). The role of digital marketing in political campaigns. *Internet Policy Review*, *6*(4). https://policyreview.info/articles/analysis/role-digital-marketing-political-campaigns

followers of these accounts may be few in number, but they possess specific characteristics.

It is common knowledge that the information in messages from influencers with a lot of followers is partly determined by advertisers, although that is often not explicitly mentioned. In the case of micro-influencer marketing, where the number of followers is small, it is suspected that followers are less wary of being influenced, since the message comes across as a genuine recommendation of a product or pronouncement of a (political) viewpoint from a close friend or acquaintance.

The lack of transparency with influencer marketing in general, and micro-influencer marketing in particular, makes this method of communication particularly useful for disseminating disinformation. The disseminators of disinformation can give very precise instructions, without the target group realising who is providing the information or why.

5.3 Chat apps

Chatting is also known as instant messaging or direct messaging. The name itself conveys an important property of chats. Information exchange in chats is faster and usually more personal than communication on social media platforms like Facebook and Twitter. Many chat apps arrange the messages in chronological order, without any influence from a recommendation algorithm.

Chat apps are very popular in the Netherlands – particularly WhatsApp, but also SnapChat, Telegram, FaceTime, Google Hangouts, Skype, Slack and Signal. Many social media platforms also offer internal chat functions, such as Facebook Messenger and Twitter Direct Messaging. Gamers also chat a lot, via apps such as Discord or directly in games such as FortNite.

Chat apps generally offer the option of recording and sending text, audio and video messages. Many contain functions for sending emojis, animated images (gifs), memes, stickers and other multimedia items, such as YouTube videos. Live video calls can be held on some chat apps, sometimes with the possibility of directly manipulating images with the help of filters.

Groups or channels are an important functionality of chat apps. They allow information to be sent to large numbers of users simultaneously. WhatsApp has limited the number of users in a group to 256. The maximum size of a group on Telegram is 200,000, but there is no limit to the number in channels. In some countries, Telegram channels have millions of participants. In Iran, for example,

these channels are used to exchange news of current events.[101] Chat apps are in fact the most important source of news for citizens in Brazil. News media in the Netherlands make only limited use of chat apps, but one that does is RTL Nieuws, which broadcasts audio news reports on WhatsApp.

It goes without saying that chat apps are an attractive medium for producers and disseminators of disinformation to spread their message. Since many groups and channels are closed, there is less chance of assertions made on them being refuted.

API for data gathering

In the controversy surrounding Cambridge Analytica, Facebook was criticised for facilitating Cambridge Analytica in gathering datasets. With a so-called application programming interface (API), Facebook enabled parties like Cambridge Analytica to collect large quantities of data from Facebook automatically, often without the knowledge of the data subjects. Following the scandal with Cambridge Analytica, Facebook restricted these possibilities.

However, various chat apps still offer this option. For example, Telegram has an extensive API with which managers of channels can collect information about participants and automatically post messages (there is more on this subject in section 5.4).[102] Signal has a similar programmable interface.[103] Inherent to these interfaces is the risk of a repetition of the data-gathering practices that occurred with Facebook.

Growing use of encryption technology

A notable development with chat apps is the use of encryption. Encryption prevents anyone who does not possess the right key from reading the information. While some chat apps only encrypt the content of messages, others even make it impossible to discover who was chatting with whom and when. This often involves the use of end-to-end encryption, which means that some or all of the information the platform transmits between the sender and the recipient is concealed from the platform's managers.

The rise of encryption is welcomed by many privacy and security experts, who regard chat apps with encryption as safer than a technology like e-mail. In their

[101] Telegram Channels (no date). *The Biggest 100 Media* https://telegramchannels.me/list/biggest
[102] Aichara, D. C. (2019). *Telegram Channel Data Extraction (User's information, chats, and specific messages) and Data Processing.* Medium. https://medium.com/game-of-data/telegram-channel-data-extraction-users-information-chats-and-specific-messages-and-data-21bb54710fd3
[103] Signal App (2014). *API Protocol* https://github.com/signalapp/Signal-Server/wiki/API-Protocol

view, chat apps with extensive encryption are preferable. The staff of the European Commission are obliged to use Signal, for example.[104]

However, encryption also makes it more difficult to identify malicious parties, such as disseminators of disinformation.[105] Another effect of encryption technology is that providers of chat apps have no knowledge of how the application is being used or of the content of messages and can therefore not be held accountable for them.

5.4 Bots

A bot is a social media or chat account that is run automatically by an algorithm, often largely without human action. Bots are increasingly capable of creating information and interacting with humans, which means that it is often unclear to the latter that they are communicating with a bot. Bots can also be used to gather information, for example through participation in a group or channel on social media or in chat apps.

Bots can be used to spread disinformation in various ways. In the first place, by posting disinformation on social media platforms. Bots can then engage in various interactions that are offered on the platform, such as liking, sharing or commenting on a message.[106] With these interactions, a bot can influence a social media platform's recommendation algorithm in such a way as to affect the frequency with which messages are shown. For the same reason, bots can be used to artificially increase the number of followers of an account. Bots can also frequently use hashtags to influence trending topics or to take command of a hashtag discussion.

If bots are discovered, the manager of a platform has the power to remove a message or an account from the platform if the use of bots is contrary to the platform's terms of service. However, malicious individuals can also take advantage of that power by having bots interact with messages or accounts that they wish to silence.

[104] Cerulus, L. (2020*). EU Commission to staff: Switch to Signal messaging app*. Politico www.politico.eu/pro/eu-commission-to-staff-switch-to-signal-messaging-app/
[105] Nieuwsuur (2020). *'Nu IS van Telegram is verwijderd, zijn ze moeilijker in de gaten te houden.'* https://nos.nl/l/2318257
[106] Hussain, M. N., Tokdemir, S., Agarwal, N., & Al-Khateeb, S. (2018). Analyzing Disinformation and Crowd Manipulation Tactics on YouTube. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1092–1095. https://doi.org/10.1109/ASONAM.2018.8508766

Bots are used on every social media platform. It is estimated that as many as 15% of all Twitter accounts are bots. According to Politicalbots.org, roughly 19 million bot accounts were active in the final week before the US presidential election in 2016.

The impact of bots is not confined to the social media platforms on which they are active. Messages from bots are sometimes used as *vox populi* in news reports from the traditional media. For example, after a large number of bots of the Russian troll factory IRA were unmasked, it was found that traditional media had adopted reports from tweets written by bots.[107] Major Norwegian news media had also adopted reports from IRA bots in their reporting in the belief that they were authentic reports from Norwegian Twitter users.[108]

Bots are also often used in combination with other dissemination technologies. For example, the IRA bots were found to be particularly effective in spreading reports on the highly partisan Russian news medium Russia Today.[109]

Recommendation algorithms can in fact also be influenced by human-controlled accounts. For example, the South Korean intelligence service was found to be manually controlling Twitter accounts in order to influence political sentiment in the country.[110] Marketing firms also generally use human-controlled accounts to influence recommendation algorithms. In a country like Spain, it only takes a few hundred accounts to have an impact on Twitter.[111]

Chatbots

A chatbot is a specific type of bot that chats directly with humans. They are best known from customer service departments. A customer often quickly realises that it is a bot because of the often simplistic manner in which the bot responds, but the technology is improving rapidly. For example, in 2020 Google launched the chatbot Meena, which, according to Google's own scoring system, performs better than other common chatbots and can imitate human interactions increasingly convincingly.[112]

---

[107] Lukito, J., Suk, J., Zhang, Y., Doroshenko, L., Kim, S. J., Su, M.-H., Xia, Y., Freelon, D., & Wells, C. (2019). The Wolves in Sheep's Clothing: How Russia's Internet Research Agency Tweets Appeared in U.S. News as Vox Populi: *The International Journal of Press/Politics.* https://doi.org/10.1177/1940161219895215

[108] NOS (2020). *Alle grote media in Noorwegen trappen in tweets van Russische trollen.* https://nos.nl/l/2325674

[109] Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. *ArXiv:1801.09288 [Cs]*. Retrieved from http://arxiv.org/abs/1801.09288

[110] Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication*, *37*(2), 256–280. . https://doi.org/10.1080/10584609.2019.1661888

[111] Kreling, T., & Modderkolk, H. (2020). *Hoe Spaanse software (onbedoeld) een gevaarlijk wapen werd voor online beïnvloeding.* Volkskrant. www.volkskrant.nl/nieuws-achtergrond/hoe-spaanse-software-onbedoeld-een-gevaarlijk-wapen-werd-voor-online-beinvloeding~b135b1bb/

[112] Adiwardana, D., & Luong, T. (2020). *Towards a Conversational Agent that Can Chat About…Anything*. https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

The expectation is that as chatbot technology advances, malicious parties will increasingly use chatbots to spread disinformation in a (semi-)automated manner in one-on-one conversations with their target.[113]

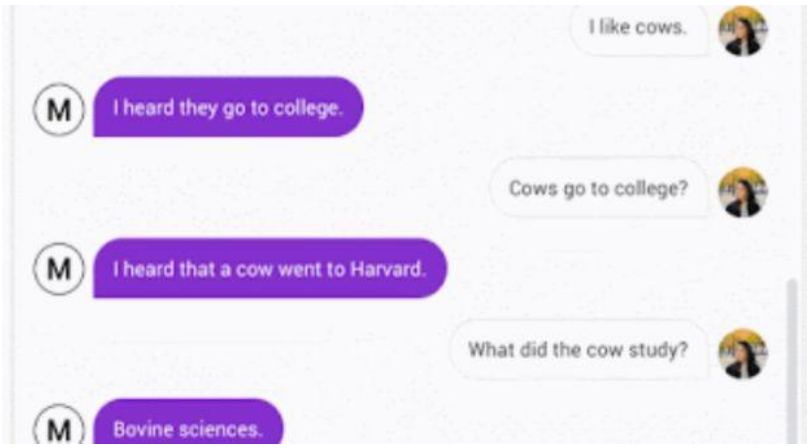

Figure 9 The chatbot Meena displays a certain sense of humour in a demonstration by Google.[114]

5.5 Search engines

Search engines are websites on which users can enter a search command and are then shown a selection of links. Well-known examples are Google Search, Microsoft Bing and DuckDuckGo. Search engines are also increasingly used as question-and-answer machines. Users can formulate their search command in the form of a question, whereupon the search engine looks for possible answers. Users of search engines usually choose higher-ranked results over results lower down the list. Instead of working solely with written commands, search engines are also increasingly capable of searching for information on the basis of a spoken command, an image or a video.

Users of social media platforms like Facebook, Instagram and YouTube also make growing use of (internal) search engines to navigate through information.

Because they are used so frequently, search engines play an important role in the dissemination of information. In the Netherlands, users also have more confidence

[113] TacticalTech. (2019). *Personal Data: Political Persuasion - The Guidebook and Visual Gallery*. https://ourdataourselves.tacticaltech.org/posts/inside-the-influence-industry
[114] Schwartz, E. (2020). *Google's New Meena Chatbot Imitates Human Conversation and Bad Jokes. Voicebot*. https://voicebot.ai/2020/02/03/googles-new-meena-chatbot-imitates-human-conversation-and-bad-jokes/

in news reports that are recommended by search engines than in information they find on social media. But the results of a search command are not always reliable, because the algorithms with which search engines operate can be manipulated.[115] Search engines also pick up reports that are popular on social media, while those reports can themselves be influenced.[116]

Search engine technology can be combined with imaging technology. For example, experiments by Amazon with the aforementioned generative adversarial networks (GANs) have shown that appropriate images can be generated from a written search command. An image of a dish of food could be generated from a list of ingredients, for instance.[117] [118]

Because visual material is playing an increasingly important role on the internet, and particularly on social media platforms, the use of image recognition and image synthesis algorithms (such as GANs) by search engines is likely to become far more common. That will also create new opportunities for spreading disinformation.

5.6 Virtual assistants

Voice-activated digital assistants – or virtual assistants – are sources of information that can be operated by the human voice. Well-known examples are Amazon's Alexa, Apple's Siri and Google Assistant, which can be asked for a weather report or for answers to various questions. Virtual assistants can be physical appliances installed in the home, but can also be a function of devices such as smartwatches, smartphones, laptops or smart TVs.

In 2019, five percent of the Dutch population had a virtual assistant in the home. That number is expected to grow rapidly in the coming years.[119] According to reports by the National Listener Survey, 34% of Dutch people use the virtual assistant to listen to the news.[120] The number of virtual assistants is also growing

[115] Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, *112*(33), E4512–E4521. https://doi.org/10.1073/pnas.1419828112
[116] Robertson, A. (2017). *It's time to stop trusting Google search already*. The Verge www.theverge.com/2017/11/10/16633574/stop-trusting-google-search-texas-shooting-twitter-misinformation
[117] Biswas, A., & Surya, S. (2020). *Converting text to images for product discovery.* Amazon Science Blog www.amazon.science/blog/converting-text-to-images-for-product-discovery
[118] Synced (2020). *CookGAN Generates Realistic Meal Images From an Ingredients List.* https://medium.com/syncedreview/cookgan-generates-realistic-meal-images-from-an-ingredients-list-250426dbfab2
[119] TNS NIPO (2019). *Gebruik smart speakers groeit explosief.* www.tns-nipo.com/nieuws/persberichten/gebruik-smart-speakers-groeit-explosief
[120] Audiomonitor (2019). *Nationaal Luister Onderzoek.* https://nationaalluisteronderzoek.nl/audiomonitor-slides/

rapidly in countries like the US, the UK, Australia and Canada.[121] The growth is accounted for not only by the popular devices from Amazon and Google, but also products of far smaller manufacturers in China.[122]

The use of virtual assistants also has to be considered in the context of disinformation, since the sources of the answers they provide are themselves vulnerable, such as Wikipedia.[123] Conspiracy theories or hoaxes can sometimes go unnoticed for lengthy periods on that platform,[124] so malicious individuals could also use those vulnerable sources to spread disinformation via virtual assistants. Furthermore, as with search engines, the algorithms with which virtual assistants operate can be manipulated.

5.7 Distributed Autonomous Applications

Distributed computing technology enables computers in a network to perform a task without being controlled by a central operator. The computers in the network decide amongst themselves which computer will perform which part of their assigned task. An example is Blockchain, whereby a network of computers maintains a ledger and the computers determine between them whether changes in the ledger account are permitted. A familiar application of this technology is the virtual currency Bitcoin.

Since the introduction of Bitcoin in 2008, many new forms of distributed computing have been developed. Distributed alternatives have been developed for various everyday online applications, such as the video platform D-Tube and the blog platform Steem. A characteristic feature of these applications is that once information has been added it is almost impossible to remove, and then only by the joint action of a majority of the computers in the network. However, in the absence of any organisational structure or central control it is very difficult to accomplish in practice.

This absence of a moderator makes this type of application attractive for malicious parties who wish to abuse it, such as disseminators of disinformation. With these

---

[121] Newman, N. et al. (2019). *Reuters Institute Digital News Report 2019.* Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR_2019_FINAL.pdf
[122] Emerce. (2020). *Consument kiest vaker voor Chinese slimme luidspreker.* www.emerce.nl/nieuws/consument-kiest-vaker-chinese-slimme-luidspreker
[123] Kinsella, B. (2019). *Voice Assistants Alexa, Bixby, Google Assistant and Siri Rely on Wikipedia and Yelp to Answer Many Common Questions about Brands.* Voicebot https://voicebot.ai/2019/07/11/voice-assistants-alexa-bixby-google-assistant-and-siri-rely-on-wikipedia-and-yelp-to-answer-many-common-questions-about-brands/
[124] Kumar, S., West, R., & Leskovec, J. (2016). Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. *In Proceedings of the 25th International Conference on World Wide Web (pp. 591–602).* Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883085

platforms they avoid the risk of their information being removed. The platforms also provide a high degree of anonymity for users.[125]

Distributed Autonomous Applications are not very popular. According to website comparison site SimilarWeb, D-Tube had 300,000 visitors in March 2020, just a fraction of the reach of YouTube, which had thirty billion visitors in the same month.[126] Nevertheless, this type of platform could play an important role in certain niches.

Distributed computing technology could in fact also be used to combat disinformation. By documenting original and authentic information in distributed applications, users would be able to identify its source.[127] The possibility of having information (almost) permanently available on internet therefore creates both opportunities and threats with respect to disinformation.

5.8 Games

Digital games are popular in the Netherlands, particularly among young people. According to the Netherlands Youth Institute, 35% of primary school pupils and 27% of 12- to 16-year-olds play a computer game every day.[128]

Games are usually based on fictional scenarios, but like films they can purport to be telling a true story. Games often use historical situations as their context or background, but do not always portray the course of history accurately.[129]

Because of the popularity of games and their exciting and sometimes even addictive nature, they offer interesting possibilities for spreading disinformation or particular narratives. Like films, games often portray stereotypes and highly politically biased story lines, such as the United States as the victor in the battle between good and evil.

[125] Polyakova, A., & Meserole, C. (2018). *Disinformation Wars.* Foreign Policy https://foreignpolicy.com/2018/05/25/disinformation-wars/
[126] Similarweb.com (no date). www.similarweb.com/website/d.tube/ en www.similarweb.com/website/youtube.com/
[127] Huckle, S., & White, M. (2017). Fake News: A Technological Approach to Proving the Origins of Content, Using Blockchains. *Big Data*, 5(4), 356–371. https://doi.org/10.1089/big.2017.0071
[128] NJi. (2019). *Gamen - Cijfers.* www.nji.nl/nl/Databank/Cijfers-over-Jeugd-en-Opvoeding/Cijfers-per-onderwerp/Gamen
[129] Veugen, C. (2014). Using Games to Mediate History. In L. Egberts & K. Bosma (Red.), *Companion to European Heritage Revivals* (pp. 95–111). Springer International Publishing. https://doi.org/10.1007/978-3-319-07770-3_5

Games can also be used to gather personal data, which can then be used for disinformation purposes. For example, Facebook persuaded users to share their personal data with Cambridge Analytica by means of simple games.[130]

Political parties can use games to win over voters or to encourage people to share information. Users of the Trump-2020 app earn points for sharing messages from Trump on Twitter, for instance.[131] This use of game elements could also be applied for dissemination of disinformation.

5.9 Cross-media storytelling

Cross-media storytelling is a method by which a sender can reach a recipient repeatedly via different channels. These channels can be social media platforms, streaming video or chat apps, but also devices such as smartphones, TVs and computers. Cross-media storytelling can therefore combine the strengths of a variety of the technologies described above.

On the one hand, this technology comprises instruments with which individual recipients or target groups can be identified and monitored on the various channels. On the other hand, with the technology the use of those channels can be coordinated in such a way that recipients are constantly being reached.[132] Disseminators of disinformation can exploit these possibilities. And because the same message appears to be coming from different sources, it can appear more credible. But cross-media storytelling also offers other advantages for disseminators of disinformation. For example, research into the IRA found that topics often appeared on Reddit a week earlier than on Twitter. It is suspected that the IRA used Reddit to test the effectiveness of particular messages.[133]

[130] TacticalTech. (2019). *Personal Data: Political Persuasion - The Guidebook and Visual Gallery.* https://ourdataourselves.tacticaltech.org/posts/inside-the-influence-industry
[131] Trump, D. (2020). *Trump 2020 App IS HERE!* www.youtube.com/watch?v=JRuQ5JMMqtM
[132] Chester, J., & Montgomery, K. C. (2017). The role of digital marketing in political campaigns. *Internet Policy Review, 6*(4). https://policyreview.info/articles/analysis/role-digital-marketing-political-campaigns
[133] Lukito, J. (2020). Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017. *Political Communication, 37*(2), 238–255. https://doi.org/10.1080/10584609.2019.1661889

6 Existing measures

This last chapter of the quick scan presents a concise overview of measures that have already been taken to combat the threat posed by disinformation for public debate and the democratic process. We focus on measures adopted by the Dutch government, the European Union and a number of the major platform companies.

6.1 Measures taken by the Dutch government

Objective and basic principles
The Dutch government's policy on disinformation is geared to protecting the stability and quality of the democratic legal order and the open society.

The measures taken by the government to combat disinformation are guided by the following basic principles:

- Constitutional values and fundamental rights such as freedom of speech, press freedom and the right to information are paramount;
- Independent journalism and a pluriform media landscape are essential for a healthy democracy;
- Media literacy and digital literacy are important elements of the strategy to counter the impact of disinformation;
- Citizens must judge the value of information themselves. Transparency about the origin of information is of fundamental importance in that respect;
- Internet services bear their own responsibility. Where their self-regulation falls short, regulation can be considered;
- The development of scientific knowledge about the existence of disinformation is welcomed;
- The government supports coordination at the European and wider international level.[134]

Policy is focused more on mitigating the impact of disinformation than on actively refuting or disproving it. The government feels that the task of actively refuting disinformation lies not with itself, but is primarily the responsibility of non-governmental actors, such as independent media, online platforms and scientists.

[134] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2019). *Kamerbrief over beleidsinzet bescherming democratie tegen desinformatie* www.rijksoverheid.nl/documenten/kamerstukken/2019/10/18/kamerbrief-over-beleidsinzet-bescherming-democratie-tegen-desinformatie

But the government does see a role for itself if political and economic stability or national security is threatened and in communicating its policy to the public.[135]

Three action lines
The measures taken by the Dutch government to counter disinformation are clustered in three action lines:
1. Preventive actions designed to prevent disinformation from having an impact and from spreading;
2. Strengthening the information position in order to provide early warning of (potential) threats;
3. Reactive actions to be taken against disinformation when it appears.

At present, the emphasis of government policy is on preventive measures.

The three action lines are implemented as follows:

Preventive actions
- Strengthening the resilience of citizens against the influence of disinformation with awareness-raising campaigns and by promoting media literacy;
- Strengthening the resilience of political office holders with a game about disinformation and actions they can take to combat it;
- Increasing transparency about disinformation and measures to counter it, for example by monitoring implementation of the EU's code of practice for platform companies;
- Preserving a pluriform media landscape, for example by earmarking additional funds for investigative journalism;
- Innovation in the consumption and production of online news, for example by developing quality standards.

Strengthening the information position
- Improving the information position at national and international level, for example by participating in the EU's Rapid Alert System, through which reports of disinformation campaigns can be quickly shared;
- International collaboration, for example through the European Centre of Excellence on Countering Hybrid Threats;
- Knowledge development.

Reactive actions
- Addressing the content of disinformation – fact-checking – with non-governmental fact-checkers;

[135] Idem

- Refuting disinformation;
- Exploring the possibilities of and responsibilities for the moderation of messages on online platforms.[136] [137]

6.2 Measures taken by the European Union

The European Union has taken various initiatives to combat disinformation. In this section, we confine ourselves to a description of the EU Action Plan against Disinformation and the EU Code of Practice against Disinformation.

EU Action Plan against Disinformation

The EU Action Plan against Disinformation includes the following measures:

- Improving the capabilities of EU institutions and member states to detect, analyse and expose disinformation by investing in digital tools and specialised personnel;
- Strengthening coordinated and joint responses to disinformation campaigns, for example through the aforementioned Rapid Alert System;
- Mobilising the private sector to tackle disinformation, for example by monitoring the implementation of the EU Code of Practice on Disinformation;
- Raising awareness and improving societal resilience, for example by organising campaigns to raise awareness inside and outside the EU and by supporting independent media and fact-checkers.[138]

EU Code of Practice on Disinformation

The EU Code of Practice on Disinformation contains a list of standards for self-regulation drawn up by representatives of platform companies, social networks and advertisers to combat the spread of disinformation.[139] The signatories include Facebook, Google, Twitter, TikTok and Microsoft.

The Code of Practice includes the following guidelines:

- Improve the transparency of political advertising;
- Improve the scrutiny by online platforms of the use of advertisements aimed at spreading disinformation;
- Intensify efforts to remove fake accounts;

---

[136] Idem
[137] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2019). *Actielijnen tegengaan desinformatie.* www.rijksoverheid.nl/documenten/kamerstukken/2019/10/18/actielijnen-tegengaan-desinformatie
[138] European Commission (2018). *Action Plan on Disinformation* https://ec.europa.eu/commission/publications/action-plan-disinformation-commission-contribution-european-council-13-14-december-2018_en
[139] European Commission (2018). *Code of Practice on Disinformation* https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

- Establish a clear marking system for bots to ensure that their activities cannot be confused with human actions;
- Provide access to data for fact-checking and research.

The code of practice is monitored as part of the aforementioned EU Action Plan against Disinformation.

6.3 Measures taken by platform companies

Various platform companies have taken or initiated measures to prevent the production and dissemination of disinformation. In this section, we mention a number of measures taken by some of the major players, before briefly discussing the responses by the companies to the increase in disinformation since the start of the corona pandemic.

One measure taken by Facebook has been to draw up a policy of refusing advertisements containing misleading or fake content. To this end, it has adopted an advertising approval process, in which the images, text and positioning of an advertisement are assessed. Fake accounts are also regularly removed to prevent artificial influencing of the recommendation algorithm and the spread of disinformation. Facebook has also instituted an independent fact-checking programme to detect fake news reports and ensure they are no longer recommended.[140] For the fact-checking of Dutch reports, Facebook works with Agence France Presse (AFP) and Deutsche Presse-Agentur (DPA).[141]

Twitter has formulated an advertising policy that prohibits the spreading of ‘disruptive content’. Advertisers on Twitter must meet certain criteria and are subject to a review process. In 2018, Twitter took measures to prevent fake accounts and spam messages, including more intensive scrutiny of the automated use of the platform by suspected bots. Twitter also provides users with more information about why they are shown particular advertisements.[142]

In 2017, Facebook, Google and Twitter also announced that they would start using trust indicators developed by the Trust Project of the Santa Clara Institute of

[140] Mosseri, A. (2017). *Working to Stop Misinformation and False News.* Facebook www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news
[141] Faccebook (2020). *Update op 26 maart: Facebook kondigt factchecking-partners in Nederland aan* https://facebook.pr.co/187141-corona-nieuwsoverzicht
[142] Twitter (2019). *Twitter Progress Report: Code of Practice on Disinformation*. https://ec.europa.eu/information_society/newsroom/image/document/2019-5/twitter_progress_report_on_code_of_practice_on_disinformation_CF162219-992A-B56C-06126A9E7612E13D_56993.pdf

Applied Ethics.[143] Since then, Facebook has been using fact-checkers to tackle disinformation.[144] Consequently, a message that is assessed as 'untrue' is placed lower in the listings of news reports.[145] Instagram attaches warning labels to reports that are judged to be misleading or inaccurate.[146]

In 2020, TikTok adopted a guideline for tackling political disinformation.[147] The platform is very popular among young people and is used to share political memes, among other things. TikTok also says it combats disinformation campaigns on the platform.

Measures in response to the corona pandemic

Following the flood of misleading reports relating to the corona crisis, platform companies have recently responded to the growing public and political pressure to take tougher action against disinformation.

Twitter, for example, has prohibited messages that contradict recommendations made by health authorities; Reddit has removed disinformation about corona from search results; Google has posted warnings and positioned official information from the WHO at the top of its search results; Facebook has established a Coronavirus Information Center; and Instagram has placed government information at the top of its reporting (see figure 10). Facebook has also donated a million dollars to the International Fact-Checking Network to support fact-checking on WhatsApp. As already mentioned, Facebook Nederland has formed a partnership with Agence France Presse and Deutsche Presse-Agentur to conduct fact-checking.[148]

It can be concluded from these measures that the internet companies feel a greater responsibility than formerly for the quality of the information they disseminate, at least as regards the reporting on the corona crisis. Naturally, it remains to be seen how permanent these measures will be, the risk being that they will be scaled down as soon as the public and political pressure eases again.

---

[143] The Trust Project (2020). *News with integrity* https://thetrustproject.org
[144] Belghmidi, L. (2019). *Facebook bestrijdt samen met 21 Europese organisaties nepnieuws in aanloop naar verkiezingen.* VRT www.vrt.be/vrtnws/nl/2019/04/26/facebook-bestrijdt-samen-met-21-europese-organisaties-nepnieuws/
[145] Kreijeveld, M. (2018). *De strijd tegen nepnieuws (3): Hoe Facebook, Google en Twitter fake news niet kunnen bestrijden*. Marketingfacts www.marketingfacts.nl/berichten/strijd-tegen-nepnieuws-3-hoe-facebook-google-twitter-fake-news-bestrijden
[146] Van Poorten, B. (2020). *Social media update: Snapchat Scan-advertenties en Instagram tegen fake news.* Marketingfacts. www.marketingfacts.nl/berichten/social-media-update-snapchat-scan-advertenties-en-instagram-tegen-fake-news
[147] Nuñez, M. (2020). *TikTok Finally Bans Disinformation Campaigns In Updated Community Guidelines.* Forbeshttps www.forbes.com/sites/mnunez/2020/01/08/tiktok-finally-bans-disinformation-campaigns-in-updated-community-guidelines/
[148] NOS (2020). *Facebook zet voor Nederland factcheckers in om coronanepnieuws te bestrijden.* https://nos.nl/l/2328458

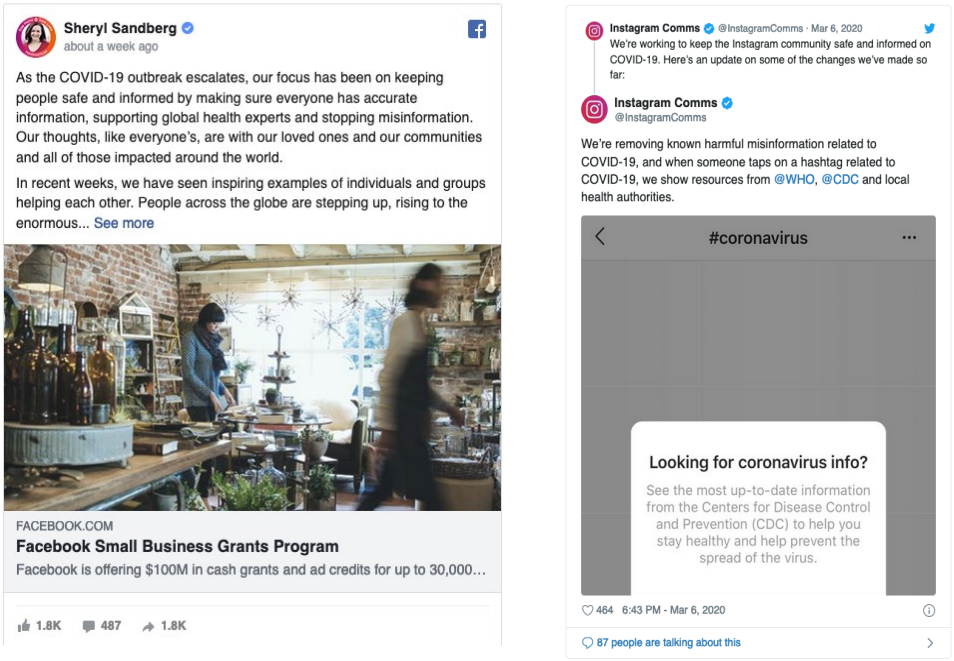

Figure 10 Announcement of measures by social media companies in response to the corona pandemic.

Quite apart from the corona crisis, there has been growing public and political pressure on social media platforms to accept responsibility for the information they disseminate. For instance, the EU's recently amended Audiovisual Media Services Directive (AVMSD) imposes obligations on video platforms to remove illegal content, such as child pornography, provocation to commit a terrorist offence or deception of consumers.[149]

[149] European Commission (2020). *Audiovisual Media Services Directive (AVMSD)* https://ec.europa.eu/digital-single-market/en/audiovisual-media-services-directive-avmsd

Part II Case studies

7 Deepfakes and psychographing

Building on the quick scan in Part I, two case studies were elaborated, one on deepfakes and the other on psychographing. The case studies are intended to provide a more coherent impression of how technology relating to disinformation could develop in the coming years and what impact those developments could have on public debate and the democratic process.

The case studies were chosen on the basis of a review of combinations of technologies:

- that are innovative; in other words, are still developing;
- that are expected to have a major impact;
- that are accompanied by asymmetry in agency between producers and disseminators of disinformation on the one hand, and its recipients on the other.

Each case study describes the current status of the technology and discusses how it is expected to develop. An impact scenario then outlines the possible consequences of the anticipated developments for public debate and the democratic process.

7.1 Case study on deepfakes

7.1.1 Current situation

As briefly described in the quick scan, artificial intelligence (AI) can be used to process and manipulate existing audiovisual material. The camera apps on smartphones that manipulate portrait photos with beauty filters are a simple and commonly used example of this. The use of artificial intelligence in audio and visual materials has grown so rapidly in recent years that it is increasingly difficult to distinguish manipulated information from unedited, authentic information.[150] [151] [152]

[150] Brundage, M., et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv:1802.07228 [cs]*. http://arxiv.org/abs/1802.07228
[151] Khodabakhsh, A., Busch, C., & Ramachandra, R. (2018). A Taxonomy of Audiovisual Fake Multimedia Content Creation Technology. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 372–377. https://doi.org/10.1109/MIPR.2018.00082
[152] U. S. Government Accountability Office (2020). *Science & Tech Spotlight: Deepfakes, (GAO-20-379SP).* www.gao.gov/products/gao-20-379sp

Various AI methods can be used to produce deepfakes by editing or manipulating video material. The software used analyses large quantities of visual material of a person and uses it to learn the shape, proportions and movements of a person's face. The producer then decides what postures should be adopted in the edited video. The deepfake algorithm generates the manipulated video image by image. The result is often combined with manipulated audio to create a realistic video, which is difficult to distinguish immediately from an authentic video.

Face swap technology is the best-known example of deepfake technology. Other examples are lip sync technology, with which a mouth's movements can be manipulated, and digital puppetry, which can be used to generate an artificial head or body. With personalised avatar creation technology, an entire virtual body can be superimposed over video images of an existing person.[153]

Various apps have recently appeared that use deepfake technology. Some examples are:

- Face2Face, with which a video can be manipulated by having the face in the video imitate another person's facial expressions in real-time;[154]
- Mug Life, with which a face can be altered in a 3D animation;[155]
- Doublicat, with which a face can be inserted in a GIF;[156] and
- HeadOn, with which faces, movements and facial expressions can be replaced in videos in real-time.[157]

At first glance, these examples appear innocent. For example, a video call with a funny filter can cheer people up during this time of quarantine and lockdown, or a person can insert their own face into amusing animations. But deepfakes also have less innocent applications. The climate action group Extinction Rebellion, for instance, used lip sync technology to produce a deepfake video of a fictitious speech by the Belgian prime minister, in which she spoke of a link between the outbreak of pandemics and the disruption of the natural environment by humans.[158]

[153] Duursma, J. (2019). *Deepfake Technologie – The Infocalypse* www.jarnoduursma.nl/wp-content/uploads/2019/09/Jarno-Duursma-_-Deepfake-Technologie-The-Infocalypse.pdf
[154] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2387–2395. https://doi.org/10.1109/CVPR.2016.262
[155] MugLife (2020). *Bring Your Photos to Live.* www.muglife.com/
[156] Reface (2020). *The Best Face Swap App.* https://reface.app/
[157] Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., & Nießner, M. (2018). HeadOn: Real-time Reenactment of Human Portrait Videos. *ACM Transactions on Graphics, 37*(4), 1–13. https://doi.org/10.1145/3197517.3201350
[158] BELGA (2020). *Extinction Rebellion publiceert deepfake-video met alternatieve speech premier Wilmès* Nieuwsblad.be www.nieuwsblad.be/cnt/dmf20200414_04921988

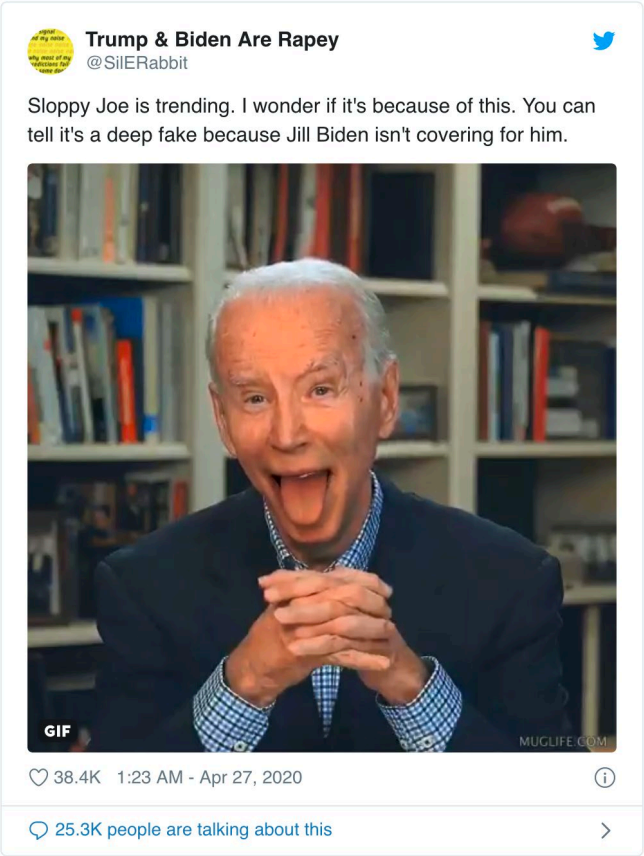

Figure 11 A video of the American presidential candidate Joe Biden that has been manipulated with Mug Life.

Developments in the field of deepfakes are advancing rapidly and many new apps and methods are emerging. The examples given here are just a sample of the many applications that are currently available.

Deepfakes are disseminated mainly on social media platforms.

7.1.2 Expected developments

The production and dissemination of deepfakes is likely to become increasingly widespread. Not only will further technological innovation make them more difficult to distinguish from authentic images, deepfake technology will also become more accessible because technology companies will bring the increasingly advanced technologies onto the market in easy-to-use apps and gadgets. This will enable

many ordinary users to produce their own deepfakes.[159] Anyone with a certain level of computer skills will be able to produce deepfakes.

According to the cyber security company Nisos, deepfakes are not yet sufficiently advanced to be used for criminal purposes. They are therefore not yet offered as a service on the dark web. However, the technology is improving rapidly and Nisos expects that deepfakes will be good enough to be used for criminal purposes within the foreseeable future.[160] Large technology companies such as Apple and Amazon are already busy perfecting the technology behind deepfakes. It is entirely possible that the resulting deepfake technology will be more than good enough to meet the requirements of criminal groups.

Growing popularity of manipulated visual material

Video sharing platforms such as YouTube, SnapChat, Instagram and TikTok are expected to become even more popular, particularly among young people. The platforms also offer augmented-reality functionalities, with which information can be added to live camera images in real-time. SnapChat foresees a future where the standard use of smartphones will be to produce live camera images.[161]

It is becoming increasingly normal to manipulate images on these platforms. SnapChat and Instagram also encourage users to apply filters, to combine images or to add text and stickers to them. The manipulation includes the use of deepfake technologies to swap or age faces, to change gender characteristics and to manipulate voices. These platforms also allow users to transmit live images. With the arrival of even faster mobile internet connectivity, such as 5G, the real-time use of deepfake technologies will also increase.

Growing importance of images in news reporting

The impact of deepfakes will also become greater as visual material becomes more important in the reporting of news. That trend is already apparent, especially on the internet. The younger generations are increasingly turning to digital sources for their news, including reporting on platforms such as Facebook. There is also a growing visual culture, as reflected in the popularity of internet services such as YouTube, TikTok, Instagram and SnapChat. News media will also probably focus more on those platforms.

---

[159] Schulz, J. (2020). *The Deepfake iPhone Apps Are Here*. Lawfare. www.lawfareblog.com/deepfake-iphone-apps-are-here

[160] Volkert, R. (2020). *Deep Fakes: Understanding the illicit economy for synthetic media.* NISOS. https://cdn2.hubspot.net/hubfs/6068438/Resources/NISOS%20-%20Deep%20Fakes%20White%20Paper.pdf

[161] Hern, A. (2020). Snapchat firm unveils platform plan to take on Google and Apple. *The Guardian*. www.theguardian.com/technology/2020/jun/15/snapchat-firm-unveils-platform-plan-to-take-on-google-and-apple

Media literacy leaves something to be desired
The growing importance of visual news coverage in combination with the greater access to deepfake technologies and the increasing popularity of manipulated visual material is a problem because a significant proportion of the Dutch population already has difficulty in determining whether a news report is genuine or fake. For example, de Volkskrant has reported that only 29% of Dutch people say they 'can distinguish genuine news from fake news' and that one in three people say 'they often no longer know what is true and what is false'.[162] In addition, 33% of the population say they never verify the accuracy of the news.[163]

According to the Monitor of Youth and Media, the digital knowledge and skills of many young people leaves something to be desired. It found large differences between students of different types of education in terms of digital literacy. The level is particularly low among students in practical education and preparatory secondary vocational education (VMBO).[164]

7.1.3 Impact scenario

Credibility of visual material is being eroded
To give an impression of the potential impact of the aforementioned developments on public debate and the democratic process, in this section we describe a possible future scenario. The example we use is not the risk of an interview with Prime Minister Mark Rutte on the NOS news bulletin being hacked using deepfake technology, since we assume that it is very likely that if that happened the news bulletin would be interrupted or the NOS would publicly repudiate the report, so that the hack would probably have little effect.

The expectation is that deepfakes could have a far greater and more diffuse effect if they are disseminated out of public view on informal and closed channels such as private chat groups on uncensored social media platforms like Parler or forums like Reddit or 8chan. There is practically no moderation on those media and there is significantly less chance of messages on them being contradicted.

The dissemination of deepfakes on informal and closed channels could also be a response to more frequent detection of deepfakes by large platforms with the help

[162] Kranenberg, A. (2017). *Nederlanders bezorgd over 'nepnieuws' - een op drie weet vaak niet meer wat waar is en wat onwaar.* Volkskrant www.volkskrant.nl/nieuws-achtergrond/nederlanders-bezorgd-over-nepnieuws-een-op-drie-weet-vaak-niet-meer-wat-waar-is-en-wat-onwaar~b6914596/
[163] Consultancy.nl (2018). *Nederlanders herkennen nepnieuws en maken zich niet zo druk om fake news* www.consultancy.nl/nieuws/17892/nederlanders-herkennen-nepnieuws-en-maken-zich-niet-zo-druk-om-fake-news
[164] Pijpers, R. (2019). *Werken aan digitale geletterdheid: van visie naar praktijk.* Kennisnet www.kennisnet.nl/publicaties/werken-aan-digitale-geletterdheid-van-visie-naar-praktijk/

of artificial intelligence, for example as a result of the Deepfake Detection Challenge launched by Facebook in 2019.[165]

In this scenario, deepfakes are produced by various groups, both professional trolls and less professional groups, conspiracy theorists and hobbyists. Because the technology required is increasingly easy to use, more and more people will be able to produce their own deepfakes. That also means that deepfakes could be disseminated across a variety of platforms and channels. Professional organisations could also make extensive use of deepfakes and look for the best way of disseminating them to achieve the maximum effect.

The growing ease with which deepfakes can be produced and disseminated via numerous informal and closed channels that are not moderated could lead to a proliferation of deepfakes on the internet. And because it is simultaneously becoming more difficult for recipients to distinguish manipulated reports from authentic ones, this could eventually, and stealthily, lead to a diminution in the significance of the distinction between authentic and manipulated material. A side effect of such a development could be that visual material from established media is also dismissed as manipulated or fake. And that could in turn lead to a loss of trust in the media. The credibility of visual material would, as it were, be eroded by constant exposure to manipulated images.

7.2 Case study on psychographing

7.2.1 Current situation

A psychograph was originally a visual representation of the personal characteristics of a person or group[166] such as values, desires, goals, interests and lifestyle. Marketers use psychographing to tailor advertisements to a particular target group. There is nothing to prevent producers and disseminators of disinformation from using this technology.

In this case study, the term psychographing refers not just to visual representations of personal characteristics, but also to a set of digital technologies that can be used to tailor messages to the personal characteristics of a target group. As we mentioned in Part I, psychographing can be regarded as an advanced form of micro-targeting.

---

[165] Facebook (2020). *Deepfake Detection Challenge* https://deepfakedetectionchallenge.ai/

[166] Wells, William D. (1975). 'Psychographics: A critical review'. *Journal of Marketing Research*. 12: 196–213. doi:10.2307/3150443. JSTOR 3150443.

Pyschographing has a long history. For decades it was based mainly on traditional target-group research using surveys, interviews and focus groups. The findings could be used to divide target groups into sub-groups, for example on the basis of the five-factor model used in psychology. That model distinguishes personal characteristics such as emotional stability, extraversion, intellectual autonomy and orderliness.[167] The idea behind the model is that if a target group is known to have a low score on extraversion, let's say, the best way of reaching it is with a calm message.

With new digital technologies, the process of researching target groups and tailoring messages to them can be automated.[168] For example, IBM claims that its Watson algorithm is capable of identifying personal characteristics from texts.[169] A company that uses IBM's services is Indivizo, which employs them to distill personal characteristics from a video of a job interview.[170] During the American presidential election in 2016, Cambridge Analytica claimed that it could conduct an effective political campaign with the help of the Facebook data of 87 million users. The company came under fire when it emerged that the users had not consented to the use of their data.[171]

Automation enables psychographing to be used on a large scale. The underlying idea is that people's (political) opinions can be influenced by presenting them with information tailored to their psychological traits, including their psychic vulnerabilities.

The claims made for the effectiveness of psychographing have been criticised in traditional media in the last few decades. The success of the new, automated methods is also disputed. But research has shown that the techniques can prompt people to click on links on websites or make online purchases more often (see figure 12). However, it is not known whether the technique also has an effect on aspects such as political preferences or voting behaviour.[172] [173]

---

[167] McCrae, R. R.; Costa, P. C.; Jr (1987). 'Validation of the five-factor model across instruments and observers'. *Journal of Personality and Social Psychology*. 52 (1): 81–90. doi:10.1037/0022-3514.52.1.81. PMID 3820081.
[168] Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112
[169] IBM (no date). *Personality Insights* www.ibm.com/watson/services/personality-insights/
[170] Indivizo (z.d.). *AI-based Personality Profiles* www.indivizo.com/personality-profiles
[171] Lapowsky, I. (2018). *Facebook Exposed 87 Million Users to Cambridge Analytica.* Wired www.wired.com/story/facebook-exposed-87-million-users-to-cambridge-analytica/
[172] Rokka, J. & Airoldi, M. (2018). *Cambridge Analytica's 'secret' psychographic tool is a ghost from the past.* The Conversation. https://theconversation.com/cambridge-analyticas-secret-psychographic-tool-is-a-ghost-from-the-past-94143
[173] Resnick, B. (2018). *Cambridge Analytica's "psychographic microtargeting": what's bullshit and what's legit.* Vox. www.vox.com/science-and-health/2018/3/23/17152564/cambridge-analytica-psychographic-microtargeting-what

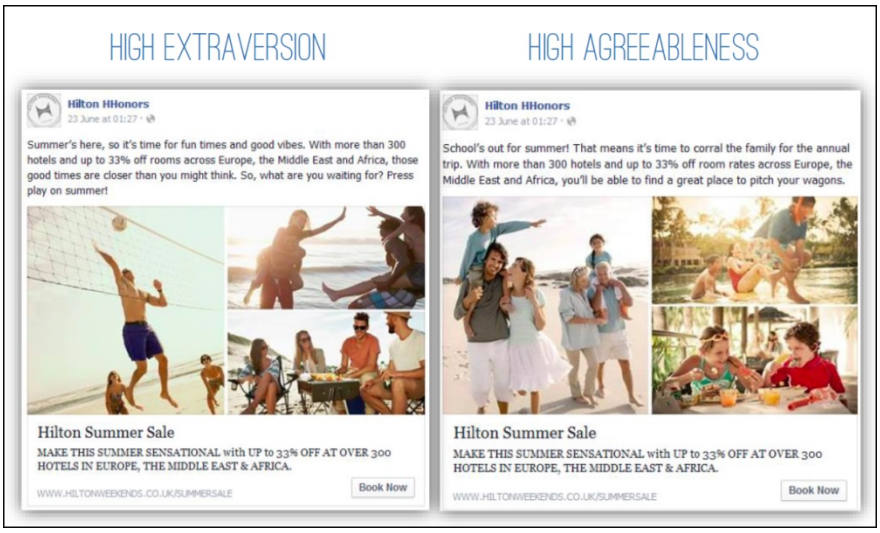

Figure 12 Advertisements on Facebook for Hilton holidays in which images and text are adapted to the estimated personal characteristics of the target group. Research by Cambridge University has shown that these advertisements were more effective than advertisements focusing on a general characteristic such as a love of travel.[174]

7.2.2 Expected developments

For the purposes of this case study we assume that the use of psychographing is effective, at least to a certain extent. After all, the technologies already seem to be capable of steering people's attention. It is therefore entirely possible that with further development the technology could also influence public attitudes to social or political issues. In that case, it is also likely that more actors will start using it.

Furthermore, the volume of data being collected about internet users online, from which a range of personal characteristics can be derived, is growing all the time. For example, the rise of the Internet of Things, whereby more and more devices, such as smart TVs and self-driving cars, are connected to the internet, means that even more data can be collected about people's offline behaviour. With biometric sensors, which are increasingly used in AR/VR equipment for example, eye and pupil movements can be monitored. The information generated can provide more insight into a person's character traits and preferences. Psychographing techniques could be further refined on that basis.

[174] LaMontagne, L. (2015). *Personality-Matched Ads: How Hilton Worldwide effectively personalized its marketing messages.* MarketingExperiments https://marketingexperiments.com/digital-advertising/hilton-worldwide-personality-matched-ads

7.2.3 Impact scenario

In this section we sketch a scenario based on an actor who possesses advanced technical resources and the motivation to conduct a long-term (disinformation) campaign. This represents what is known as an Advanced Persistent Threat (APT). Because of their technological capabilities and effectiveness, it is often suspected that groups behind an APT are allied to or supported by a state actor.

This case study is based on the assumption that an APT group has set out to covertly influence public debate and the democratic process with the help of psychographing technologies. By involving itself in sensitive social issues, the group's intention is to provoke social divisions and undermine people's confidence in established institutions.

To achieve this goal, the APT group sets out to influence individuals with strong feelings about a particular issue or who distrust the establishment. To identify those individuals, the group uses automated forms of psychographing, with the help of artificial intelligence. The algorithms used during the long-running campaign are repeatedly refreshed with the data of selected individuals and so become steadily better at finding persons whose character is similar.

The individuals are found by searching in large databases, which can be bought from brokers in the advertising sector. Data can also be gathered from public sources such as social media, news media or public government sources. But they can also be stolen by means of hacks or collected from data leaks. The General Data Protection Regulation (GDPR) provides no protection in this scenario, because an APT group operates covertly and will not feel threatened by the GDPR.

The messages disseminated by the group are designed in such a way that they can be easily shared on social media. In other words, the preferred method is to use short tweets, posts, films or catchy images. When it sends the messages, the APT group also encourages people to pass on the disinformation.

The disinformation is disseminated mainly through social media platforms. It is occasionally also picked up by the established media, since their reporting is increasingly driven in part by what is trending on Twitter or YouTube.

To create maximum unease and mistrust, the APT group disseminates messages mainly via non-public channels, such as closed groups on Facebook or Telegram, since there is little chance of the messages being contradicted on those channels. That further increases the impact of the disinformation campaign. Furthermore, in this way it is possible to provide contradictory information to opposing sides in the public debate without their realising it.

Providing the separate target groups with conflicting messages geared to their personal characteristics increases divisions in society. That leads to polarisation in the public debate and a growing unwillingness to engage in a dialogue with the other side, which undermines the democratic process.

In this case, the APT group could in fact also respond cleverly to the measures taken by platform companies to combat disinformation. For example, the group could use revelations of disinformation by fact-checkers circulating in one camp to show the other camp how naive their opponent is, and hence reinforce the discord.

The APT group could also react to the measures taken by platform companies to moderate the content, such as deciding that certain messages constitute disinformation and having them recommended less by the algorithms or removed. The APT group could then brand those measures as a form of censorship and thus further reinforce the sense of mistrust in established parties.

Part III Outlook

8 New measures

This chapter describes further measures that could be taken to prevent harm to public debate and the democratic process as a result of technological developments in the area of disinformation, and the actors that should take those measures. In that context, we build to a large extent on the case studies in chapter 7 and the findings from the expert meeting on 2 June 2020. Additional desk research was also carried out for this chapter.

We concentrate on measures that do not infringe freedom of speech and press freedom. For example, the government cannot remove misinformation purely and simply because it is misleading. There have to be additional legal grounds. Otherwise removing disinformation would be contrary to freedom of speech.

As we said in the introductory chapter, this study is not concerned with as yet unknown technological developments, since we have no way of knowing what those developments will be. In the previous chapters we have shown how technological innovations that are already emerging might evolve and what impact they could have on the production and dissemination of disinformation. The pace of technological developments in the IT domain makes it difficult to predict where they will lead. Describing them is therefore to a certain extent speculative.

For example, there is still considerable uncertainty about the effectiveness of disinformation campaigns using advanced forms of micro-targeting. Although these applications are already widely used in the advertising world, their effectiveness has not been adequately demonstrated. At the same time, it is impossible to rule out the possibility that the use of micro-targeting for disinformation purposes could prove effective enough to be regarded as a threat to public debate and the democratic process.

In the following sections we provide an overview of the most important new measures that could be taken to counter potential threats from technological developments in the field of disinformation.

8.1 Measures against widespread deepfakes

The case study on deepfakes in Part II showed that widespread production and dissemination of deepfakes could stealthily undermine the distinction between authentic and manipulated visual material.

If it becomes increasingly difficult for citizens to distinguish fake from authentic, they could become indifferent to the distinction: perhaps everything you see on internet is a little bit true and nothing is entirely true. That creates the risk that the reporting by the established media and government agencies will also no longer be regarded as reliable since it can also be manipulated.

Measures designed to increase media literacy could increase awareness that not everything that is seen and heard on the internet is true, but most people are still unlikely to be able to distinguish authentic from manipulated visual material. Because manipulation with the help of artificial intelligence is becoming increasingly refined, people will no longer be able to believe their own eyes and ears. That runs counter to the government's point of departure that citizens can judge the value of information and disinformation for themselves.

8.1.1 Detection of manipulated visual material

Guidelines should therefore be developed to help citizens to distinguish between authentic and manipulated visual material. One such instrument would be for social media companies to use artificial intelligence to detect manipulated visual material and to mark material in their messages that has or might have been manipulated.

The algorithms that are used to produce deepfakes often leave traces that can be picked up by detection techniques.[175] For example, persons in manipulated videos often do not blink or do so in an unnatural way. That could be detected with the help of artificial intelligence.[176] Various initiatives are underway to improve detection systems. For example, there are datasets available that developers of detection algorithms can use to test and train their systems, such as Faceforensics++.[177] Large technology companies have also launched various programmes to improve detection methods. Google participates in Reality Defender 2020,[178] and Amazon, Facebook and Microsoft are working together on the Deepfake Detection Challenge

[175] Hameleers, M., Powell, T. E., Meer, T. G. L. A. V. D., & Bos, L. (2020). A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication*, *37*(2), 281–301. https://doi.org/10.1080/10584609.2019.1674979
[176] Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. https://doi.org/10.1109/WIFS.2018.8630787
[177] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to Detect Manipulated Facial Images*. www.niessnerlab.org/projects/roessler2019faceforensicspp.html
[178] Reality Defender 2020 (no date). https://rd2020.org/

(DFDC).[179] Various researchers are also developing detection strategies and tools, such as the detection system Poster.[180]

As a rule, manipulated visual material is only detected after it has been disseminated. Detecting it therefore has little effect. It would be more effective to scan and filter messages for deepfakes in advance. Deepfakes are also expected to be used more frequently in live streaming, which makes real-time detection essential. It might soon be possible to install plug-ins in web browsers that can detect and block deepfakes in real-time.[181]

In response to the enhanced possibilities of detecting deepfakes, producers and disseminators might switch to the even more advanced forms of image manipulation that are appearing on the market, or develop them themselves. Platform companies should therefore also continue to invest in detection technologies in order to keep pace with the producers and disseminators of steadily more advanced deepfakes.

Hotline for malicious image manipulation
It is difficult to detect deepfakes on platforms like SnapChat, Instagram and TikTok because of the trend, described in the case study, that the manipulation of visual material is becoming increasingly normal on those platforms. The challenge then is to distinguish benevolent from malevolent visual manipulation. The question is whether technological solutions can be developed that are capable of properly interpreting the context of images and spoken text, so that visual manipulation that might be malicious can then be assessed by human moderators.

Platforms could also establish a hotline where users can report suspicions of malicious visual manipulation. Moderators could then view the images and mark or remove materials they find to be malicious. That would require sufficient investment by these companies in capacity to moderate their communities.

Platform companies are primarily responsible
Primary responsibility for detecting (malicious) deepfakes lies with the platform companies. Under existing legislation the government does not have the power to compel the companies to detect deepfakes. Article 15 of the European e-Commerce Directive prohibits the imposition of such an obligation on internet

[179] Facebook (2020). *Deepfake Detection Challenge* https://deepfakedetectionchallenge.ai/
[180] Sohrawardi, S.J. et al. (2019). Poster: Towards Robust Open-World Detection of Deepfakes. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2613–2615. https://doi.org/10.1145/3319535.3363269
[181] Duursma, J. (2019). *Deepfake Technologie – The Infocalypse* www.jarnoduursma.nl/wp-content/uploads/2019/09/Jarno-Duursma-_-Deepfake-Technologie-The-Infocalypse.pdf

companies.[182] But given the public interest in reliable reporting, the government could insist that platform companies pursue an active detection policy. To be effective, any such measure should preferably be taken by the European Union.

8.1.2 Authentication of visual material

A second instrument that could help citizens to distinguish authentic from manipulated visual material is authentication of the material. A digital signature, for example, makes it easier for citizens to determine whether material is from a reliable source.[183] It gives members of the public at least some assurance in the search for reliable reporting.

An important requirement for this is the existence of a reliable system for registering hallmarks. A potential pitfall of this method is that it is not watertight and that by digitally signing unreliable messages people will, consciously or otherwise, collaborate in the dissemination of disinformation.

Various parties are currently engaged in initiatives in this area. The BBC, the New York Times, Google, Facebook, Microsoft and other parties have joined forces in the Trusted News Initiative, a system for authenticating reports from sources they consider to be reliable.[184] The government could make authentication standard practice with all of its own reporting, as well as insisting that public news sources that receive funding or are co-financed by the government always authenticate their reports.[185] The Dutch public broadcasting organisation (*Nederlandse Publiek Omroep*, NPO) is already taking initiatives in this domain.[186]

There are also methods for preventing the spread of manipulated visual material. For example, platform companies that provide filters to manipulate images could automatically embed hallmarks in images that are edited with those filters. If others then disseminate the images with the pretence that the material is authentic, the claim can be verified by the recipient.

---

[182] EUR-lex (2000). *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')* https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031
[183] Jacobs, B. (2019). *Teken tegen nepnieuws.* iBestuur https://ibestuur.nl/weblog/teken-tegen-nepnieuws
[184] Cooper, D. (2020). *News outlets will digitally watermark content to limit misinformation.* Engadget www.engadget.com/bbc-fake-news-111000601.html
[185] Van Boheemen, P., Munnichs, G., Kool, L., Diercks, G., Hamer, J., & Vos. A. (2020). *Cyberweerbaar met nieuwe technologie.* Rathenau Instituut. www.rathenau.nl/nl/digitale-samenleving/cyberweerbaar-met-nieuwe-technologie
[186] Takken, W. (2020). *Martijn van Dam: 'de publieke omroep moet ook online verbindend zijn'*. NRC. www.nrc.nl/nieuws/2020/06/08/npo-moet-ook-online-verbindend-zijn-a4002042

Another option is controlled capture software, which records the time and location at which an image was produced so that the data cannot be altered later.[187] With Truepic, for example, the time, the location and the identity of the smartphone with which a photograph was taken can be registered.[188]

8.2 Measures against influencing through micro-targeting

The case study on psychographing showed that with micro-targeting various target groups in society can be reached with different – and possibly conflicting – social and political messages that are concealed from others. In this way, the political mood in the country and the political opinions of citizens can be influenced.

8.2.1 Address more than just political advertisements

The current policy debate about micro-targeting in the Netherlands focuses mainly on its use in political advertising campaigns. For example, the State Commission on the Parliamentary System (Remkes Commission) called for a statutory requirement of transparency to compel political parties to be open about their use of digital instruments. According to the commission, citizens must be able to recognise political advertisements as such, and also see why they are receiving a particular message and who is paying for the advertisement.[189]

In line with this, the Minister of the Interior and Kingdom Relations said in a letter to the House of Representatives: 'To prevent improper influencing of fair and free elections, election campaigns must be transparent. I am therefore inserting rules into the Political Parties Act (*Wet op de politieke partijen*) designed to safeguard and improve the accountability of those campaigns, to prevent deception and to provide clarity about who has paid for an advertisement. The purpose of these rules is to protect democracy against (…) risks that digital information and communication technologies could pose for elections. Their regulation is intended to make the campaigns transparent for voters'.[190]

---

[187] Duursma, J. (2019). *Deepfake Technologie – The Infocalypse* www.jarnoduursma.nl/wp-content/uploads/2019/09/Jarno-Duursma-_-Deepfake-Technologie-The-Infocalypse.pdf
[188] Truepic (no date). *Photo and video verification you can trust* https://truepic.com/
[189] Staatscommissie parlementair stelsel (2018). *Lage drempels, hoge dijken: Democratie en rechtsstaat in balans.* www.staatscommissieparlementairstelsel.nl/documenten/rapporten/samenvattingen/12/13/eindrapport
[190] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2020). *Kamerbrief over voortgang voorbereiding Wet op de politieke partijen*. www.rijksoverheid.nl/documenten/kamerstukken/2020/06/11/kamerbrief-inzake-voortgang-voorbereiding-wet-op-de-politieke-partijen

The Remkes Commission in fact noted that the way micro-targeting is used in the US is impossible in the Netherlands because parties do not have access to registers of voters.[191] The General Data Protection Regulation (GDPR) also prohibits the use of data about political opinions without the consent of the individuals concerned.

Although the measures proposed by the minister mark an important step in preventing possible deception of the voter with digital advertising campaigns, we must express a number of reservations here.

For example, we anticipate that the prohibition of the use of data about political preferences in the GDPR can be avoided. For example, the GDPR will probably make little impression on an Advanced Persistent Threat group (see section 7.2) that wishes to use data about political preferences to influence political campaigns. Of course, the group could be prosecuted on the grounds of the GDPR if it does actually use the data for that purpose.

The GDPR could also be avoided by discovering people's political preferences by proxy on the basis of other data – such as their postcode, the make of car they drive or the newspaper they subscribe to. They could then be provided with specific, slanted information based on their estimated preferences. In this case, the data used would not have to relate to the specific individual's political preferences, which would fall under the GDPR.

Furthermore, we feel that an approach that focuses mainly on political advertising is too narrow. The case study on psychographing demonstrated that disinformation campaigns using micro-targeting can also be intended to stir up political divisions and can foster radicalisation and polarisation. They could then have a major impact on the political mood in the country.

Micro-targeting can also be used to spread conspiracy theories, thus damaging confidence in the established media, the judiciary and the political institutions among some groups in society. That could surreptitiously undermine the legitimacy of the democratic rule of law.[192]

The Minister of the Interior and Kingdom Relations has said that with respect to the dissemination of disinformation by means of micro-targeting, she does not want to focus solely on its use in political campaigns, to which end she intends to amend

[191] Staatscommissie parlementair stelsel (2018). *Lage drempels, hoge dijken: Democratie en rechtsstaat in balans.* www.staatscommissieparlementairstelsel.nl/documenten/rapporten/samenvattingen/12/13/eindrapport
[192] Cohen, J.E. (2020 Forthcoming). Tailoring Election Regulation: The Platform Is the Frame. *4 Geo. Tech. L. Rev.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3573127

the Act on Political Parties. The minister is also in favour of 'greater transparency about the origin and the methods of dissemination of disinformation on internet services'. She is reviewing whether 'statutory rules could compel transparency'. As the minister herself has said, it is not yet clear how that might be done. [193]

8.2.2 Regulation by platform companies

The previous section has shown that to tackle disinformation campaigns using micro-targeting, the role of the platform companies must also be considered. After all, it is they that enable other, malicious parties to use micro-targeting for the purpose of disseminating disinformation. The current regulation of platform companies is insufficient to prevent these negative effects.

The EU code of practice is not enough

The most important existing instrument for regulating platform companies is the EU Code of Practice on Disinformation. By signing the code of practice, a number of major platform companies indicated their willingness to endeavour to prevent the dissemination of disinformation. But the results have been disappointing.

The ERGA, the European umbrella body of media watchdogs, observes that the code of practice has a number of shortcomings.[194] First, platform companies report on the implementation of measures themselves, making independent supervision impossible. Furthermore, the terms in which reports are formulated are too general, with information aggregated at EU level, and there is uncertainty about the definitions of key concepts such as 'political advertising'. The code has also not been signed by some popular platforms, including WhatsApp and Messenger.

The ERGA therefore recommends tightening up the code. For example, it calls for the efforts of the platform companies to be monitored at the national level by independent regulators. The ERGA proposes that the regulators should publish internationally comparable reports based on uniform definitions and indicators. Every platform company – at least above a certain size – operating within the EU should be required to accept this form of co-regulation, which would involve self-regulation by the platforms together with powers for the regulators to impose coercive measures in the event of non-compliance.

---

[193] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2020). Kamerbrief ontwikkelingen beleidsinzet bescherming democratie tegen desinformatie. www.rijksoverheid.nl/documenten/kamerstukken/2020/05/13/kamerbrief-ontwikkelingen-beleidsinzet-bescherming-democratie-tegen-desinformatie

[194] European Regulators Group for Audiovisual Media Services (2020). *ERGA Report on disinformation: Assessment of the implementation of the Code of Practice* https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf

The expert meeting organised for this study produced a similar picture. There was broad agreement at the meeting that platform companies are doing too little to prevent the spread of disinformation. According to the experts, they are also doing too little to tackle more extreme and criminal statements such as hate speech and incitement of violence. In the experts' view, self-regulation has not been effective enough and stricter regulation is therefore required.

Restricting possibilities for micro-targeting

In the first place, platform companies could – like other developers of commercial advertising systems – build monitoring systems into the services they develop for designing, executing and analysing online advertising campaigns. They would then be able to monitor who is using their services and for what purposes, screen their customers in advance (due diligence) and take preventive measures if they suspect that customers are abusing their services. This would enable them to detect the use of the services for campaigns that foster polarisation or radicalisation, or other forms of disinformation. That would require the platform companies to monitor such activities. They could then take action against customers who use their services for harmful purposes.

Platform companies could also adopt technical measures to limit the possibilities for abusing advertising technology. They could take measures to prevent the use of micro-targeting to influence selected target groups, for example by curtailing the possibilities for selecting target groups on the basis of personal characteristics. This could be achieved by restricting the automatic extraction of data by means of APIs, which would make it technically impossible to select target groups on the basis of political preferences or from data that are known to indicate political preferences.

According to Crain and Nadler, the advertising systems should also be more transparent for internet users, who often don't know what their advertising profile is or why it applies to them. Users are also often unable to see who is sending an advertisement. Users should also be given more control over the data that are used by advertisers in order to prevent their data from being used for purposes that harm their own interests. Crain and Nadler also propose that the possibilities for micro-targeting should be restricted by establishing a minimum for the size of a target group.[195]

The measures to restrict the possibilities for micro-targeting mentioned here could be at the expense of the revenue model of the platform companies because they could make those companies less attractive for certain categories of advertisers. It

[195] Crain & Nadler (2019). Political Manipulation and Internet Advertising Infrastructure. *Journal of Information Policy*, *9*, 370. https://doi.org/10.5325/jinfopoli.9.2019.0370

is therefore by no means certain that the platform companies would adopt those measures of their own volition. In that case, the government – and preferably the EU – could insist on them, or impose additional requirements through legislation.

Some parties even advocate a ban on personalised advertisements and micro-targeting. Important reasons given for that are the serious infringement of the privacy of internet users and the manipulation of their online behaviour on the basis of the data collected about them. The proposed ban would address all the various problematic aspects and is seen as more effective than regulating each problem separately.[196]

The European Parliament has also adopted a motion calling for a ban on personalised advertisements. It says that the ban should be incorporated in the Digital Services Act. Paul Tang, the author of the amendment, argues that personalised advertisements are based on an unwarranted infringement of privacy, that they are annoying and that it is asking too much of many users to have to repeatedly reject tracking cookies.[197]

8.3 Transparency about recommendation algorithms

As set out in the quick scan, the recommendation algorithms of platform companies are usually aimed at retaining the user's attention for as long as possible, which is most successful with sensational content that matches the previously identified preferences of the user. Accordingly, these algorithms reinforce social and political preferences and social divisions.

The recommendation algorithms work in a similar way to the technologies for micro-targeting. With the help of the algorithms, the many thousands of data that platform companies collect about their users are analysed. Users are shown content that is geared to the preferences and personal characteristics that emerged from that analysis. An important distinction is that recommendation algorithms do not produce content themselves, but use messages from others. Another difference is that the majority of the platform companies are not consciously trying to undermine public debate or the democratic process with their algorithms. However, all things considered the radicalising and polarising effect of the recommendation algorithms can cause such harm – and the platform companies do not currently seem inclined

[196] Edelman, G. (2020). *Why Don't We Just Ban Targeted Advertising?* www.wired.com/story/why-dont-we-just-ban-targeted-advertising/
[197] Kist, R. (2020). *Europarlementariër Paul Tang: 'Persoonlijke advertenties zijn een smet op het internet'.* NRC. www.nrc.nl/nieuws/2020/06/19/overwinning-voor-paul-tang-in-strijd-tegen-gepersonaliseerde-advertenties-techreuzen-a4003409 en https://paultang.nl/en/forbid-personalised-ads/

to prevent those harmful effects. Instead, they claim to perform a neutral role as an intermediary between senders and recipients of messages.

According to Julie Cohen, an American professor of law and technology, it follows from the mechanics of the recommendation algorithms that platform companies are anything but a neutral conduit for content posted by others. The algorithms display content that reflects the preferences, wishes and concerns that follow from the user profiles, reinforce those preferences and hold the attention of users with sensational reporting. Accordingly, says Cohen, manipulation of users is 'endemic'.[198]

Building a reflection period into platform services

The points made above raise the question of what possibilities platform companies have to combat the harmful effects of their recommendation algorithms. One way of preventing the spread of sensational or more extreme messages is by building in a moment for reflection by the users of platform services.

Because these messages are disseminated in part by users who 'like' the information or share it with others, a place to start is with their online behaviour. Users often share messages impulsively, without really giving it any thought. By inserting a mechanism that forces them to wait a few seconds before they can forward a message, they might act less impulsively.[199] In that context, Twitter is experimenting with a warning if users forward an unopened link whose reliability they themselves have not been able to assess.[200]

The government – preferably the EU – could demand that platform companies build in such a reflection period, and could in fact also require it of managers of chat apps.

Transparency about the use of algorithms

A more far-reaching measure to combat the polarising and radicalising effect of recommendation algorithms would be to critically analyse, and if necessary modify, the use of those algorithms. But platform companies are not transparent about the algorithms they use. They regard them as sensitive trade secrets. The Dutch government endorses that. For example, the Minister of the Interior and Kingdom

[198] Cohen, J.E. (2020 Forthcoming). Tailoring Election Regulation: The Platform Is the Frame. *4 Geo. Tech. L. Rev.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3573127
[199] Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-009
[200] Twitter Support (2020). https://twitter.com/twittersupport/status/1270783537667551233

Relations has said that the recommendation algorithms are the intellectual property of the platform companies and how they design them is a matter for themselves.[201]

However, various parties advocate greater transparency from the platform companies about their algorithms. The European High Level Expert Group on Fake News and Online Disinformation, for example, has called for more transparency about the use of recommendation algorithms by platforms.[202] And the British House of Commons' Digital, Culture, Media and Sport Committee appealed in a research report on disinformation for the creation of an independent regulator of platform companies who would have access to the algorithms they use.[203] An NGO, the Electronic Frontier Foundation, goes a step further and argues that internet users should be able to personally change the algorithms that determine what content they see and should be able to say what sources they trust.[204]

Research into the mechanics of algorithms

A first step towards more transparency would be for platform companies to give scientific researchers access to their recommendation algorithms in order to investigate how the algorithms work and what impact they have on the online behaviour of internet users. This would provide greater insight into how they work and their potential negative effects.[205] [206] [207] If necessary, governments should compel access to that information for scientific researchers.

8.4 Measures aimed at closed and encrypted channels

As the case studies in Part II showed, producers and disseminators of deepfakes and groups that use micro-targeting to launch disinformation campaigns might use closed or encrypted groups and channels on social media platforms and chat apps which are not moderated. In reaction to the increasing possibilities of detecting

[201] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2019). *Antwoord op vragen van het lid Baudet over het rapport 'Politiek en Sociale Media Manipulatie'*. www.tweedekamer.nl/kamerstukken/kamervragen/detail?id=2019Z20342&did=2019D46745
[202] European Commission (2018). *Final report of the High Level Expert Group on Fake News and Online Disinformation* https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation
[203] House of Commons (2019). *Disinformation and 'fake news': Final Report* https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf
[204] York, J., Greene, D. & Gebhart, G. (2019). *Censorship Can't Be The Only Answer to Disinformation Online*. Electronic Frontier Foundation www.eff.org/nl/deeplinks/2019/05/censorship-cant-be-only-answer-disinformation-online
[205] Ausloos, J. (2020). *Technologiereuzen moeten zeggen hoe ze ons gedrag bepalen en zo dwingen we dat af*. VRT www.vrt.be/vrtnws/nl/2020/06/25/de-macht-van-technologiereuzen-en-hoe-ze-aan-banden-te-leggen/
[206] European Commission (2018). *Final report of the High Level Expert Group on Fake News and Online Disinformation* https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation
[207] Bruns, A. (2019). After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

deepfakes, disseminators of deepfakes could decide to move to those groups or channels; organisers of (covert) micro-targeting campaigns will generally use them.

Although these closed or encrypted groups and channels have a smaller reach than the public channels of social media platforms, they can still have millions of users. Messages can also be transmitted on multiple channels simultaneously.

User limits
The risk is that producers and disseminators of disinformation will have free rein. One way of preventing the related threats to public debate and the democratic process is to establish a maximum number of users for each group or channel. Some platform companies already adopt such limits. A major drawback of such a measure is that it would also restrict the use of such channels, which are welcomed from the perspective of democracy. Take the important public function of closed and encrypted channels in non-democratic countries, for example. Telegram channels, which do not have limits to the number of users, are widely used by Iranians to discuss current affairs. In Brazil, chat apps are the most important source of news for many citizens.

Warning system
Another way of combatting the impact of disinformation on closed and encrypted channels is to create an independent national warning system to detect disinformation campaigns concerning sensitive social issues. This option was suggested by various participants during the expert meeting.

There is a strong chance that sooner or later messages on closed and encrypted channels will enter the public domain. When that happens, the messages can be detected and responded to. For example, they could be publicly refuted and internet users could be referred to sites with reliable information and to fact-checkers. An example of the latter is the live blog that the Dutch public news channel NOS launched during the corona crisis, which contains informative videos about current developments relating to the virus and the efforts to combat it.

The warning system could in any case ensure that internet users know what is happening and can be kept informed of current disinformation campaigns on sensitive public issues. This is akin to the warnings that employers and banks issue to their employees and customers about common ransomware and phishing attacks.

A warning system of this nature would have to include a hotline where internet users can report suspicions of disinformation and a website with links to sites providing information that is deemed reliable. In light of the public interest, the

government could facilitate these facilities – naturally without undermining the independence of the warning system.

Platform companies could also play a role in referring users to sites that provide information that is regarded as reliable. For example, they could increase the visibility of information from verified sources. During the corona crisis, for instance, the Dutch-language websites of Facebook, Twitter, YouTube and Google have been posting information from the RIVM high in their reporting in order to draw maximum attention to it among internet users. Although this appears to be only a temporary measure, platform companies could make referrals to reliable information sites a more permanent feature.

The recent appeal by the European Commission for the creation of national and international consortia of scientists, fact-checkers, journalists and other relevant stakeholders appears to be in line with this suggestion. One function of these national hubs should be the early identification of disinformation campaigns.[208]

Detection of malevolent chatbots and monitoring by providers
As already mentioned, producers and disseminators can also use (semi-) automated chatbots to spread disinformation.

Suppliers of chat apps could use detection techniques – even without breaking the encryption of the content of the chat messages – to determine whether an account sends more chat messages than is humanly possible within a particular period. If it does, the account could then be removed.

8.5 Fact-checking remains very important

Fact-checking and contradicting misleading information remains crucial for mitigating the impact of disinformation on public debate and the democratic process. After all, the reason it is attractive for producers and disseminators of disinformation to operate covertly on closed and encrypted channels is that it is then more difficult for others to unmask deepfakes or to refute inaccurate information.

As already mentioned, the government does not see a primary role for itself in actively contradicting disinformation. It regards it as primarily the responsibility of journalists and scientists to check the truth of reports and to contradict fake news.

[208] European Commission (2020). *2020 CEF Telecom Call - European Digital Media Observatory (CEF-TC-2020-2)* https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom/apply-funding/2020-edmo

Given the public interest in fact-checking, however, the government could help to facilitate it, for example by providing additional financial support for the Dutch Journalism Fund (*Stimuleringsfonds voor de Journalistiek*) and the Dutch Fund for In-depth Journalism (*Fonds Bijzondere Journalistieke Projecten*) to enable them to hire fact-checkers. In that case, the independence of the fact-checkers would have to be guaranteed.

This indirect facilitation by the government of fact-checking of disinformation would be similar to the way in which the European Commission has supported efforts to combat disinformation by creating a platform to support the work of independent fact-checkers.[209]

Platform companies could also play a role in fact-checking, specifically by labelling messages that possibly contain disinformation and by referring users to fact-checking sites which explain why particular reports are (or might be) inaccurate, so that this contradictory information reaches a wider audience. Platform companies could also provide financial support for fact-checking sites.

Technologies are in fact already being developed to (partially) automate the fact-checking process. For example, Dutch and Flemish researchers have created FactRank, a tool that automatically identifies claims made in parliamentary debates or in tweets by politicians that are check-worthy. The use of such tools could allow fact-checkers to work more quickly. [210]

8.6 Investing in media literacy remains very important

Finally, it emerged from both the interviews and the expert meeting that tighter regulation of platform companies and technological measures to tackle disinformation will have limited effect without continued investment in media literacy. That is in line with government policy, but its importance cannot be emphasised strongly enough.

Invest in media literacy
It goes without saying that the better able people are to judge the value of information (or disinformation), the weaker the social and political impact of misleading reporting will be.

[209] European Commission (2018). *Action Plan on Disinformation.* https://ec.europa.eu/commission/publications/action-plan-disinformation-commission-contribution-european-council-13-14-december-2018_en
[210] Universiteit Leiden (2020). www.universiteitleiden.nl/nieuws/2020/05/lancering-factrank

Although the Netherlands scores better than many other countries in terms of media literacy, the findings from research are not entirely reassuring. A significant proportion of the Dutch population still have difficulty in properly assessing the reliability of information.[211]

Research by Kantar clearly demonstrated that a number of groups in society require additional support in that respect. The elderly and the less-well educated appear to be particularly vulnerable. This relates mainly to their understanding of how the media present – and in the process often slant – reality, and their ability to find and process information and to reflect on their own use of media.[212]
This means that media literacy needs to receive greater attention, not only in education – which focuses mainly on young people – but also in other domains. The Dutch Media Literacy Network (*Netwerk Mediawijsheid*) is in fact already pursuing that objective.

It is also important to bear in mind that with the development of more advanced forms of disinformation (such as sophisticated applications of deepfake technology or more advanced forms of micro-targeting), it will only become more difficult for many people to recognise disinformation. The technological developments are progressing so rapidly that part of the population in unable to keep up with them.

Not solely rational
It also has to be remembered that an overly rational approach to dealing with disinformation could be self-defeating.

An important finding of the expert meeting was that an overemphasis on the truth of information – for example in fact-checking – will not work for everyone. In many situations it remains difficult for people to verify the reliability of sources of information. This is due to factors such as the fact that scientific insights can change over time and that there is no single scientific answer to many of the issues facing society.

Furthermore, the ‘untrue’ nature of disinformation is not always relevant. In spreading disinformation, internet users may also be expressing anger or dissatisfaction over social or political issues, for example. A loss of confidence in the government could also be a factor. With respect to combatting disinformation, it

[211] Kranenburg, A. (2017). *Nederlanders bezorgd over 'nepnieuws' - een op drie weet vaak niet meer wat waar is en wat onwaar.* Volkskrant www.volkskrant.nl/nieuws-achtergrond/nederlanders-bezorgd-over-nepnieuws-een-op-drie-weet-vaak-niet-meer-wat-waar-is-en-wat-onwaar~b6914596/
[212] Plantinga, S. & Kaal, M. (2018). *Hoe mediawijs is Nederland?* www.mediawijzer.net/wp-content/uploads/sites/6/2018/09/Rapport-Mediawijsheid-volwassenen-2018.pdf

is therefore also important to realise that messages can be assessed on aspects other than just their reliability or truth.[213] [214] Efforts to increase media literacy should therefore also devote more attention to the wider context in which (dis)information plays a role.[215]

[213] Marwick, A.E. (2018). Why Do People Share Fake News? A Sociotechnical Model of Media Effects. 2 *GEO. L. TECH. REV.* 474

[214] Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the 'Post-Truth' Era. *Journal of Applied Research in Memory and Cognition, 6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

[215] Bessems, K. (2020). *Socioloog Harambam: 'We zetten complotdenkers te gauw weg als gekkies'.* Volkskrant. www.volkskrant.nl/nieuws-achtergrond/socioloog-harambam-we-zetten-complotdenkers-te-gauw-weg-als-gekkies~b1942c88/

9 Conclusions

In this concluding chapter we summarise the most importing findings of this study into the potential impact of technological developments as they relate to disinformation and into further measures that could be taken to prevent threats to public debate and the democratic process from disinformation.

9.1 A disturbing picture

We began this study with the observation that disinformation has not had a major social or political impact in the Netherlands in recent years, with the possible exception of the flood of misleading reports that have been disseminated regarding the coronavirus outbreak. But it is still too soon to pass judgement on the latter's significance for Dutch society's resilience against disinformation.

9.1.1 Wide-ranging technological possibilities for the production and dissemination of disinformation

The rapid technological developments in the field of IT could alter that situation within the foreseeable future. This study provides a broad overview of technological developments that could play a role in the production and dissemination of disinformation in the years to come. The survey is far from reassuring. The possibilities that technologies such as text synthesis, voice cloning, deepfakes, micro-targeting and chatbots create for producers and disseminators of disinformation to mislead internet users are many and varied. They range from manipulated video images that are barely distinguishable from authentic material and surreptitious influencing of voting behaviour with the help of advanced forms of micro-targeting to the spread of misleading information by means of (semi-) automated one-on-one conversations with chatbots.

Many of the technologies described are still being further developed. Among other things, this is leading to even more advanced deepfakes – which are even more difficult to distinguish from authentic visual material. That also undermines the capacity of citizens to properly evaluate information and disinformation.

Furthermore, the operations of the groups involved in the production and dissemination of disinformation are often opportunistic. They exploit vulnerabilities

in society, join in debates on sensitive social issues in the media and employ the technological options that will have the greatest effect. It is also difficult to combat the groups concerned, since it is often hard to discover who is behind a disinformation campaign, and even if their identity is suspected, proving it is another matter.

9.1.2 Combatting disinformation with technology is essential, but not enough

On the other hand, the technological developments could also yield new or improved instruments to combat disinformation. For example, through the use of artificial intelligence to detect deepfakes, with early monitoring of malicious use of micro-targeting, or by means of automatic detection of the malevolent use of chatbots.

However, a possible effect of countermeasures is that they will in turn lead to the development of even more advanced deepfakes, which are more difficult to detect, or more refined forms of micro-targeting. Countermeasures could also prompt producers and disseminators of disinformation to shift their activities to closed groups and channels to avoid control by moderators.

There is another reason why efforts to combat disinformation could fall behind. The new technologies appear to offer fewer possibilities for tackling disinformation than for producing and disseminating disinformation. For example, production and dissemination technologies benefit greatly from automation, while efforts to counter disinformation often still require human intervention to assess whether information is actually misleading.

9.2 Possible new measures

The previous chapter contained a discussion of a number of measures that could be taken to counter the threat to public debate and the democratic process from new technological developments in the field of disinformation. It was found that more is needed than the measures that are already being taken against the production and dissemination of disinformation. Better use could be made of the new opportunities that technological developments offer for tackling disinformation, for instance. At the same time, many of the measures we suggest are extensions of existing measures or their underlying principles.

In the following section we briefly summarise the new measures that were suggested.

9.2.1 Measures against deepfakes

Investing in detection of deepfakes
Platform companies could invest in an active detection policy aimed at combatting deepfakes, in order to compete in the potential race against producers and disseminators of increasingly advanced deepfakes.

Hotline for malicious image manipulation
Platform companies like YouTube, SnapChat, Instagram and TikTok, on which deepfakes are omnipresent, could institute a hotline that users can use to report (suspicions of) malicious image manipulation.

Authentication of visual material and other messages
The digital authentication of visual material and other messages would enable internet users to investigate whether material is from a source that they regard as reliable. This calls for a reliable system of registering digital hallmarks. The government and the large technology companies could take the lead in this.

9.2.2 Restricting possibilities for micro-targeting

Monitoring the use of advertising technology
Platform companies could build a monitoring function into their services to prevent abuse of the advertising technology they supply.

Restricting the technical possibilities of advertising technology
Platform companies could impose restrictions on how advertisers select target groups and monitor the responsible use of the advertising technology they supply.

Providing transparency for internet users
Platform companies could provide their users with more insight into their advertising profile and why it applies to them, as well as how the profile is used by advertisers.

9.2.3 Measures against the harmful effects of recommendation algorithms

Building a reflection period into platform services
To counter the harmful effects of the dissemination of sensational messages, platform companies could build a brief reflection period into the use of their services to deter users from sharing (dis)information impulsively.

Providing transparency about recommendation algorithms
To prevent the harmful effects of recommendation algorithms, platform companies could be transparent about how their algorithms work, starting by providing access to them for scientific researchers.

9.2.4 Warning system for closed and encrypted channels

To combat the spread of disinformation on closed and encrypted channels, an independent national warning system could be established to identify and alert people to disinformation campaigns relating to sensitive public issues. The government and the platform companies could facilitate this warning system.

9.2.5 Critically analysing the platform companies' revenue model

Measures such as detecting deepfakes, restricting the technical possibilities of advertising technology and providing transparency about how recommendation algorithms work could conflict with the revenue model of the platform companies. They might therefore have little inclination to take these measures. In that case, the government could adopt more stringent measures, such as compelling greater transparency about the use of recommendation algorithms or critically reviewing the revenue model of the platform companies.

9.2.6 Investing in fact-checking remains important

Because fact-checking provides an important assurance for internet users who are searching for reliable information, the government and platform companies could continue investing in facilities for fact-checkers.

9.2.7 Investing in media literacy remains important

Technological measures and stricter regulation of platform companies could reduce the production and dissemination of disinformation. But they will not eliminate disinformation. Little is known about what happens on closed and encrypted groups and channels managed by platform companies. And not all platform companies will be willing to take action to counter disinformation. There will still be safe havens on the internet – and that means there will still be space for disinformation.

The government must therefore continue to invest in media literacy. Internet users will continue to be confronted with disinformation and it will help if they are better equipped to deal with it.

At the same time, one should evidently not expect too much from media literacy, for example because new technological applications will make disinformation campaigns increasingly sophisticated and hence more difficult for internet users to see through.

9.3 Conclusion: platform companies are primarily responsible, but government can intervene

With many of the measures set about above, primary responsibility for tackling disinformation lies with the platform companies. For some measures, that responsibility also extends to the managers of chat channels. But given the public interest in preventing the potentially harmful effects of disinformation for public debate and the democratic process, the government could decide to act if platform companies neglect their responsibility. It could, for example, insist that the companies pursue an active detection policy to prevent deepfakes or monitor responsible use by advertisers of the possibilities of micro-targeting.

And if pressure doesn’t work, measures could be made compulsory. Those measures could be at the expense of the platform companies’ revenue model. Whether the government should adopt them will depend in part on the seriousness of the threat to public debate and the democratic process ensuing from aspects such as the polarising effect of recommendation algorithms or disinformation campaigns by advertisers facilitated by platform companies. To have sufficient effect, any compulsory measures should logically be taken at EU level.

Appendix 1: Questions for interviews

General questions

1. What is your role/function?
2. What do you understand by disinformation?
3. What are your feelings about the current situation in the Netherlands as regards disinformation?
    i. What threats do you see?
    ii. How resilient is Dutch society against disinformation?
    iii. How do you feel about the current international situation?
4. What developments do you expect in relation to the production and dissemination of disinformation in the next five years? In your opinion, what are the major threats? What are your biggest concerns?
5. In your opinion, what are the most important measures (technological and non-technological) that can be taken in the coming years to counter the threats and strengthen the resilience of Dutch society? Who is responsible for what?
6. In your opinion, how can it be avoided that possible measures impair important societal values such as freedom of speech and press freedom?

Relevant technologies

7. In our desk research we have identified the following technologies that could be relevant for the production and dissemination of disinformation:
    - Text synthesis
    - Voice cloning
    - Image synthesis, including deepfakes
    - Augmented and virtual reality and avatars
    - Memes
    - Database technology / big data / open data
    - Social media platforms, including recommendation algorithms and super apps
    - Chat apps, including encryption technology
    - Bots
    - Micro-targeting, including programmatic advertising, dynamic prospecting, campaign software, natural language processing, sentiment monitoring and influencer marketing
    - Search engines
    - Virtual assistants
    - Distributed autonomous organisations / blockchain
    - Games
    - Interactive TV and live streaming

8. In your opinion, are there any technologies missing from the list?
9. Do you regard any of these technologies as less relevant?
10. Further to the earlier ‘General questions’, which technologies do you believe will have the greatest impact?
11. Could you explain your reasons?

Measures against disinformation (further to the earlier ‘General questions’)

12. In your opinion, what measures should be taken to counter the risks arising from disinformation? What technological measures should be adopted or developed?
13. Which measures do you consider to be the most effective?
14. In that context, who is responsible for what? Which parties are best equipped to take the measures?
15. Can you give an example of a good practice?

Conclusion

16. Do you have anything to add to your earlier comments?

Appendix 2: Participants in interviews

Name	Organisation
Noëlle Aarts	Radboud University
Jarno Duursma	Trendwatcher
Joris van Hoboken	University of Amsterdam
Linus Neumann	Netzpolitik Podcaster
Cees van Riel	Rotterdam School of Management
Adam Segal	Council on Foreign Relations
Anonymous	AIVD
Anonymous	Marketing expert

Appendix 3: Participants at expert meeting

Name	Organisation
Peter Burger	University of Leiden
Ufuk Esmer	BKB
Henriette Kieviet	Dutch Media Literacy Network
Peter Olsthoorn	Freelance journalist
Claes de Vreese	University of Amsterdam
Anonymous	Authority for Consumers and Markets

Appendix 4: List of technologies

General technologies

Technology	Description
Database technology	Large-scale collection and analysis of (personal) data
Artificial intelligence	Self-learning algorithms and systems

Production technologies

Technology	Description
Text synthesis	Algorithms that generate readable and logical (news) messages
Voice cloning	Manipulation of voice messages using artificial intelligence
Image synthesis and deepfakes	Generation and modification of videos using artificial intelligence
AR, VR and avatars	Presentation of information in a virtual environment
Memes	Images intended to be widely shared on social media

Dissemination technologies

Technology	Description
Social media platforms	Online platforms such as Facebook, Twitter and TikTok
Micro-targeting	Reaching specific target groups with a message tailored to them
- Campaign software	(Partially) automatic control of micro-targeting
- Dynamic prospecting	Automatic selection of target groups
- Programmatic advertising	Automatic tailoring of messages to target groups
- Psychographing	Automatic analysis of personality traits
- Influencer marketing	Dissemination of messages via accounts on social media with many followers
Chat apps	Sharing of (encrypted) messages, one-to-one or in small groups
Bots	(Partially) automatic control of accounts on social media
Search engines	Platforms that enable the internet to be searched, analyse search behaviour and display advertisements
Virtual assistants	Voice-controlled devices with which search engines can be consulted, among other things
Distributed autonomous applications	Online platforms without central control
Games	Online games
Cross-media storytelling	Reaching a specific person or target group via various channels and devices

© Rathenau Instituut 2020
This work or parts of it may be reproduced and/or published for creative, personal or educational purposes, provided that no copies are made or used for commercial objectives, and subject to the condition that copies always give the full attribution. In all other cases, no part of this publication may be reproduced and/or published by means of print, photocopy, or any other medium without prior written consent.

Open Access
The Rathenau Instituut has an Open Access policy. Reports, background studies, research articles and software are all open access publications. Research data are made available pursuant to statutory provisions and ethical research standards concerning the rights of third parties, privacy, and copyright.

Contact details
Anna van Saksenlaan 51
P.O Box 95366
2509 CJ The Hague
070-342 15 42
info@rathenau.nl
www.rathenau.nl

Board of the Rathenau Institute
Gerdi Verbeet
Noelle Aarts
Felix Cohen
Hans Dröge
Laurence Guérin
Janneke Hoekstra
Erwin Muller
Rajash Rawal
Peter-Paul Verbeek
Melanie Peters - secretary

Het Rathenau Instituut stimuleert de publieke en politieke meningsvorming over de maatschappelijke aspecten van wetenschap en technologie. We doen onderzoek en organiseren het debat over wetenschap, innovatie en nieuwe technologieën.

Rathenau Instituut