Rathenau Instituut

Harmful Behaviour Online

An investigation of harmful and immoral behaviour online in the Netherlands

Report

Authors

Mariëtte van Huijstee, Wouter Nieuwenhuizen, Mathilde Sanders, Eef Masson, and Pieter van Boheemen

Cover image
Shutterstock

Preferred citation:
Rathenau Instituut (2022). *Harmful Behaviour Online – An investigation of harmful and immoral behaviour online in the Netherlands*. Den Haag (authors: Huijstee, M. van, W. Nieuwenhuizen, M. Sanders, E. Masson, and P. van Boheemen)

Original publication
Rathenau Instituut (2021). *Online ontspoord – Een verkenning van schadelijk en immoreel gedrag op het internet in Nederland*. Den Haag (auteurs: Huijstee, M. van, W. Nieuwenhuizen, M. Sanders, E. Masson en P. van Boheemen)

To our deepest regrets, our director Melanie Peters died on 11 August 2021. She was the director of Rathenau Instituut at the time of the original publication in Dutch. Her preface for the original report was maintained for this translated publication.

Research for the original publication was conducted between December 2019 and July 2021.

Preface

Paedophile hunting. Phishing. Cyber addiction. Revenge porn. Disinformation. These are just a few examples of harmful and immoral behaviour online. Certain properties of the internet – the virality of online messages, the (perceived) anonymity of internet users, and the immediacy with which a video can be viewed worldwide – facilitate these kinds of behavioural phenomena. They can be enormously harmful for both individuals and society.

But what kind of behaviour are we talking about? How prevalent is it? And what can the Dutch government do about it? At the request of the Research and Documentation Centre for the Dutch Ministry of Justice and Security (WODC), the Rathenau Instituut has studied harmful and immoral behaviour online. We have interviewed and talked to experts in the fields of policymaking, scholarship and professional practice and reviewed academic and journalistic sources and policy documents. Based on our findings, the expertise acquired in previous research and analysis by the Rathenau Instituut, we introduce a taxonomy of six categories of harmful and immoral behaviour online divided into 22 different phenomena, from online manipulation of information to hate speech and self-harm.

Our study reveals that, sooner or later, all internet users in the Netherlands may encounter harmful and immoral behaviour online. People lack adequate protection online, and their fundamental rights are at stake. For a long time, the internet appeared to be a domain of self-regulation and self-reliance. To counteract harmful and immoral behaviour, however, we need an active government, one that not only reacts to derailments but that also proactively intervenes in the online environment to prevent harm and to protect people's fundamental rights. We present a strategic agenda that will enable the Dutch government, in collaboration with businesses, civil society organisations and the public, to get a grip on harmful and immoral behaviour online.

The Rathenau Instituut has been studying the impact of technology on society for 35 years. The purpose of the present report is to contribute to the societal debate about what constitutes desirable and permissible behaviour online. New phenomena continue to emerge online and moral standards are subject to change. These factors make it all the more necessary to have an active government and a public debate on harmful and immoral behaviour online.

Dr. ir. Melanie Peters †
Director Rathenau Instituut

Summary

Introduction
The internet has certain properties that tend to derail online behaviour. A person who would never insult a passer-by on the street may have no trouble doing so on Twitter. Someone who would never steal from the local supermarket may feel less inhibited about stealing credit card information online. In their book Evil Online (2018), Dean Cocking and Jeroen van den Hoven describe the internet as an environment in which harmful and immoral behaviour is inspired, facilitated and encouraged. This book made the Dutch Ministry of Justice and Security wonder what the status of such behaviour is in the Netherlands.

The Ministry's Research and Documentation Centre (WODC) asked the Rathenau Instituut to answer the following research question: What is the nature and scale of harmful and immoral behaviour online in the Netherlands, what are the underlying mechanisms and causes, and what options for action are available to the Ministry, and the government as a whole, for limiting harmful and immoral behaviour online?

Our report addresses online behaviour that takes place in a moral twilight zone, and in which the government is currently hesitant to act. We looked at online behaviour that can be designated as harmful and/or immoral. This behaviour is harmful not only to individuals but also to larger groups or society as a whole. Some of the behaviours that we discuss in this study violate certain fundamental rights and laws and are therefore unlawful or illegal. Yet it turns out that it is much more difficult for people to judge whether something is acceptable in an online environment. The online world is not necessarily more lawless or more of a free-for-all than the offline world, but it is more easily experienced as such.

In this report, the Rathenau Instituut uses a taxonomy to present a unique overview of harmful and immoral online behaviour in the Netherlands. This taxonomy can serve as a framework for a coordinated approach by the national government, in collaboration with the business community and stakeholders in civil society. A further aim is to contribute to the societal debate on what constitutes desirable and permissible behaviour online. We know that moral standards are subject to change and that public debate about these standards is necessary.

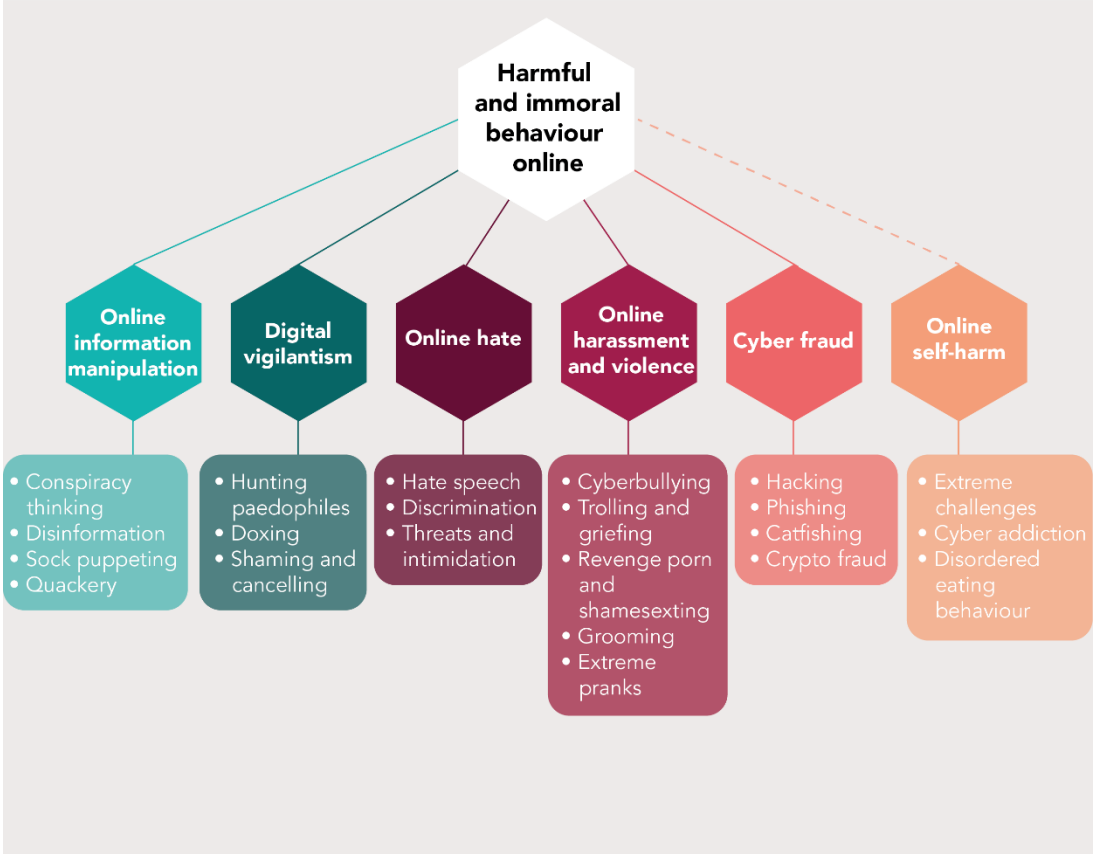
Approach
The report addresses the following sub-questions:

1. What is the taxonomy of online behaviours and online phenomena that can be harmful to individuals or groups, and thus affect the moral fabric of society?
2. What is the nature of these problematic behaviours and phenomena in the Netherlands?
3. What is the scale of problematic behaviours and phenomena in the Netherlands, in terms of stakeholders, victims and damage to society?
4. How are these problematic behaviours and phenomena, and the resulting damage to society, linked to the operation, underlying mechanisms and design of the online environment? In other words: how does the online world act as a facilitator and catalyst for harmful statements and behaviour on the internet and social media?
5. What options for action have already been developed, nationally and internationally, for limiting harmful and immoral behaviour online and the societal damage it causes, and what lessons can be learned from them?
6. What options for action does the Dutch government have?

To answer these sub-questions, we combined the following methods: literature review, interviews, workshops and meetings with experts from the fields of policymaking, professional practice and scholarship. A total of 56 such experts contributed to the study.

Taxonomy, nature and scale

This study is the first to map all aspects of harmful and immoral behaviour online in the Netherlands. The Rathenau Instituut developed a taxonomy of six categories of harmful and immoral behaviour online, listing 22 different phenomena that all internet users in the Netherlands may encounter sooner or later.



Bron: Rathenau Instituut

Figure 1 Taxonomy of harmful and immoral behaviour online

Taxonomy of harmful and immoral behaviour online[1]

The harmful behaviour listed in this taxonomy can severely impact individuals, groups and society as a whole. It can range from a teenage girl starving herself because she joins an extreme challenge with other adolescents or female journalists and scientists being discouraged from speaking out online in fear of harassment, to societal disruption due to the spread of conspiracy theories and disinformation.

Interviews with experts and the literature on the nature and scale of the phenomena listed in the taxonomy make clear that all Dutch people run the risk of becoming involved in this behaviour as a victim, perpetrator or bystander. Anyone can be affected by the harmful and immoral behaviour outlined in this report. However, for certain phenomena, some groups are more at risk than others, depending on their age, gender, race, sexual orientation, religious beliefs or level of education. It is difficult to generalise based on the available data.

1 See the glossary at the start of this report for definitions of the phenomena.

The study shows that, to date, accurate definitions and systematic measurements are lacking for various phenomena. It is not useful to try to determine which phenomenon is the most worrying, as this depends on the criteria chosen: the number of victims, the severity of the harm, or the possible harm in the future. We conclude that all phenomena are worrisome in their own way, for society as a whole, for individuals or for groups of individuals.

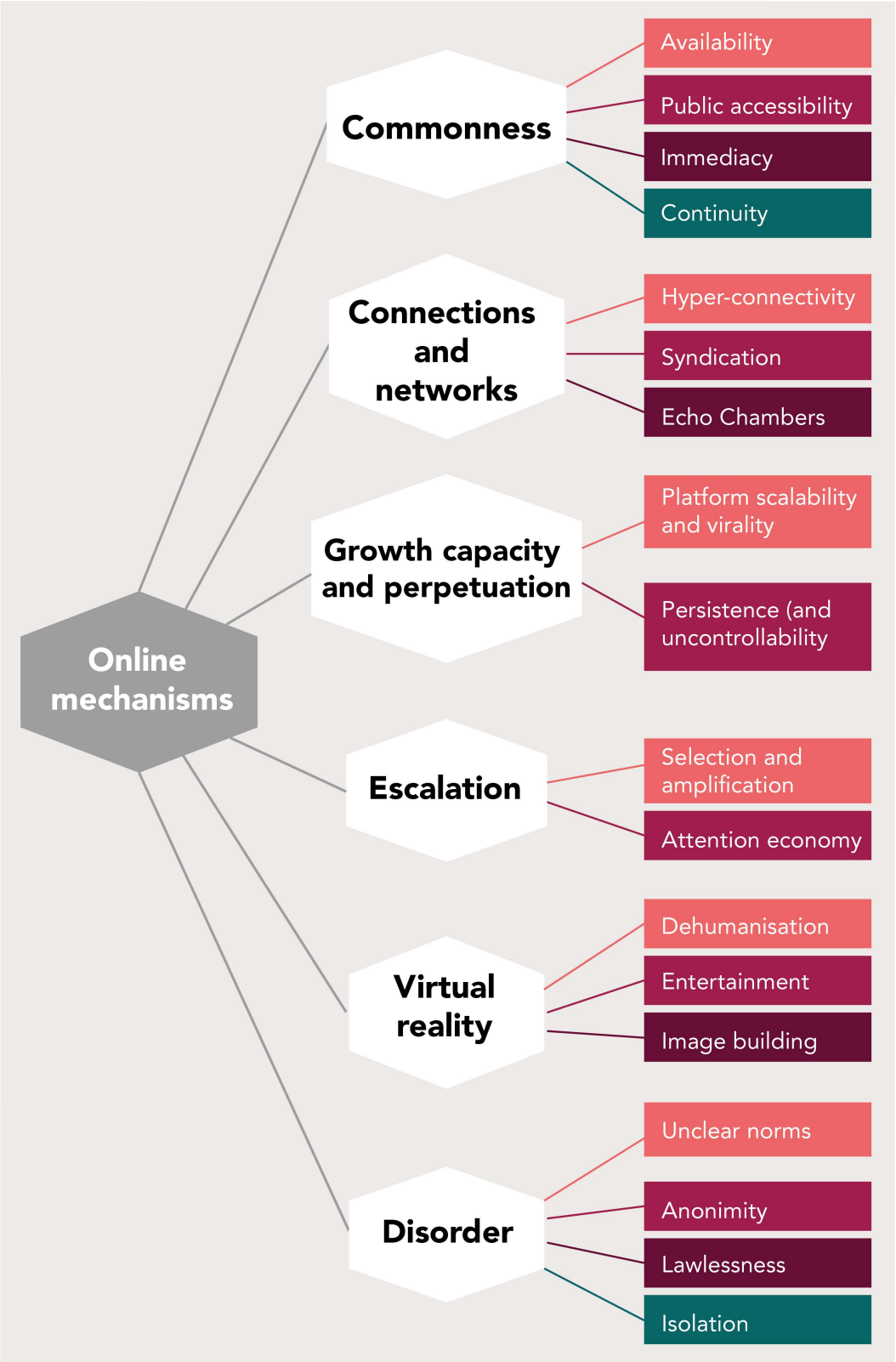
Mechanisms

Certain mechanisms of the internet's online environment influence human behaviour. These online mechanisms may cause people to deal with values and rules differently online than offline. Besides the mechanisms of the internet, many other factors influence human behaviour, including social, psychological, cultural and economic ones. All these factors play a role in the development of harmful and immoral behaviour online. This report focuses on mechanisms characteristic of the internet.

The study identified a total of 18 online properties and mechanisms that play a role in inspiring, facilitating and driving harmful and immoral behaviour online: 1) availability, 2) public accessibility, 3) immediacy, 4) continuity, 5) hyper-connectivity, 6) syndication, 7) echo chambers, 8) platform scalability and virality, 9) persistence (and uncontrollability), 10) selection and amplification, 11) attention economy, 12) dehumanisation, 13) entertainment, 14) image building, 15) unclear norms, 16) anonymity, 17) (apparent) lawlessness, 18) isolation. These mechanisms are grouped under six descriptive characteristics of the internet:

1. Commonness
2. Connections and networks
3. Growth capacity and perpetuation
4. Escalation
5. Virtual reality
6. Disorder

An overview of all mechanisms and their classification can be found in Figure 2 below.



Bron: Rathenau Instituut

Figure 2 Overview of online mechanisms

The case studies in this report illustrate that the same mechanisms can play a role in very different phenomena, and that the mechanisms occur in combination. For example, syndication (the ease of finding like-minded people online) and virality (rapid, uncontrollable distribution of content online) play a role in the online shaming case, the disinformation case and the disturbed eating behaviour case. Intervening in the mechanisms, such as requiring transparency about the recommendation algorithms for online content or lifting the anonymity of internet users in certain environments, makes sense in preventing or reducing harmful and immoral behaviour online. But such interventions require careful consideration and societal debate. After all, the mechanisms of the internet can also lead to socially desirable behaviour and social merits. Anonymity online, for example, enables whistle-blowers to report societal malpractices. Intervening in these mechanisms may also limit or nullify these positive effects.

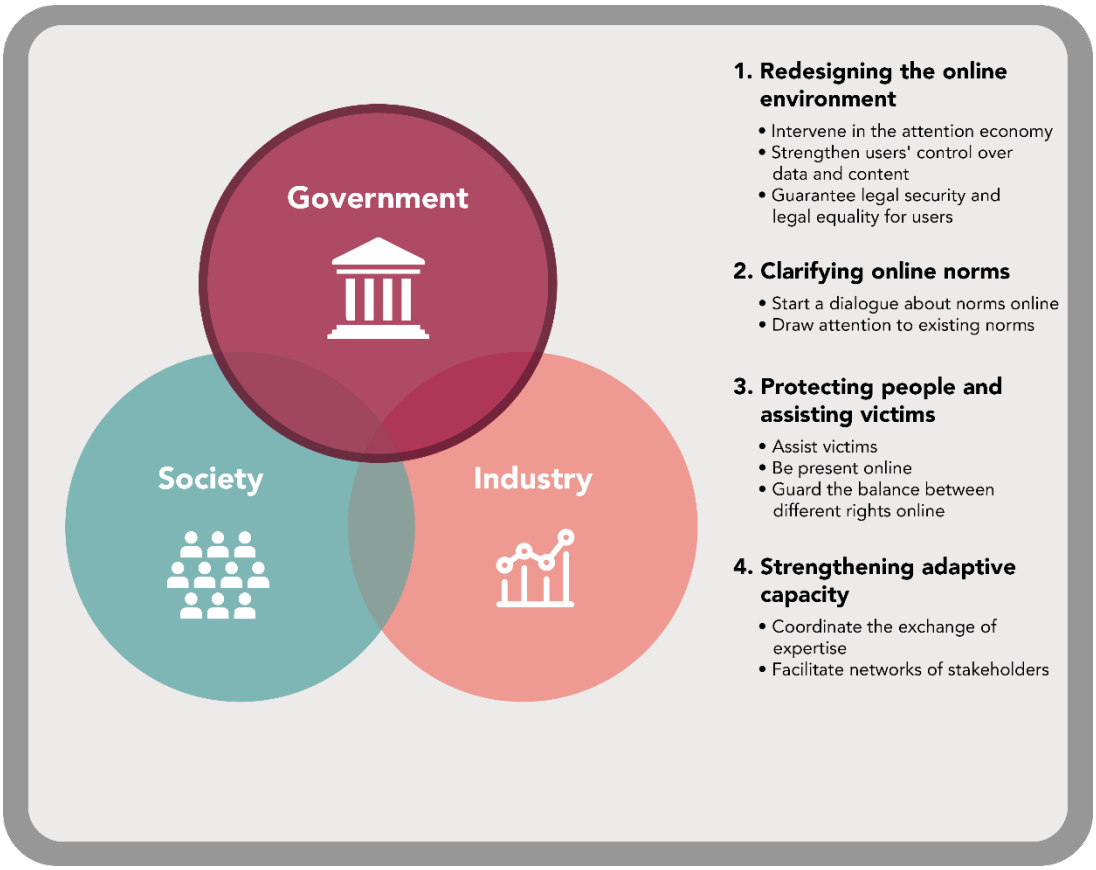
Options for action

So far, the internet has been a domain of self-regulation and self-reliance, where the government has taken no responsibility for oversight and users have managed by themselves. However, our study shows that fundamental rights are at stake; citizens are insufficiently protected on the internet. Businesses, civil society organisations and citizens need an active government to counter harmful and immoral behaviour online, and to promote socially desirable behaviour online.

This report provides an overview of existing measures that governments, businesses, social workers and others have already taken to tackle harmful behaviour online. The overview reveals which interventions already work and seem promising when it comes to reducing or preventing harmful and immoral behaviour online. But it also reveals the shortcomings and where additional interventions are needed. The most important observation is that many of the current initiatives are mainly *reactive* in nature. They are aimed mainly at combating the symptoms of harmful and immoral behaviour rather than at the underlying mechanisms. In this respect, we see differences between various stakeholders. Governments and large platform companies in particular have not been very proactive. In the case of platform companies, this is not surprising. After all, tinkering with mechanisms means choosing an alternative form of platform design. This results in uncertainties about business models, and because companies operate in a competitive market, it is primarily other, small businesses and stakeholders that are experimenting with alternative forms of design.

Our analysis of existing measures shows that governments mainly take action when behaviours get out of hand and, therefore, need to be restrained. Up to now, their interventions have mainly been reactive. The overview of online mechanisms in this report can help governments and other stakeholders to be more proactive.

We introduce a strategic agenda for the Dutch government based on interviews and discussions with experts in the fields of policymaking, scholarship and professional practice, a review of academic and journalistic sources and policy documents, and the expertise gained by the Rathenau Instituut in previous research and analysis. In this agenda, we identify four themes in which the Dutch government can cooperate with stakeholders from industry and society to play a guiding, coordinating and facilitating role in tackling harmful and immoral online behaviour and promoting a safe online environment.



Bron: Rathenau Instituut

Figure 3 Strategic agenda for tackling harmful and immoral behaviour online

The first theme – *Redesigning the online environment* – contains options for the Dutch government to address the online mechanisms that contribute to harmful and immoral behaviour online. For example, the report makes a number of suggestions to intervene in the online attention economy. The second theme – *Clarifying online norms* – deals with the role of the Dutch government, industry and society in renewing the social agreements on norms and values online. The options for action under this theme are intended to bring about broader awareness and understanding

in society at large of harmful and immoral behaviour online. The third theme – *Protecting people and assisting victims* – contains suggestions for the Dutch government, law enforcement and executive agencies to better respond to the phenomena of harmful and immoral behaviour online and the damage they cause. For example, we make a number of suggestions for the government to be more visible and present online. The fourth theme – *Strengthening adaptive capacity* – offers suggestions for the Dutch government to gain and maintain a grip on harmful and immoral online behaviour, which is constantly changing. These options for action are aimed at future-proofing the strategic agenda.

Contents

Glossary

The following definitions are based on academic research and journalistic sources. Source references can be found in Chapter 3, where we present our taxonomy of harmful and immoral behaviour online.

Cancelling: online naming and shaming that calls for someone to be excluded from their community as a form of social punishment.

Catfishing: extreme form of online dating deception that involves falsely representing oneself to a potential romantic partner, without the intention of meeting in person.

Conspiracy thinking: the belief that certain events or situations are not accidental but have been secretly manipulated behind the scenes by powerful forces with evil intentions.

Cyber addiction: excessive and uncontrolled online activity with prolonged internet use, especially in social networking, online gaming and use of pornography sites.

Cyberchondria: unnecessary panic or anxiety induced by excessively or repeatedly reviewing morbid or alarming content during health-related searches online.

Cryptofraud: a form of deception that sometimes involves pyramiding trading in which people are persuaded to buy or sell cryptocurrency in order to boost its price.

Cyberbullying: repeated and intentional online bullying by a group or individual against a victim who has difficulty defending themself.

Challenges: encouraging people to complete certain (dangerous) tasks and then share a video of themselves doing so online.

Disinformation: the dissemination of information that is ‘inaccurate’ or ‘misleading’ with malicious or harmful intent**.**

Digital vigilantism: a form of collective action, moral censure or rebuke (e.g. using online shaming, harassment or doxing) targeting individuals who exhibit undesirable social behaviour.

Doxing: the public release of an individual's private, sensitive, personal information, for example their home address, telephone number, passport number or employer's contact information, family members' contact details, and photographs of their children.

Extreme pranks: a form of interpersonal humiliation involving a three-way relationship between the one who humiliates, the victim and the witnesses. Online, the pranks often take the form of recorded 'offline' pranks, with the camera zooming in on the victim's response (confusion, shock, distress or embarrassment).

Griefing: intentionally annoying other players in online games by manipulating certain game elements in a way that affects other players.

Grooming: the process whereby an adult develops a sexually abusive relationship with a minor through the use of cybertechnology, for example via social media. It is also referred to as Online Grooming (OG).

Hate speech: all forms of expression that spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance.

Hacking: activities involved in attempting or gaining unauthorised access to IT systems.

Quackery: the unauthorised practising of medicine by someone who claims to be able to cure an illness with a useless or even harmful remedy.

Paedophile hunting: a form of digital vigilantism whereby people pose as children to 'trap' paedophiles and then either punish them themselves or report them to the police.

Phishing: fraudulently acquiring information about persons and organisations by emailing users a fake version of a popular website to trick them into providing sensitive details.

Pro-ana coach: person who encourages young girls (minors) with an eating disorder to lose more weight, usually for the purpose of obtaining sexually explicit material from them.

Sexting: a practice whereby people distribute or share sexually explicit messages, photos or videos of themselves on mobile phones or other devices.

Sextortion: a form of extortion in which someone threatens to disseminate images of a sexual nature without the victim's consent in order to procure additional images, money, or sexual acts.

Shaming: a form of digital vigilantism in which public moral criticism is expressed online in response to violations of social norms.

Shame sexting: making and/or distributing sexually explicit images or videos without the subject's consent.

Stalking: repeatedly intimidating, harassing and sometimes threatening victims.

Sock puppeting: a false online identity (sock puppet) used for purposes of deception. A sock puppet can be used on social media for catfishing or trolling, for example.

Trolling: intentionally disrupting online communities by behaving in a way that is deemed unacceptable, such as calling people names, picking fights or making negative comments about others. There is also a broader interpretation of trolling as a concept, namely the use of fake accounts to spread disinformation and influence the public debate.

Revenge porn: the unauthorised possession, disclosure and distribution of stolen sexual images, for example by hackers, partners, ex-partners, child abusers, rapists and human traffickers.

Disturbed eating behaviour (eating disorders): psychological disorders characterised by disordered eating behaviour and/or compensatory behaviour (self-induced vomiting, laxative misuse). People with an eating disorder have a distorted body image, are obsessed with their weight or body shape, and are terrified of gaining weight.

1 Introduction

The internet has certain properties that tend inspire, facilitate and catalyse harmful behaviour online. A person who would never insult a passer-by on the street may have no trouble doing so on Twitter. Someone who would never steal from the local supermarket may feel less inhibited about stealing credit card information online. The book *Evil Online* (Cocking & van den Hoven, 2018) describes the internet as an environment in which harmful and immoral behaviour is inspired, facilitated and encouraged. This book made the Dutch Ministry of Justice and Security wonder what the status of such harmful behaviour is in the Netherlands. The Ministry asked its Research and Documentation Centre (WODC) to examine the following key question: *What is the nature and scale of harmful and immoral behaviour online in the Netherlands, what are the underlying mechanisms and causes, and what options for action are available to the Ministry, and the government as a whole, for limiting harmful and immoral behaviour online?*

The WODC asked the Rathenau Instituut to undertake this study. The Rathenau Instituut's mission is to map out the effects of technologies on society and to propose options for action aimed at protecting the public interest. The principle underpinning all our research is that public values must be protected in the face of technological advances. We ask: which values are at stake, and what is the role of government, citizens and the business community in protecting these values? Based on our mission, Rathenau Instituut has gained extensive experience in studying the harmful effects of online technologies and proposing options for action.

The main question in this study can be broken down into a number of sub-questions, which are addressed in successive chapters of this report.

1. In general terms, what is the taxonomy of online behaviours and online phenomena that can be harmful to individuals or groups, and thus affect the moral fabric of society?
2. What is the nature of these problematic behaviours and phenomena in the Netherlands?
3. What is the scale of problematic behaviours and phenomena in the Netherlands, in terms of stakeholders, victims and societal damage?
4. How are these problematic behaviours and phenomena, and the resulting damage to society, linked to the operation, underlying mechanisms and design of the online environment? In other words: how does the online world

act as a facilitator and catalyst for harmful statements and behaviour on the internet and social media?

5. What options for action have already been developed, nationally and internationally, for limiting harmful and immoral behaviour online and the societal damage it causes, and what lessons can be learned from them?
6. What options for action appear to be appropriate for the Dutch government?

The concepts ‘online’ and ‘harmful and immoral’ require further clarification.

1.1 Online

Our daily lives are bound up with the internet in countless ways. As a result, it can be hard to distinguish between online and offline at times. What we do online also has an impact offline, and vice versa. However, it is possible to distinguish between actions that could not occur without the internet (such as managing a company website, posting photos and chatting on social media, or gaming online) and behaviour that does not require the internet (such as going for a walk in the woods). There is also behaviour that existed before the advent of the internet (such as bullying), but that now has an online equivalent (cyberbullying). In this study, we look at harmful and immoral behaviour online. In other words, we study cyberbullying as opposed to all forms of bullying, and online discrimination as opposed to discrimination in a more general sense. We also consider the difference between offline and online behaviour and the underlying mechanisms that characterise the internet and influence online behaviour.

1.2 Harmful and immoral

In this study, we investigate online behaviour that can be designated as harmful and/or immoral, where the social and moral boundaries are not always clear to online users. This behaviour is harmful not only to individuals but also to larger groups or society as a whole. As a society, we have laid down moral standards in laws and rules, in human rights conventions and in implicit social agreements. Rights, including fundamental rights, are valid both online and offline. For example, people wanting to express themselves online (for example in a pro-ana blog or tweet) enjoy freedom of expression, but others (for example those who get carried away by pro-ana content and whose health is consequently impaired) also have rights and interests. Freedom of expression does not protect every utterance: hate speech, for example, is not protected.

Some of the behaviours that we discuss in this report may violate certain fundamental rights and laws and are therefore unlawful or illegal. While a court may find the boundaries of what is permissible online clear, it appears to be much harder for people to judge when something is acceptable online. The online world is thus not necessarily more lawless or more of a free-for-all than the offline world, but it is more easily experienced as such. Differences in moral standards and moral confusion – the 'moral fog' (Cocking & van den Hoven, 2018) – can have all kinds of harmful consequences in the online environment. This study focuses on behaviour in this moral twilight zone. The government is still hesitant to act and in search of an appropriate way to protect fundamental rights online. In undertaking this study, the Rathenau Instituut is helping to develop a set of tools for government and contributing to the societal debate on what is desirable online. Moral standards are subject to change and public debate about these standards is necessary.

1.3 Scope

This study is about harmful and immoral behaviour online, with the internet facilitating, encouraging or inspiring such behaviour. We do not examine all potentially harmful or immoral online behaviour, nor do we consider all potentially moral, positive, altruistic online behaviour. Neither do we include all of the mechanisms that may play a role in harmful and immoral behaviour online, such as social, psychological or economic factors. We focus solely on mechanisms that are specific to the internet. We have categorised the phenomena that fall within the scope of the study into a taxonomy consisting of six categories and a total of 22 phenomena of harmful and immoral behaviour online. We have also identified 17 underlying mechanisms and developed a strategic agenda for the government based on four themes.

We would ask our readers to bear the following in mind:

1. This is an exploratory study that covers a broad scope. Its findings should therefore not be regarded as exhaustive, but as an initial step towards raising awareness of the dark side of the internet and how we, as human beings and as a society, should deal with it.
2. The fact that this study focuses on harmful online behaviour does not mean that the internet only facilitates harmful behaviour. Online connectivity also has many positive sides to it, such as the potential to reach larger social groups and to facilitate more direct interaction between government and citizens. These positive aspects are beyond the scope of this study, but should be taken into account when designing measures to counter the internet's harmful effects.

New terms and new phenomena associated with harmful and immoral behaviour online continue to emerge, as we found out in the course of this study. For example, ‘paedophile hunting’ was in the public spotlight when we began the study, but interest in it had waned by the time we had finished. As we wrapped up our research, crypto-speculation emerged as a topic of interest. Even so, we expect our taxonomy, overview of underlying mechanisms and options for action to remain relevant for a long time to come.

1.4 Reader’s guide

This report consists of the following sections:

- A description of our methodology, including an explanation of each phase of the study (Chapter 2).
- Several case studies of harmful and immoral behaviour online (interspersed between chapters). The case studies illustrate various phenomena and how the underlying mechanisms operate. They also clarify the roles of the various different stakeholders involved in or affected by online behaviour.
- A taxonomy of harmful and immoral behaviour online outlining the nature and (where possible) scale of the relevant phenomena in the Netherlands (Chapter 3). The taxonomy consists of 22 phenomena, divided into six categories. This chapter answers research questions 1 to 3.
- An overview and discussion of the online mechanisms behind harmful and immoral behaviour online (Chapter 4). Here, the focus is on the mechanisms designated as specific to the online environment. This chapter answers research question 4.
- An overview of existing initiatives that are meant to prevent or counter harmful and immoral behaviour online (Chapter 5). We discuss these in terms of type of approach and from the perspective of the responsible stakeholder (e.g. government, business, civil society organisation or social welfare service). This chapter answers research question 5.
- A strategic agenda suggesting options for action to be taken by the Dutch government in collaboration with stakeholders in the private sector and in society to prevent, limit and/or remedy harmful and immoral behaviour online (Chapter 6). This chapter answers research question 6.

2 Approach

This study investigates various sub-questions using a combination of methods in each case. After establishing the research design, the Rathenau Instituut defined the scope of the study. Input was provided by the advisory committee and officials in various ministries.

We then conducted a literature review to determine the nature, scale and causes of harmful and immoral behaviour online. The review covered academic literature, policy papers and reports commissioned by policymakers, journalistic reports and online platform codes of conduct. In addition, we interviewed experts from the fields of scholarship, policymaking and professional practice. A theme-by-theme analysis of all these sources resulted in a taxonomy of harmful and immoral behaviour online. Finally, we prepared an overview of options for action and solution categories, once again drawing on the literature and on an exploratory workshop with scholars, policymakers and professional practitioners. All findings were then submitted to a number of experts at a validation meeting, which led to their further refinement and enrichment. In various phases of the study, we consulted the advisory committee established by the Research and Documentation Centre (WODC) (see Appendix 1). A total of 56 experts from the fields of scholarship, policymaking and professional practice contributed to the study. A more detailed description of our methods follows.

2.1 Exploration

After establishing the research design, we kicked off the study with an exploratory workshop with officials from ministries, law enforcement and social welfare organisations. The aim of the workshop was to gain a better understanding of the knowledge requirements and to help us frame our research topic. The programme and participants are given in Appendix 2.

2.2 Literature review

The case studies described in the book *Evil Online* served as the starting point. This produced a list of phenomena and mechanisms that we used as search terms in databases of scholarly literature and news media. We used the snowball method on this list to find more relevant articles and reports. A source was considered

relevant if it examined the nature of the phenomena, presented statistical data on the scale of these phenomena in the Netherlands, explained the causes and mechanisms, or offered ideas about possible actions that could be taken. An analysis of guidelines and rules of conduct imposed by online platforms also yielded dozens of terms that fall under the heading ‘harmful and immoral behaviour online’. The results of the literature review have been integrated into all of the chapters and the sources are listed in the bibliography.

2.3 Interviews

We supplemented the literature review by conducting 15 interviews with experts (see Appendix 3). We selected the interviewees to ensure that as a whole, the interviews would address the full breadth of the phenomena and the underlying causes and mechanisms. The interview guide is included in Appendix 4. The results of the interviews have been incorporated into the various chapters.

2.4 Workshop

On 13 April 2021, the Rathenau Instituut research team organised a workshop on options for tackling harmful and immoral behaviour online. The literature review and the interviews led to five solution categories that received considerable support but had not been developed into actual initiatives. The aim of the workshop was to flesh out these solution categories by encouraging a dialogue between staff members from different ministries, law enforcement and social welfare organisations, researchers, representatives of civil society organisations and others with relevant expertise. There were 22 participants in the workshop, divided into the five solution categories (see Appendix 5):

- online monitoring and assistance
- conversation about norms online
- value-sensitive platform design
- technical solutions
- enforcement of laws and rules online.

The findings have been incorporated into the strategic agenda presented in Chapter 6.

2.5 Validation meeting

On 26 May 2021, the Rathenau Instituut research team organised an expert meeting to validate the research results. The meeting was attended by the researchers and the staff of various executive agencies and civil society organisations.

Prior to the meeting, the participants were sent a summary of the study (approximately 20 pages). They were given the opportunity to comment on the research results during the meeting. The focus there was on the options for action arising from the analysis given in the report. The purpose of this exercise was to work with the attendees on prioritising options for action and reflecting on the role that different parties can play in tackling harmful and immoral behaviour online. Seven people attended the validation meeting (see Appendix 6). The findings have been incorporated throughout the report.

2.6 Advisory committee

This report was produced at the request of the Ministry of Justice and Security's Research and Documentation Centre (WODC). At the start of the study, an advisory committee was installed consisting of a representative of the Ministry of Justice and Security, the Ministry's Research and Documentation Centre (WODC) and three experts (for a list, see Appendix 1). The advisory committee played an advisory role and met with the research team on four occasions in different phases of the study. Its recommendations have been incorporated into the report at the Rathenau Instituut's discretion. The Rathenau Instituut is solely responsible for the contents of the report.

Case: Online shaming

This case is about online shaming, a manifestation of digital vigilantism. It is an adaptation of a real-life case. We start by describing the events and then reflect on the roles of the various stakeholders. The online mechanisms that are instrumental in this case are shown in bold and will be explained in more detail in Chapter 4.

Case

Manu was sexually abused last year by a man who still holds an important position in society. The police advised Manu not to report the abuse due to a lack of evidence. Manu himself has reason to believe that this man is victimising others, but he does not know that for certain. He has lost faith in law and order. After much hesitation, Manu decides to go public on social media to warn others and prevent the perpetrator from claiming any more victims.

Although Manu has only a few followers on social media, his post is shared by someone with **much more reach**. More victims start sharing their stories about the same perpetrator. Manu is shocked by the sudden **media attention** that the case is generating in newspapers and on TV but feels supported by the stories that others are sharing. It turns out that other people contacted the police with allegations against the same man, but it never came to an investigation or prosecution. People start to turn against the man on social media and someone publishes his home address online (doxing). He receives threats and his employer orders him to take a temporary leave of absence (cancelling).

Manu is also accused of being a liar and attention-seeker, triggering him to relive the trauma of believing that the abuse is his own fault. The social media platform decides to **hide posts about the case in users' timelines** because they violate its rules about slander, libel or defamation. A police spokesperson says on a popular talk show that no charges have ever been filed against the accused. Victims express outrage on social media because the police had in fact advised them not to report the crime. Their faith in law and order dwindles further.

Reflection

Several stakeholders play a role in this case. Besides Manu and the person he has accused, they also include bystanders, social media platforms and the police, all of whom influence the consequences of Manu's actions. It is difficult to distinguish perpetrator from victim in this example because both Manu and the man he has accused can be viewed as both victim and perpetrator. Manu is a victim of sexual abuse, but he can also be regarded as a 'perpetrator' because of his role in disclosing

the allegations. Manu is not out to hurt the accused, but at the same time he wants to prevent him from victimising others. This shows that harmful online behaviour is not necessarily clear-cut and often difficult for bystanders to judge.

The case illustrates how quickly situations can escalate on the internet due to the **uncontrollability** and **persistence** of information once it is online. Various online mechanisms play a role here, such as the **scalability** and **virality** of online platforms, the **public accessibility** of the internet and the **uncivil behaviour** that people often display there. In this particular case, it is especially important to bear in mind that people often lose control of the situation online. Once his message goes viral, Manu no longer controls how it is disseminated. Other people track down the details of the man Manu is accusing and threaten him, while the many posts expressing doubt about his story cause Manu himself to relive his trauma.

The **immediacy** of the internet – with online behaviour having an instantaneous impact – leaves the accused neither time nor opportunity to defend himself. Condemnation is swift: his employer places him on suspension and the traditional media quickly pick up the story. This turn of events differs significantly from what would have happened if the police and public prosecutor had investigated his behaviour. The fact that it is happening online also allows various stakeholders to immediately jump into the fray. Bystanders support Manu's side of the story or cast doubt on it by sharing other people's posts or by posting themselves. **Syndication**, the ease of finding like-minded people online, plays a major role here. Depending on the online social environment that people inhabit, they may be expected to behave in certain ways. By speaking out against injustice, they demonstrate that they are on the right side of the moral divide.

In this case, the social media platform decided to hide messages that were spreading about Manu's case, thus intervening in their **amplification** online. This type of balancing act, between freedom of expression and potential harm, is difficult for platforms and they are subject to frequent criticism. It is in part because the police decided not to conduct an investigation after talking to Manu that he eventually took matters into his own hands. Feelings of dissatisfaction and misgivings as to whether the criminal justice system will in fact deliver justice are factors in every form of digital vigilantism.

3 Taxonomy of harmful and immoral behaviour online

In this chapter, we describe the nature and scale of harmful and immoral behaviour online based on a taxonomy developed by the Rathenau Instituut. To get to grips with the changing and multifaceted topic of 'harmful and immoral behaviour online', the research team developed a taxonomy that categorises types of harmful and immoral behaviour and classifies specific phenomena.

We began by combining some phenomena identified in the research that proved difficult to differentiate. We then sorted the phenomena by theme and considered the main motive behind the behaviour (for example, taking matters into one's own hands, sadism) and the victims' traits. This was an iterative process, with the research team working on the taxonomy until the end of the study.

We start this chapter by briefly explaining the taxonomy. We then present the six categories of the taxonomy and describe the nature and scale of different forms of harmful behaviour online.

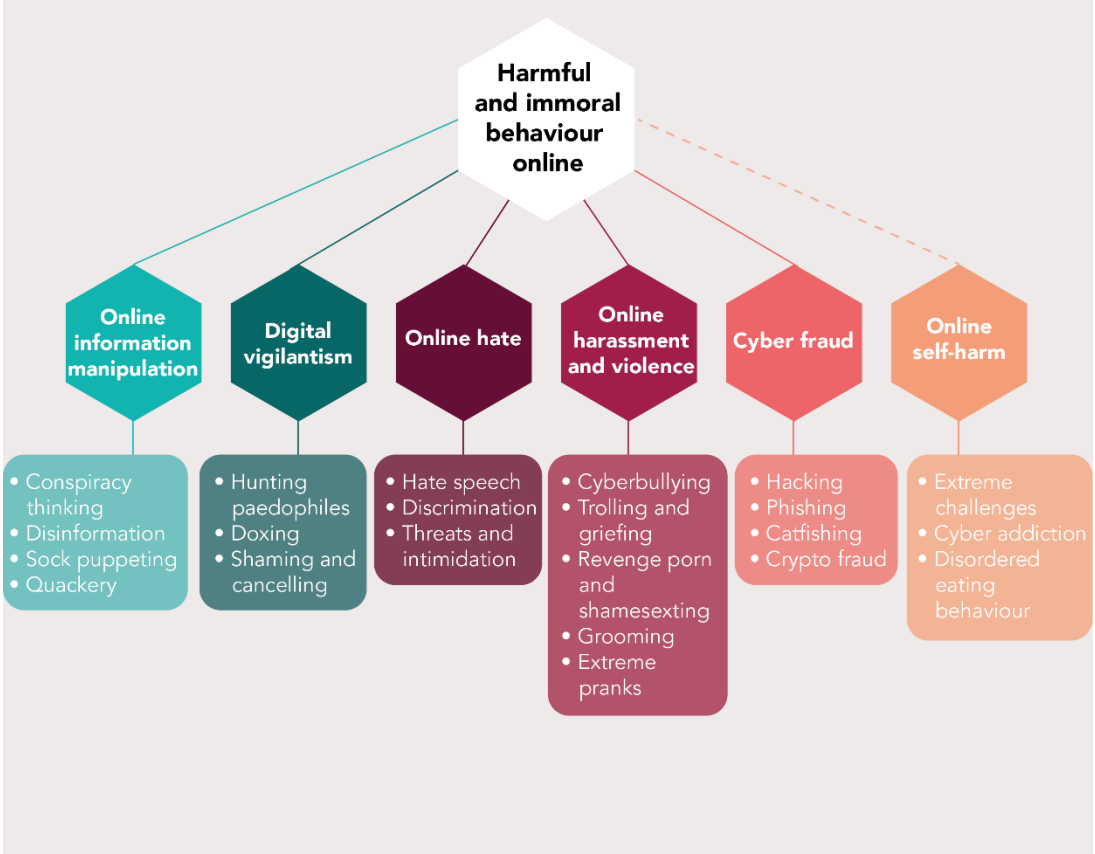
3.1 Taxonomy

The taxonomy presented in Figure 3.1 divides various forms of harmful behaviour online into six main categories:

1. Online information manipulation;
2. Digital vigilantism;
3. Online hate;
4. Online harassment and violence;
5. Cyber fraud;
6. Online self-harm.

Within these categories, we identify 22 phenomena (see glossary at the beginning of this report).

The taxonomy includes behaviours that are not always unlawful and are often displayed with impunity. Sometimes they are online versions of offline behaviour, for example threatening or intimidating others. Sometimes, however, they are new forms of internet-enabled behaviour, such as doxing, phishing, hacking or sock puppeting.



Bron: Rathenau Instituut

Figure 3.1 Taxonomy of harmful and immoral behaviour online

Below, we discuss the characteristics of harmful and immoral online behaviour by category and by phenomenon. We then turn to the scale of each phenomenon, where figures are available.

3.2 Online information manipulation

Online information manipulation is defined here as the dissemination of all kinds of information (text, images, audio) online that is presented as factual but is nevertheless false. It may be a deliberate act (for example spreading disinformation) or an unintentional one (for example spreading conspiracy theories). The intentional or unintentional dissemination of false information is misleading and may cause unrest and confusion. It also makes it more difficult to gauge the value of information (Rathenau Instituut, 2018a, 2021b). The category ‘information manipulation’ comprises **conspiracy thinking, disinformation, sock puppeting and quackery**. These phenomena are not always illegal in themselves, although they can be an element of criminal behaviour. Information manipulation does not

necessarily target specific individuals or groups, and it often affects society as a whole.

Conspiracy thinking

We use the term ‘conspiracy thinking’ when people believe that certain events or situations have been secretly manipulated behind the scenes by powerful forces with evil intentions (COMPACT Education Group, 2020). Conspiracy theorists generally believe that nothing is as it seems and that many events did not happen by chance but as part of an evil, premeditated plan. In doing so, they reject the possibility that reality is chaotic and complex. They have a tendency to blame ‘conspirators’ for what has gone wrong.

The term ‘conspiracy thinking’ is a sensitive one: it is value-laden and can be perceived as judgmental towards people who believe in conspiracies. Scholars point out that it is often used to delegitimise a person’s position on a given issue (Husting & Orr, 2007). Dutch sociologist Jaron Harambam argues in his PhD dissertation that there is an inherent risk in discrediting conspiracy theorists (Harambam, 2017, p. 75). Sometimes new evidence comes to light that proves conspiracy theorists right (Mortimer, 2017). That is why the Rathenau Instituut (2018a) is reluctant to refer to people as conspiracy theorists and uses this term with prudence.

Conspiracy theories spread faster **online**, but the phenomenon of conspiracy thinking is centuries old. As early as 68 CE, some Romans rejected the idea that the Emperor Nero had committed suicide and suspected a conspiracy behind his death (Champlin, 1998). What distinguishes online conspiracy thinking from its offline counterpart is that conspiracy theorists can easily connect and share information online, further reinforcing their beliefs. It is also easy for groups of conspiracy theorists to find other conspiracy theories online that strike them as plausible (Bessi et al., 2015).

People who disseminate conspiracy theories online are often suspicious of traditional media and science (Rathenau Instituut, 2018a, 2018b). They often feel that these institutions are elitist and do not represent them (Harambam, 2017). It is therefore important to them to communicate their convictions. There is a difference between (unintentionally) spreading conspiracy theories and (intentionally) sowing doubt about certain events, as the purpose of the first may not be malicious.

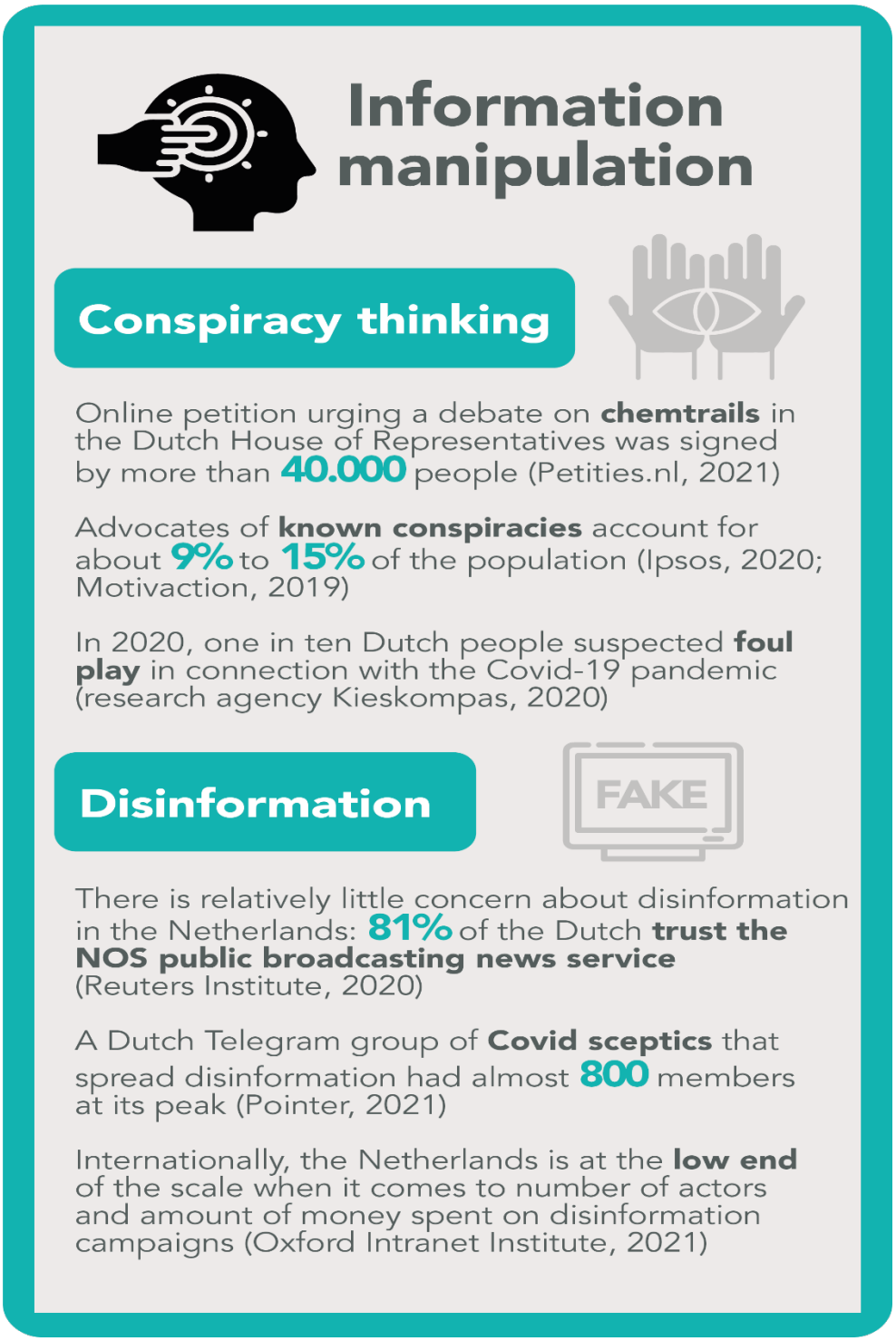
The **harm** caused by conspiracy thinking can be considerable: it can foster anti-democratic sentiments and undermine trust in broader societal structures (Sternisko et al., 2020). Conspiracy theorists can also harm individuals by accusing them of

secret evil intentions, by threatening them (Schildkamp & Rodenburg, 2021) or by committing criminal acts inspired by conspiracies. In the United States, kidnappings, pursuits and a murder have been linked to the QAnon conspiracy theory (*The Guardian*, 2020). The terrorist who opened fire on mosque-goers in Christchurch, New Zealand, was inspired by online conspiracy theories about white supremacy (*The Independent*, 2019), similar to the German terrorist who invoked racist conspiracy theories in 2020 (AP, February 2020). Conspiracy thinking can thus cause real damage, both to individuals and society. Conspiracy thinking in itself is not illegal, but certain manifestations or behaviours that derive from it may well be.

Scale

All manner of conspiracy theories have been circulating online in the Netherlands since the dawn of the internet, for example questioning the veracity of the Moon landing, the Holocaust and the 9/11 terrorist attacks in the USA. One well-known conspiracy theory is that the condensation trails we see in the sky are actually ‘chemtrails’, chemical agents sprayed by aircraft for nefarious purposes. Advocates of this theory have started an online petition that they hope will force the Dutch House of Representatives to debate the subject. More than 40,000 people signed the petition (Petities.nl, 2021). Alongside these older conspiracy theories, Covid-19 has also sparked new suspicion. In the Netherlands, such mistrust manifests itself not only online but also, and increasingly, in the offline world, for example in protest marches and setting fire to 5G transmission masts (NCTV, 2021).

It is proving difficult to ascertain from surveys just how large a following extremist conspiracy theories such as the QAnon movement actually have. Advocates of the QAnon conspiracy theory seek to overthrow the democratic state by force (Bellingcat, 2021). In the USA, the total number of QAnon followers is estimated to be between 3% and 14% of the population (Shanahan, 2021).



Bron: Rathenau Instituut

Figure 3.2 Online information manipulation

We have no estimates for the Netherlands, but there are some indicators of the scale of conspiracy thinking in general. According to the Dutch public broadcasting service NOS, some 40,000 QAnon-related posts in Dutch Facebook groups resulted in more than half a million interactions (i.e. likes, shares, comments) (Bouma, 2020). Viruswaarheid ('Virus Truth'), the action group that criticises and distrusts the government's Covid-19 policy, has about 10,000 Twitter followers out of a total of 2.9 million Dutch Twitter users. If their accounts are taken down, conspiracy theorists often move to other platforms such as Bitchute, Gab, Telegram or Parler.

Research by the Dutch newspaper *NRC Handelsblad* and the University of Amsterdam shows that despite the measures introduced by mainstream platforms Facebook, Instagram, Twitter and YouTube, the number of messages posted by popular conspiracy theorists and virus sceptics did not decline significantly between July 2020 and January 2021 in any country (Kist & Van den Bos, 2021; Motivaction, 2021). Conspiracy theories have particularly strong followings among young adults who have little faith in mainstream news channels (such as newspapers, radio and television news or current affairs programmes). Various studies focusing on those who believe in well-known conspiracies suggest that this group accounts for approximately 9% to 15% of the population (Ipsos, 2020; Motivaction, 2021). Among those who believe that the coronavirus is a biological weapon, there is an overrepresentation of young people, the lower educated and adherents of ultra-left and ultra-right political parties (Ipsos, 2020; Motivaction, 2021). This is consistent with findings on trust in science and other institutions such as the media (Rathenau Instituut, 2018b). According to the research agency Kieskompas, one in ten Dutch people suspect foul play in connection with the pandemic (Visser, 2020).

Disinformation

Disinformation differs from conspiracy thinking. Conspiracy theorists often genuinely believe the information they are disseminating, whereas those who intentionally spread disinformation do so to cause harm (COMPACT Education Group, 2020). It is easy to spread disinformation **online**. It is not always clear what quality control measures online platforms and websites apply, as the Rathenau Instituut concluded in a previous study on the digitisation of news (Rathenau Instituut, 2018a).

Disinformation often refers to the dissemination of information that is 'inaccurate' or 'misleading' (Rathenau Instituut, 2018a), rendering the term difficult to define. In the scholarly community, there have been growing calls in recent years to take a person's intention into account when appraising their behaviour (Gelfert, 2018). Something should only be classified as disinformation if the person disseminating it

does so with malice aforethought. But even this definition is difficult to work with in practice, as it is not always clear whether someone's intentions are malicious. The Rathenau Instituut (Rathenau Instituut, 2018a, pp. 33–34) makes the following distinction (based on a proposal by the Council of Europe):

1. **Disinformation:** information that is false and deliberately created to harm a person, social group, organisation or country.
2. **Misinformation:** information that is false but not created nor spread with the intention of causing harm, for example following an attack or other shocking event.
3. **Mal-information:** information that is based on reality, used to inflict harm on a person, social group, organisation or country.

Research on the circulation of fake news on Facebook shows that people over the age of 65 are more than seven times more likely to share fake news than young people. In other words; any measures meant to limit the damage of online disinformation should not only target young people (A. Guess et al., 2019).

In the Rathenau Instituut's view, disinformation's biggest threat is that it will undermine the public debate and the democratic process (Rathenau Instituut, 2020b). There may be various motives behind the dissemination of disinformation. For state actors, disinformation is a means of creating confusion and social unrest in their own or another country. But those who disseminate disinformation may also have more opportunistic or economic motives, for example if they can earn money by making fake news go viral. While it is not illegal to circulate disinformation, doing so may well be associated with hate speech, defamation, fraud or other criminal offences. Disinformation has been a priority issue for both social media platforms and politicians in recent years.

Scale

As yet, there has been little empirical research in the Netherlands on disinformation (Rathenau Instituut, 2018a; ROB, 2019). Little is known about the scale of the phenomenon, and there are only a few studies that investigate who is behind it and what impact it has (Prij & Janssens, 2020). A further problem is that sources are difficult to compare due to confusion (and disagreement) about the definition of disinformation. As a result, every study has a different way of measuring the scale of the phenomenon (Common & Kleis Nielsen, 2021).

For a general impression, we therefore need to turn to international studies. For example, there are peer-reviewed studies showing that fake news accounts for only 0.15% of daily media consumption in the USA (Allen et al., 2020). We know that three out of four Americans did not visit any fake news websites in the run-up to the 2016 presidential elections, while a quarter of Americans did do so at least once (Guess et al., 2020). Finally, it has been shown that the largest group of American visitors to fake news websites consists of people who use the internet intensively

but who are also very engaged and loyal users of mainstream news websites (Nelson & Taneja, 2018).

The Netherlands appears to have less disinformation circulating and a less polarised media landscape than the United States, at least for now (Rathenau Instituut, 2018a). According to the Digital News Report, the Dutch place a relatively large measure of trust in the media and are not much concerned about fake news (Reuters, 2020). We can account for this in part by considering the relatively robust position of the public broadcasting service, the quality benchmark for other news media (see Figure 3.2). Few Dutch people get their news solely from social media; they also turn to the television, the radio and newspapers. The traditional news media are well represented in the top twenty social media feeds of the Dutch (Möller et al., 2019).

Internationally, the Netherlands is also at the low end of the scale when it comes to 'cyber troop capacity', i.e. the number of actors, tools, permanent teams involved in and the amount of money spent on disinformation campaigns (University of Oxford, 2020, p. 18). There is almost no evidence of foreign disinformation, and only a few documented examples of Dutch operators using Russian disinformation tactics, for example (Rogers & Niederer, 2019). However, the incident in 2021 in the Dutch House of Representatives in which a comedian pretended to be Leonid Volkov (a close associate of Russian opposition leader Aleksei Navalvy) proves that foreign deception is not unthinkable here either. There was also a Dutch Telegram group that had almost eight hundred Covid-sceptics spreading disinformation (Pointer, 2021b).

Sock puppeting

A sock puppet is a sock used as a hand puppet. **Online**, it refers to a false online identity used for purposes of deception (Oleshchuk, 2020), often taking the form of fake accounts on online platforms. Sock puppets are unwelcome in many online communities and user terms and conditions may state that users may not pretend to be someone else. It is typically an online phenomenon because the internet lends itself particularly well to assuming a false identity.

Sock puppeting is sometimes associated with disinformation because it allows people to spread misleading information under another identity. In November 2020, for example, several media outlets reported that a white Republican candidate for the US Congress had pretended to be a black man and Trump supporter on Twitter (Espinoza, 2020). He had attempted to use a sock puppet to influence the public debate. In this particular instance, the ruse was exposed because the Republican candidate shared a message intended for his sock puppet account seemingly by

accident, thus exposing the ruse. He denied these accusations on Twitter. This example shows how easy it is for people to spread misleading information using sock puppets and how difficult it is to prove that they have done so.

Sock puppeting is a form of online information manipulation that can be deployed in many other forms of harmful behaviour, such as phishing (see 3.6, Cyber fraud) and trolling (see 3.5, Online harassment and violence). The **harm** done by sock puppeting depends on the actor's intentions. It can range from personal economic or reputational damage to societal damage. The latter occurs when the public debate is influenced by misleading information, as in the case of disinformation. Sock puppeting may be illegal if it is used to commit criminal offences such as slander, libel, defamation and hate speech. It may also be a criminal offence in its own right to steal someone's identity and use it as a sock puppet. Sock puppeting does not always entail identity fraud, however. It often involves fake profiles using computer-generated photographs, for example.

Scale

We have been unable to find any figures concerning the scale of the sock puppeting phenomenon in the Netherlands. There is virtually no literature on the subject. Given the lack of data, we refer here to what we know about the scale of phishing, catfishing and trolling (see below).

Quackery

Quackery is the unauthorised practising of medicine by someone who claims to be able to cure an illness with a useless remedy (Geerts & Den Boon, 1999).

Patients often go **online** to seek information about their illness and possible (alternative) treatment methods (Delgado-López & Corrales-García, 2018). Both the variety of information available and the possibility of connecting with fellow patients make the internet a valuable tool for them. But there quackery lurks: nothing is properly regulated, and it can be difficult to judge online information on its merits – much more so than offline. For example, research into online information about cancer treatments shows that the internet is commonly used to promote unproven and unsafe cancer therapies (Delgado-López & Corrales-García, 2018).

Nevertheless, as with the term 'conspiracy thinking', we would advise exercising restraint when referring to 'quackery' because it can very easily be associated with alternative medicine. Alternative medicine should only be regarded as quackery if it 1) rejects conventional medicine, 2) recommends harmful therapies without an appropriate warning, 3) costs a lot of money and 4) invokes supernatural forms of healing (Offit, 2013).

The **harm** caused by quackery to individuals can be considerable, for example if they undergo unreliable therapies based on such information or refrain from seeking care on the advice of quacks. For example, in 2020, the Covid-19 pandemic led to the dissemination of dangerous information about unauthorised, potentially very harmful drugs (Freckelton QC, 2020). In addition, searching obsessively for health-related information can lead to cyberchondria, which is anxiety induced by reviewing morbid or alarming content during health-related searches online (Aiken, 2016).

In 2019, a homeopathic doctor from Eindhoven was fined after advertising an unauthorised flu remedy online (*Eindhovens Dagblad*, 2019). Following administrative proceedings, the Dutch Health Care Inspectorate imposed the fine because the advertisements were misleading. Quackery can also be a criminal offence under Section 96 of the Dutch Individual Health Care Professions Act [*Wet op de Beroepen in de Individuele Gezondheidszorg*] if care practices damage a person's health.

Scale
We have been unable to find any figures concerning the scale of quackery in the Netherlands. The Rathenau Instituut (2018a, p. 40) has noted that of all the online hoaxes or fake warnings on the internet, almost a third were related to the 'health risks' of, for example, food, household items or insects. According to the World Health Organisation, false information about Covid-19 spreads quickly through social media and is hampering the fight against the virus (Laato et al., 2020). A recent study in the USA by the Center for Countering Digital Hate found that only 12 individuals (accounts) were responsible for spreading around 65% of anti-vaccination misinformation on social media (Bond, 2021). Some of these accounts are active on multiple platforms promoting natural health and selling supplements, workshops and books. Covid 19 has given these entrepreneurs a market opportunity, according to the Center's chief executive (Brumfiel, 2021).

Easy access to the internet, anonymity and low costs are making it increasingly attractive to search for medical information online (Zheng et al., 2020). The Netherlands ranks second in Europe when it comes to the percentage of the population searching for health information online (Eurostat, 2021). In 2020, 76% of the Dutch population did so, up from 53% in 2011 (Eurostat, 2021).

3.3 Digital vigilantism

Digital vigilantism is a form of collective action against persons who exhibit undesirable social behaviour. The motivation is **moral censure and taking the law into one's own hands**. Online, this may manifest itself in naming and shaming and

doxing, for example. While it is not in itself illegal to take matters into one's own hands, such behaviour is often associated with criminal acts (such as the use of violence).

Shaming is a tactic used to make socially accepted norms explicit, for example by confronting someone via the Internet with the racist nature of their statements. This makes it difficult to determine whether shaming is 'justified', and when moral boundaries are being crossed. Online vigilantism can assume the character of 'do-it-yourself policing' if people feel that the justice system is failing to call certain individuals or groups to account. It is a form of surveillance by those who feel it is their duty to point out to others what they consider to be morally reprehensible behaviour. In the online environment, such initiatives can emerge rapidly and spontaneously (see Chapter 4). The public availability of large amounts of personal information makes it possible to monitor and track others continuously.

We distinguish three forms of digital vigilantism: **paedophile hunting**, **doxing** and **shaming and cancelling**.

Paedophile hunting

Paedophile hunting is the 'hunting down' of persons suspected of being paedophiles. Paedophile hunters use **online** anonymity to pretend that they are children so as to 'trap' paedophiles (Hadjimatheou, 2019). In doing so, they sometimes also use physical violence.

Criminologist Katerina Hadjimatheou points out that the term 'paedophile hunting' is ethically problematic because it dehumanises people. The word 'hunting' suggests that it is permissible to hunt people in the same way as animals (Hadjimatheou, 2019). The term is popular with groups that use (online) violence against paedophiles precisely *because* it dehumanises and incites hatred while lending moral legitimacy to their own actions.

Paedophile hunters believe that their aims are consistent with those of the state. They believe that they are helping to uphold law and order. Unlike others who take the law into their own hands, paedophile hunters have faith in the justice system: they hand people over to the police, believing that the traditional institutions will deliver justice. In fact, many paedophile hunters do not think they are taking the law into their own hands, but see themselves as investigative journalists. Some groups are highly professionalised, train their members and accept donations (Hadjimatheou, 2019).

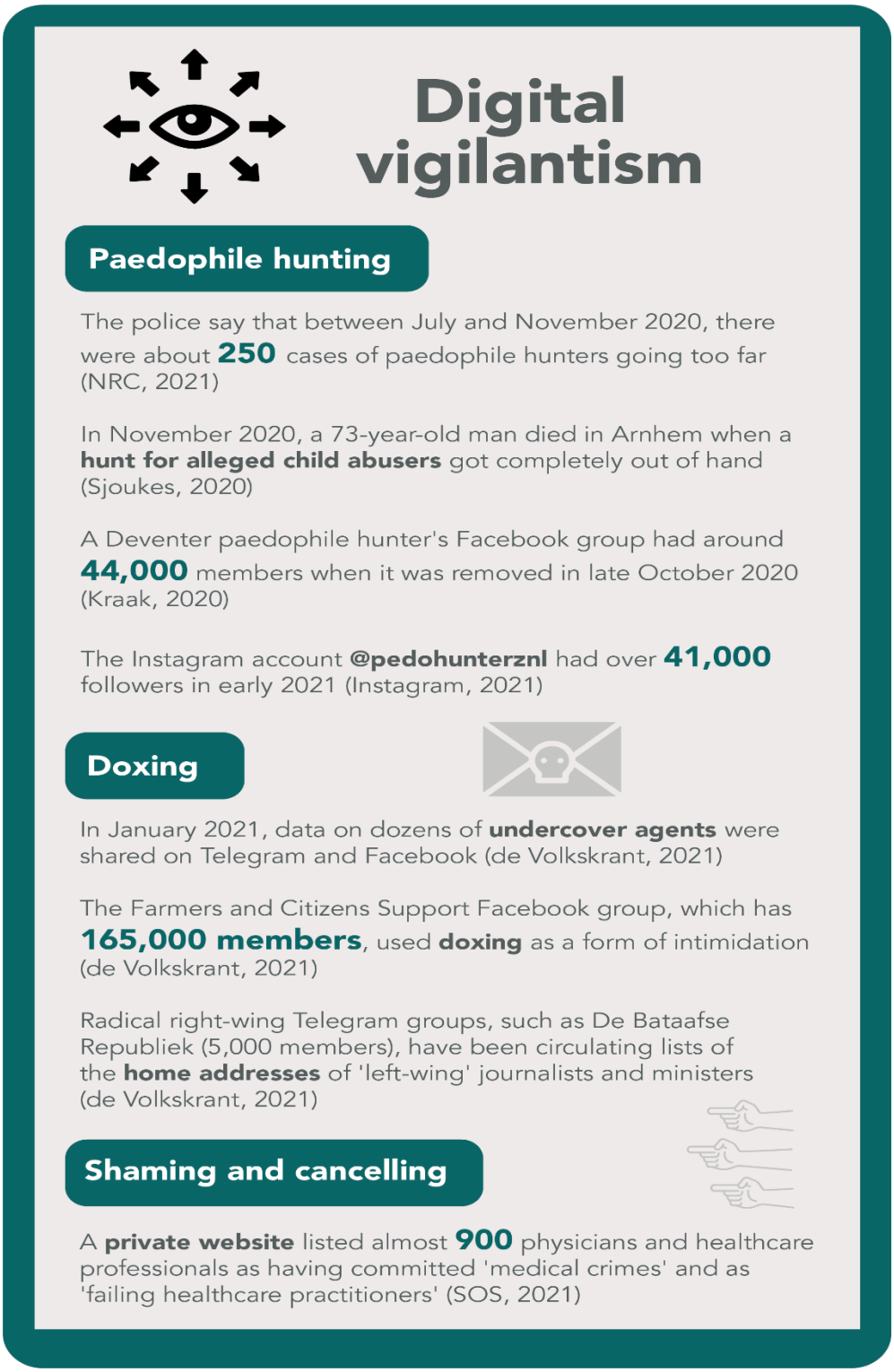
Paedophile hunting helps protect actual or potential victims, but it can also cause a great deal of **harm** – both to individuals and to society. Suspected paedophiles who are 'hunted' online can be 'condemned by the masses' before they have had a proper hearing. This can have major consequences for their standing in society, their personal relationships and their job – even if they are innocent. Moreover, they may become victims of physical violence. In addition, paedophile hunters' actions may create precedents when it comes to citizens taking the law into their own hands. They undermine the rule of law and cause more damage to society as a result. Paedophile hunting is not a crime in itself, unless it involves the use of violence, for example. In May 2021, the Arnhem District Court sentenced five suspects to up to five years in prison for brutally assaulting an alleged paedophile (*RTL Nieuws*, 2021).

Scale

'Do-it-yourself policing' is an international trend in which the Netherlands is at the forefront (Denef et al., 2017). The number of people who embark on their own investigations appears to be increasing steadily, in part due to the democratisation of information, research tools and knowledge (de Vries, 2018).

Paedophile hunters have been active in the Netherlands for at least a decade. There has been no empirical research into paedophile hunting in the Netherlands, but there are signs that the phenomenon is spreading (Herweijer & Ververs, 2020). According to Arnout de Vries of applied research organisation TNO, this is partly because adults have become more alert to online child abuse. During the pandemic, parents have been more likely to monitor their children's online behaviour, which means they are more aware of possible abuses. On the other hand, De Vries also thinks the increase is a response to the lack of law enforcement online (Herweijer & Ververs, 2020).

About 90% of people who get involved in such investigations have good intentions (de Vries, 2018). Still, sometimes things go wrong. In November 2020, for example, a 73-year-old man died in Arnhem when a hunt for alleged child abusers got completely out of hand (Sjoukes, 2020). The police say that between July and November 2020, there were about 250 cases of paedophile hunters going too far (Veldhuis & Ingabire, 2021).



Bron: Rathenau Instituut

Figure 3.3 Digital vigilantism

One of the accounts with most reach on Instagram for sharing information and videos of confrontations is @pedohunterznl, which had over 41,000 followers in early 2021 (Instagram, n.d.). A Deventer paedophile hunter's Facebook group had around 44,000 members when it was removed in late October 2020 (Kraak, 2020). The rise in the number of paedophile hunters has coincided with an annual increase in the number of child abuse images circulating online, also recently discovered on Pornhub, the largest mainstream pornography platform (Grant, 2020a, 2020b).

Doxing

Doxing is the public release of an individual's private, sensitive, personal information, for example their home address, telephone number, passport number or employer contact information, family members' contact details, or photographs of their children (MacAllister, 2016). The term 'doxing' probably comes from the online hacker group Anonymous and refers to 'docs' or 'documents'.

Like many other forms of immoral and harmful behaviour, doxing is often not a stand-alone activity but a strategic component of other forms of online harassment, such as threats and revenge porn. Because doxing makes their personal details public, victims are also exposed to the risk of harmful behaviour offline.

The information used to dox someone is often publicly available online, without hacking being required. Suppose, for example, that someone lists his employer on the professional social network LinkedIn. This information can be combined with other public sources to serve as a weapon in a hate campaign.

People's reasons for doxing are not straightforward. It may be purely for fun, but it may also serve political purposes or be a means of 'self-regulation' in certain communities. An example of the latter is the hate campaign #gamergate, which targeted women in the gaming community. The campaign was driven by male members of the community who were opposed to more diversity among gamers. Doxing was widely used to threaten and intimidate women, for sexist reasons (Wingfield, 2014).

Doxing is **harmful** because victims' private information can easily be exploited online, for example to threaten someone or harass someone's employer. People who engage in public debate may also be afraid of doxing, sometimes causing them to censure themselves online.

Doxing is a legally complex matter because many countries, including the Netherlands, have no legal grounds for prosecuting this behaviour (MacAllister, 2016). Dutch Minister of Justice and Security, Ferdinand Grapperhaus, has

announced that he wants to introduce legislation prior to the House of Representatives' summer recess of 2021 to tackle the phenomenon. The Public Prosecution Service is also investigating whether specific cases of doxing can be classified as criminal (ScienceGuide editorial board, 2021).

Scale

Very few empirical studies of doxing have been published in the international academic literature, let alone in the Netherlands. The phenomenon has been reported since the early 2000s, however, and has become much more visible during the Covid-19 pandemic.

For example, the Netherlands' National Coordinator for Counterterrorism and Security (NCTV) reported in its quarterly publication *Terrorist Threat Assessment Netherlands* that anger and frustration about the coronavirus restrictions have led to doxing as an intimidation tactic by activists who put the personal data of police officials and politicians online (Von Piekartz, 2020). In January 2021, data on dozens of undercover agents were doxed, and after riots erupted in response to the government's decision to install an evening curfew, the online hunt for these agents intensified in Telegram and Facebook groups, as reported by the Dutch newspaper *de Volkskrant* (Von Piekartz & Bahara, 2021). The Farmers and Citizens Support Facebook group, which has 165,000 members, used doxing as a form of intimidation by threatening to publish the personal details of dozens of undercover agents (Von Piekartz & Bahara, 2021). Radical right-wing Telegram groups, such as De Bataafse Republiek (5,000 members), have been circulating a list of the home addresses of 'left-wing' journalists and ministers for some time now, sometimes accompanied by a call to form a 'vigilante group' to 'deactivate' these persons 'by non-violent means' (Von Piekartz & Bahara, 2021). And in March 2020, certain active Twitter users found the front doors of their homes plastered with stickers from Vizier Op Links, an anonymous platform with about 16,000 followers that tries to disrupt the daily lives of left-wing thought leaders, activists and politicians.

Examples from the USA also feature in the press, such as the Twitter account @YesYoureRacist, which tries to track down the identity of racists. When this account was used in August 2017 to appeal for help in tracing the protesters in Charlottesville, the number of followers grew from 65,000 to almost 400,000 within a few days (van Houwelingen, 2017). This illustrates just how quickly the phenomenon can spread.

Shaming and cancelling

Online shaming is a practice of public moral criticism in response to violations of social norms (Billingham & Parr, 2020). People who engage in online shaming are not always out to embarrass someone. More often, they want to draw attention to and challenge a social practice or norm and mobilise others to their cause.

Public shaming can help to uphold and validate existing social norms, even in the absence of laws and regulations (Billingham & Parr, 2020). For example: it can be an effective way to alert people to abuse or to 'expose' racism and sexism (Billingham & Parr, 2019). On the other hand, this type of behaviour can also be used to deprive women of their voice in the public debate by denigrating them publicly and by humiliating young girls for their sexuality (Levey, 2018). Shame-sexting (creating and distributing sexually explicit images or videos without consent) is described in section 3.5 and is not discussed here.

The internet makes shaming easier because it is easy for groups of people to come together and for communities to self-regulate **online**. Suppose that someone in a Facebook community posts something that is perceived as racist. One member's condemnation may spur all the members of that community to turn against that person. The internet also makes shaming very difficult to monitor and control.

In a recent article in the *European Journal of Philosophy*, the authors propose five moral constraints on public shaming (Billingham & Parr, 2020). If all of these constraints are met, shaming *may* be morally justifiable. That does not mean that everyone has the 'right' to shame, but it does offer a basis for moral justification. Shaming can sometimes serve a 'noble purpose', in other words, but the means to that end is itself difficult to justify.

Table 1 Five moral constraints on online public shaming

Moral constraint	Explanation
Proportionality	Public shaming is proportionate when its negative consequences are not excessive in comparison with its positive consequences.
Necessity	Public shaming is justifiable if there is no other course of action that serves the same purpose while imposing future burdens.
Respect for privacy	Public shaming must respect rights to privacy. It should not involve dredging up irrelevant information from the past or highly sensitive information that would preferably remain private.
Non-abusiveness	Public shaming must not involve threats, sexist or racist abuse, or other forms of hate speech.
Reintegration	Public shaming must not exclude individuals from reintegration back into the community, and shamers must be aware of the risks of exclusion.

Source: Billingham & Parr (2020)

Billingham & Parr's study (see Table 1) shows that the moral justification for public shaming is a complex matter. It is important to consider the intentions of people who shame online. Do they give others the opportunity to learn from and modify their behaviour? Is it possible for victims to return to the community after being shamed?

Because it is much harder to control the consequences of an action online (for example, because its reach cannot be monitored), it is also much harder for all five constraints to be met in the online environment. Bystanders play a critical role in public shaming; there is no traditional victim-perpetrator relationship, but rather collective action. Moreover it is difficult to hold people accountable online for the negative consequences of shaming. It is, after all, a form of collective action with shared responsibility.

A well-known example of online public shaming is the #Metoo movement. It allowed women who had experienced sexual harassment and whose complaints had been ignored by existing institutions to speak out (Mendes et al., 2018). Using the internet, they sought to obtain some form of justice (Powell, 2015). The movement has made organisations more aware of sexually transgressive behaviour and more responsive to allegations by victims (Leopold et al., 2021). On the other hand, some hashtag users have made false allegations, damaging the people they have accused.

Online public shaming can cause significant harm to victims. A public accusation may result in their expulsion from a particular community or prevents them from doing their job. Public shaming in itself is not a criminal offence, but in the Netherlands libel and slander (malicious falsehood) can be qualified as offences. Libel is defined here as the dissemination of negative statements about a person with the aim that others can hear those statements. Libel becomes slander if the perpetrator knows that the statement or assertion in question is false.

Cancelling is a type of online public shaming that calls for some form of social castigation in which an individual is excluded from a community. Cancelling goes a step further than shaming because it is explicitly aimed at undermining someone's authority. The boundary is fluid. The #MeToo movement, for example, led to the resignation or withdrawal from public life of quite a number of public figures.

A well-known example of public shaming and cancelling in the Netherlands concerns the D66 party's MP Sidney Smeets, who retired from the House of Representatives following allegations of sexually transgressive behaviour towards minors. Smeets claimed that he had never broken the law, whereas his 'shamers' invoked certain ethical norms that are not necessarily enshrined in criminal law. Their intention was to address the issue of ethical norm violations by people in positions of power.

In the USA, Apple employees started an online petition calling for the resignation of a new employee (Ghaffary, 2021) who had been accused of making sexist comments in the past. The employee resigned after being put under pressure internally. Because statements made online live on for years afterwards, the internet makes it easier to call someone to account for past statements.

Scale

One of the best-known examples of online public shaming is the #Metoo movement, which began on Twitter on 24 October 2017. The hashtag was used 12 million times in the first 24 hours (CBS, 2017). We have not been able to find data on the scale of this movement in the Netherlands. We will not discuss the scale of shame-sexting and other forms of sexual public shaming here because we classify these under 'online harassment and violence', and not under 'digital vigilantism'.

In the Netherlands, Stichting Online Shaming (SOS) has taken up the fight against online public shaming. In a case brought to trial by SOS, for example, the courts in January 2021 banned the website ZwarteLijstArtsen.nl, which has acted as a public 'pillory' for medical professionals for a decade (NOS, 2021a). A private website, it listed almost 900 physicians and healthcare professionals, often with photographs, and portrayed them as committing 'medical crimes' and as 'failing healthcare practitioners' (SOS, 2021).

According to national newspaper *NRC Handelsblad*, the Covid-19 pandemic appears to be increasing the intensity and scale of online public shaming. The newspaper does not cite figures, but points to public shaming as a common means of inducing others to comply with the coronavirus restrictions, for example by circulating photographs of offenders online (Van Noort, 2020).

Similarly, SOS does not have figures on the total scale of shaming and cancelling in the Netherlands. The term 'public shaming' is very broad and victims often do not realise that this is what they are going through. Depending on the type of public shaming, they may contact various other helplines, such as the Dutch children's hotline [*Kindertelefoon*], the Dutch social welfare service for crime, abuse, accident or disaster victims [*Slachtofferhulp Nederland*], EOKM Expertise Centre for Online Child Abuse, or Helpwanted.nl for victims of online sexual abuse.

EOKM and Helpwanted.nl received 'a few' reports of Muslim girls who are criticised online for not wearing a headscarf or for wearing a low-cut top in public. EOKM thinks that the number of reports is just the tip of the iceberg, because people confronted by public shaming are afraid of 'victim blaming'. They themselves are often ashamed of the images posted and do not dare to ask for help.

3.4 Online hate

Online hate consists of **hate speech**, **discrimination**, **threats and intimidation**. Online hatred is directed at individuals, but is often also meant to harm underlying groups. A form of **xenophobia** (aversion to all things foreign) is the main driver behind online hate. Online hate stems from aversion to certain groups of people, even if it is directed at individuals. Victims of online hate often suffer the consequences of various forms of immoral and harmful behaviour, for example racism, stalking and doxing.

Victims of online hate are condemned on the basis of various aspects of their identity. Someone who is, for example, black, female and lesbian may experience online hate that is racist, sexist and homophobic. In sociology, gender studies and law, when vectors of inequality intersect, this is known as 'intersectionality'. Intersectionality plays a major role in online hate. Men may face threats and harassment, but they are far less likely than women to experience sexual harassment (Vogels, 2021).

Hate speech

There is no uniform definition of hate speech. The Council of Europe defines it as follows: ‘hate speech covers all forms of expressions that spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance’ (Council of Europe, 2021).[2] Online hate speech is **harmful** both to the victims and to society at large, as it creates an unsafe environment for all. People may become more cautious in their online comments for fear of falling victim to online hate.

Online hate speech differs from offline hate speech, according to research by UNESCO (2015). First, **online** hate content can stay available for a very long time in different formats across multiple platforms. Second, hate speech is itinerant; even if it is removed, it can quickly reappear on another platform elsewhere. Third, online anonymity makes enforcement difficult; although the police can often identify perpetrators, they may not have enough capacity to follow through. Finally, the transnational reach of the internet makes it difficult to tackle hate speech with national legislation when the hosting platform is headquartered in another country (Gagliardone et al., 2015).

The Dutch Criminal Code contains three sections on hate speech: Sections 137c, 137d and 137e. ‘Acts’ such as threats do not qualify as hate speech in the Netherlands because they are dealt with in a separate section of the Criminal Code (see Section 3.4). In June 2020, the political parties GroenLinks and ChristenUnie submitted a bill raising the penalties for hate crimes, i.e. crimes committed with discriminatory intent (Bhikhie, 2020).

Scale

According to the Council of Europe’s European Commission against Racism and Intolerance, reported cases of online hate speech are only the ‘tip of the iceberg’ (ECRI, 2019). About one tenth of all tweets directed at female politicians in the Netherlands are said to display hatred or aggression. This observation emerged from research conducted by the Utrecht Data School and the weekly magazine *Groene Amsterdammer* (Saris & Van de Ven, 2021), which investigated 339,932 tweets directed at all women on Dutch electoral lists between 1 October 2020 and 26 February 2021. The conclusion was that MPs who are female and a member of a minority religion or of colour are subject to a great deal of hate speech (Saris & Van de Ven, 2021).

Female journalists and scientists are also targets of hate speech. Female columnists, for example, are more likely to encounter hate speech than their male

2 https://www.coe.int/en/web/freedom-expression/hate-speech

counterparts, according to a 2017 survey by national newspaper *de Volkskrant*. Of thirty female columnists working for five newspapers and two opinion magazines, two thirds had been threatened online once or multiple times. Half of the columnists that were surveyed, reported that they occasionally or frequently felt intimidated by online comments (Linnemann & Melchior, 2017). Scientists are also increasingly troubled by online hate speech or 'vitriol', according to the president of the Royal Netherlands Academy of Arts and Sciences, Ineke Sluiter. She cites well-known scientists who have been targeted, such as Marion Koopmans, whose Twitter feed is awash with condemnations and threats after every media appearance (Digan, 2021).

Another well-known (and somewhat earlier) example of hate speech against female politicians in the Netherlands is the 'Get Lost Day' organised as a Facebook event on 6 December 2016 against the Dutch politician Sylvana Simons. Some 39,000 people marked themselves as interested in the event, and at least 16,000 planned to attend (Wiegman, 2016). Hate speech against Simons continued during the March 2021 elections. Traditional media sometimes turn off commenting on reports about politicians (RTL, 2021). Almost a third of all tweets directed at politician and climate activist Kauthar Bouchallikht contained sexist or Islamophobic hate speech. On one particularly bad day, she received a hate message every three minutes (Saris & Van de Ven, 2021).

Discrimination

Below, we discuss three common forms of online discrimination: sexism, racism and homophobia. Discrimination is a criminal offence under Sections 137c, 137d and 137e of the Dutch Criminal Code.

Sexism
Sexism refers to actions or attitudes that discriminate against people based solely on their gender (European Institute for Gender Equality, 2016). **Online** sexism is often associated with other forms of immoral and harmful behaviour online, such as threatening and cyberbullying. The International Center For Research On Women (ICRW) uses the term 'technology- facilitated gender-based violence' for all forms of cyberbullying, online harassment and verbal or other violence based on someone's sexual or gender identity (Hinson et al., 2018). Women are often the victims.

According to Amnesty International (2018), online sexism and misogyny are often intended to intimidate or belittle women. Approximately 7.1% of tweets received by women are either 'problematic' or 'abusive' (Amnesty International, 2017). Passive and indirect sexist comments presented as jokes can also be harmful to women's well-being, according to a 2015 Harvard study (Fox et al., 2015). Online sexism is

harmful because it creates an unsafe situation for direct victims and contributes to an environment that is less safe and less free for women in general (Plan International, 2020).

Sexism is not explicitly addressed in the Dutch Criminal Code, unlike in Belgium and France for example. It is a criminal offence as a form of discrimination under Sections 137c, 137d and 137e of the Criminal Code.

Racism

Online racism, also known as cyber racism, differs from offline racism in that the racist statements are made **online**, and that racists can easily engage with one another and unite into groups online (Bliuc et al., 2018). People with racist views use the internet to validate their views and to gain a sense of belonging with like-minded others. The internet empowers them. People who exhibit racist behaviour can easily share their views there, partly because the internet offers them anonymity. Common motives behind online racism are to hurt people of colour, to advocate racial conflict and to normalise racist ideology in the public debate (Bliuc et al., 2018).

The **harm** suffered by victims of online racism has been documented in detail (Bliuc et al., 2018), but specific Dutch research on this topic appears to be scarce. A study of racism in online gaming communities, for example, shows how men of colour attempt to cope with online racism by remaining silent and not speaking out for fear of attracting more hate (Ortiz, 2019). The role of bystanders in rendering social norms explicit also plays a major role in online shaming, as we have shown in section 3.3.

Homophobia

Homophobia is the fear of, or hatred towards, people who are emotionally or sexually attracted to people of the same gender. By extension, however, the term is also used to describe fear or hatred of people with non-cisgender[3] identities, for example. Discrimination is deemed to occur when someone's actions are motivated by homophobia.

One form of harmful and immoral homophobic online behaviour is 'outing'. Outing is when a person discloses another person's gender identity or sexual orientation without the latter's consent. This can be extremely **harmful**, for example in countries where LGBTQ+ people are persecuted or in families where they are not accepted. In August of 2020, Flemish public broadcaster VRT devoted a lengthy article to online chat groups whose members wanted to 'track down' people from the LGBTQ+ community (VRT, 2020). In some instances, members even offered to

3 Cisgender means 'non-transgender': people born male who identify as male or born female who identify as female. Their gender identity thus matches the sex they were assigned at birth.

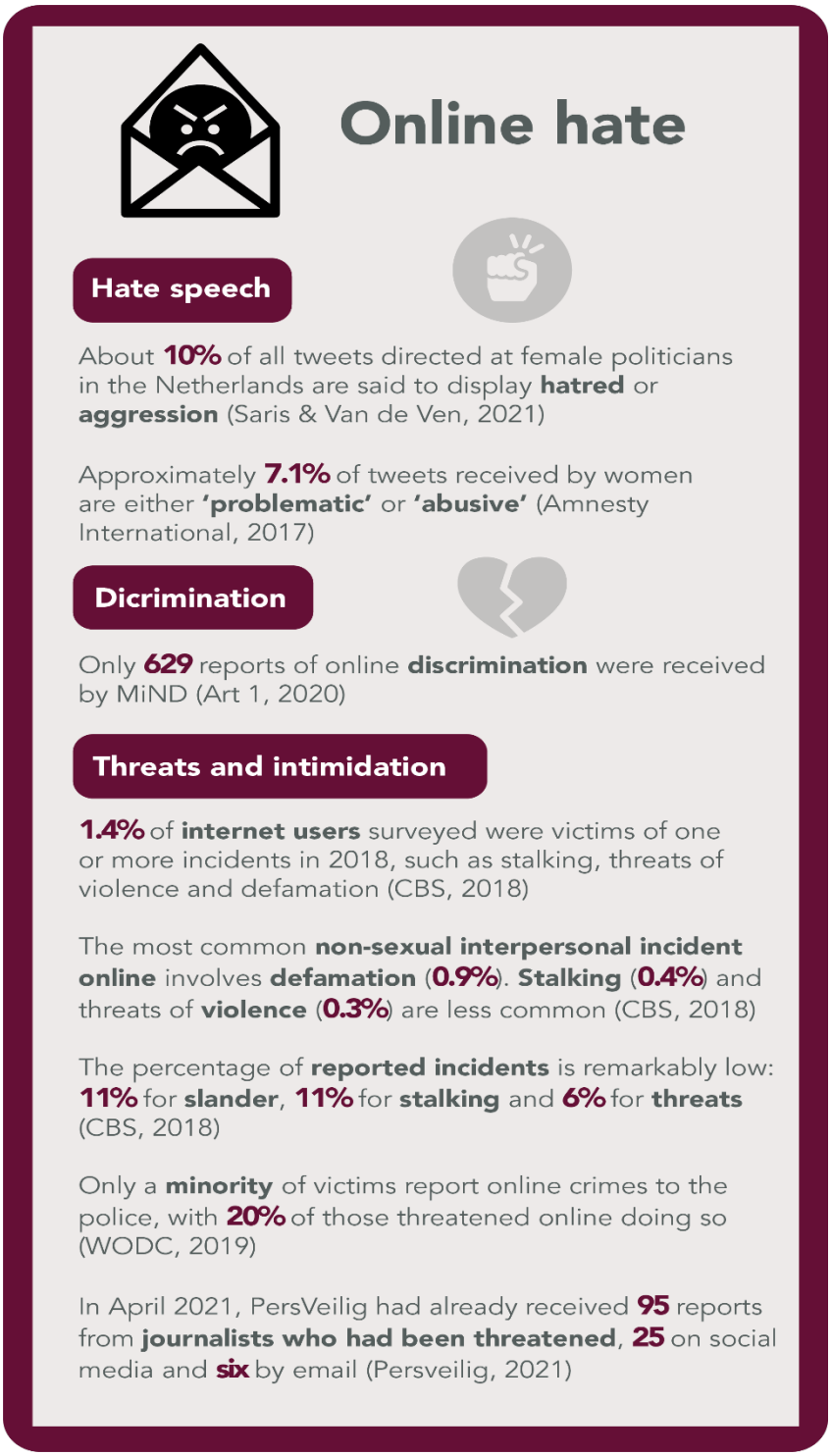
pay others to physically attack certain people. This sort of behaviour makes both the online and the offline world unsafe for victims.

Scale

Figures on the overall scale of discrimination in the Netherlands reveal only a fraction of the total. Many incidents are not recognised as discrimination or are not reported to the authorities (MiND, 2020).

According to the Dutch internet discrimination helpline, MiND, the police registered a total of 5,487 incidents of discrimination (both online and offline) in the Netherlands in 2019 (MiND, 2020), whereas in 2016 only 4,376 incidents were reported (Mink & Van Bon, 2017). The number of incidents is therefore growing. The municipal antidiscrimination services saw a slight decline in reports of discrimination, from 4,761 in 2016 to 4,382 in 2019 (Mink & Van Bon, 2017). Of all these reports, only a fraction (slightly more than one tenth) consisted of online discrimination (see Figure 3.4).

In 2019, discrimination based on ethnic origin was the most common form (2,156 incidents registered by the police and 1,922 by the antidiscrimination services), followed by sexual orientation (1,603, police), anti-Semitism (768, police), disability (552, antidiscrimination services) and gender (515, antidiscrimination services) (MiND, 2020, p. 3).



Bron: Rathenau Instituut

Figure 3.4 Online hate

Threats and intimidation

Because groups can easily mobilise **online**, certain behaviours can quickly escalate. As a result, a group of people may target one individual, who may find the group's behaviour intimidating (Blackwell et al., 2017). The harassment can then take different forms, ranging from repeated posting or messaging and calling someone's employer to threatening to disclose certain photos or information. It can be intimidating simply to be followed on social media by certain accounts.

Young women in particular face threats of sexual assault online (Plan International, 2020). Our interviews reveal that victims of group hate online often encounter it in many different forms. Under Section 285 of the Dutch Criminal Code, threatening someone with certain serious offences is a crime against that person's freedom. Whether or not the victim actually felt threatened is irrelevant. Young women in particular face threats of sexual assault online (Plan International, 2020). Our interviews reveal that victims of online group hate often encounter it in many different forms.

Scale
There are few figures available indicating the overall scale of online threats and intimidation in the Netherlands. We do know who the main victims are, however. Among adolescents aged 12 to 18, 5% have been subjected to non-sexual forms of threat or intimidation. Girls in this age group are targeted more often than boys (7% as opposed to 4%) (CBS, 2018).
Approximately 1.4% of internet users were victims of one or more incidents in 2018, such as stalking, threats of violence and defamation (CBS, 2018). The most common non-sexual manifestation is defamation (0.9%). This includes gossip, distributing photographs or videos, and bullying. Stalking and threats of violence are less common at 0.4% and 0.3% respectively (CBS, 2018).

The percentage of reported incidents is remarkably low: 11% for defamation, 11% for stalking and 6% for threats. The most common reason for not reporting stalking, defamation and threats is the assumption that it will not help, with victims often commenting that it is not a police matter (CBS, 2019c). Research by the Research and Documentation Centre (WODC) also shows that only a minority of victims report online offences to the police, with 20.2% of those threatened online doing so (Sipma & Leijsen, 2019).

Research by Amnesty International (see infographic 3.4) shows that in 2017, around 23% of female survey respondents in eight countries had experienced online taunts or threats at least once, ranging from 16% in Italy to 33% in the USA (Amnesty, 2017). The Netherlands was not included in this study.

Finally, it is striking that the number of online and offline threats against journalists is increasing (SVDJ, 2021). In the Netherlands, a new helpline, PersVeilig, recorded a total of 121 incidents in 2020, 36 of which occurred on social media and nine by e-mail. On 19 April 2021, PersVeilig had already received a total of 95 incident reports, 25 occurring on social media and six by e-mail.[4]

3.5 Online harassment and violence

The category 'online harassment and violence' includes the phenomena of **cyberbullying, trolling and griefing, shame-sexting, sextortion and revenge porn, grooming,** and **extreme pranks**. Their common denominator is that they all involve intentionally hurting and harming individuals, with **sadism** or a deliberate wish to hurt being the main motive. Online harassment and violence includes harassing or sexually harassing others online without this being motivated by vigilantism or xenophobia. Such behaviour can be used in combination with other phenomena (such as sock puppeting or hacking). As with online hate, online harassment and violence are common but incidents are rarely reported to the authorities.

Cyberbullying

Online bullying is often referred to as cyberbullying. Cyberbullying occurs when a group or individual engages in repeated and intentionally harmful behaviour online against a victim who has difficulty defending themself (Juvonen & Gross, 2008). This makes cyberbullying an all-purpose term for many forms of harmful behaviour online. Stalking, doxing and sock puppeting are sometimes also considered cyberbullying. This is a logical overlap, since bullying can take many different forms. Systematically and deliberately excluding someone from online social media communities is also considered cyberbullying.

Much of the research into cyberbullying focuses on children and adolescents. Of all the forms of harmful and immoral behaviour online, cyberbullying is the one that researchers have studied the longest. When adults engage in repeated, intentionally harmful behaviour online, we often do not refer to it as bullying but as harassment or hate speech, for example. Socio-psychological factors, such as loneliness and insecurity among adolescents, play a role in both online and offline bullying. What makes bullying **online** different from bullying offline is that it is harder

4 Source: e-mail from PersVeilig representative received on 19 April 2021

for victims to escape it. Because the act of bullying is no longer tied to a physical location, victims may also feel unsafe in their own home.

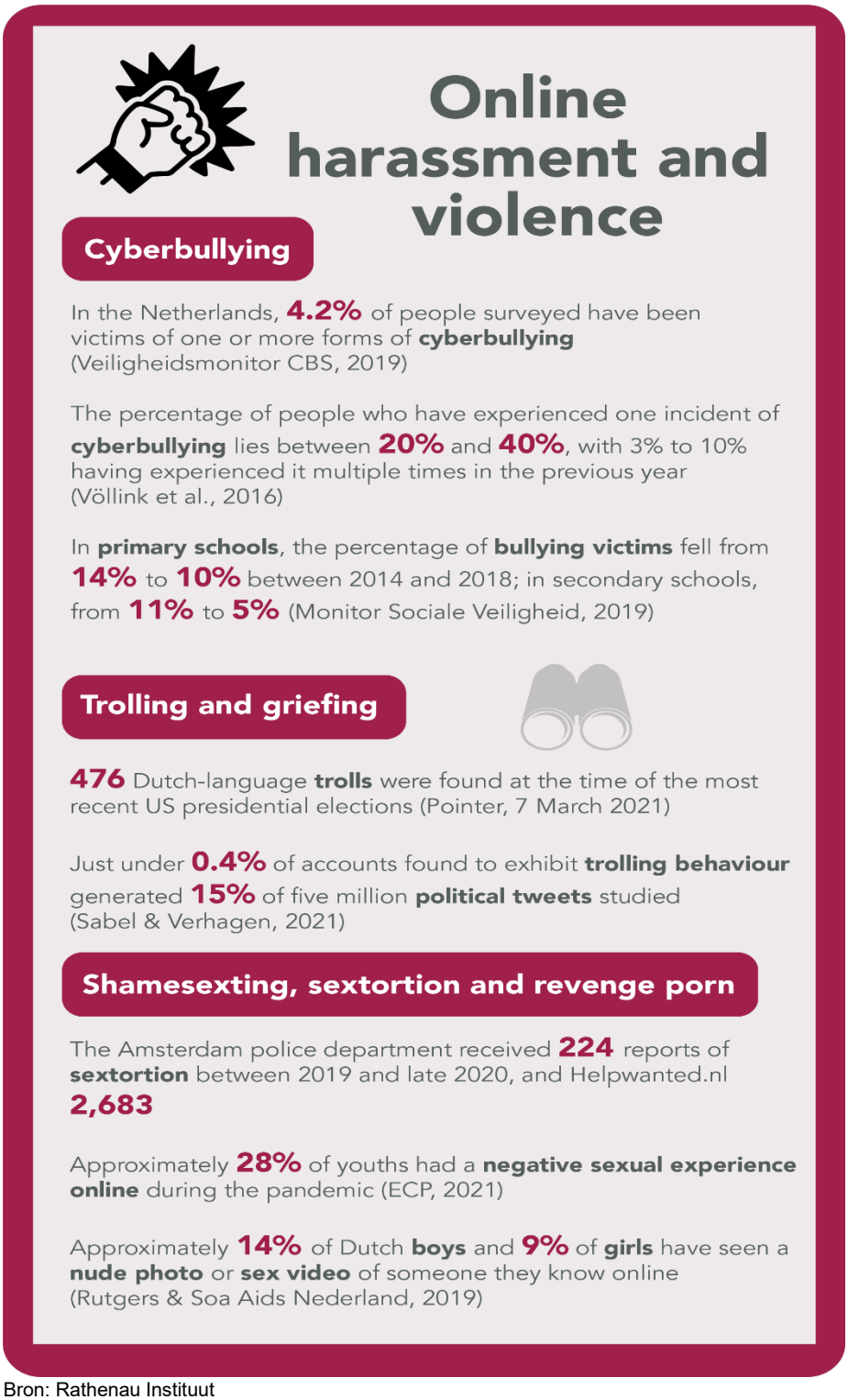
Bystanders often do not intervene because it is harder to interpret online posts or messages. Is this really cyberbullying or does the victim think it's funny too? The asynchrony of the internet also plays a role in bystanders not intervening, as it is not always clear whether a situation has already been 'resolved'. Unlike offline bullying, cyberbullying does not necessarily occur when others are online to witness it; posts and messages may be online for a long time before bystanders notice them (Cleemput et al., 2014).

Cyberbullying can be enormously **harmful** for victims. For example, a study dating from 2017 shows that adolescents who are victims of cyberbullying are significantly more likely to have suicidal thoughts and to attempt suicide than adolescents who are not bullied (Nikolaou, 2017). Cyberbullying also has a negative impact on adolescents' mental health and can cause depression. Like offline bullying, cyberbullying is not a criminal offence (Shaikh et al., 2020).

Scale
Researchers have been studying cyberbullying since the turn of the present century. Many studies show that cyberbullying is a common problem. Since the incidence of cyberbullying is not always measured in the same way, recorded rates vary widely. For example, the percentage of adolescents who said they had been victims of cyberbullying in 2012 ranged from 4% to 57% (Dehue et al., 2012). The percentage of people who have experienced one incident of cyberbullying lies between 20% and 40%, with 3% to 10% saying that they had experienced it multiple times in the previous year (Völlink et al., 2016).

Statistics Netherlands' Safety Monitor [*Veiligheidsmonitor*], a biennial national survey of crime victimisation, shows that in 2019, 4.2% of Dutch people were victims of one or more forms of cyberbullying, up from 3.1% in 2017 and 2012 (CBS, 2019c). The same agency's statistics reveal that only 22.2% of these victims reported cyberbullying to the police (CBS, 2021).

The Social Safety Monitor [*Monitor Sociale Veiligheid*] that Dutch schools conduct every year suggests that bullying (both online and offline) was on the decline in primary and secondary schools between 2014 and 2018 (see Figure 3.5). Among primary school pupils, almost 8% were reported to have been bullied by email or in a chat app or text message. Cyberbullying is more common in secondary education than in primary education (*NJi*, 2019).



Bron: Rathenau Instituut

Figure 3.5 Online harassment and violence

Trolling and griefing

The narrowest definition of trolling describes it as intentionally disrupting online communities by behaving in a way that is deemed unacceptable, such as calling people names, picking fights or making negative comments about others (Cheng et al., 2017).The term trolling has been used since the early twenty-first century to describe forms of anti-social behaviour in online communities, for example on Wikipedia or in online forums. The media often describe trolling as behaviour exhibited by people with sadistic and sociopathic personalities. Recent research, however, shows that ordinary people can engage in trolling and that a person's mood, the context of an online discussion and other factors have a significant bearing on people's 'trolling behaviour'.

People often troll without a thought for how their behaviour will affect their victims, and usually on the pretext of humour. Victims who defend themselves against trolling can expect to be trolled even more; for many people, trolling is a kind of game to see how far they can go (Phillips, 2015). As early as 2002, research showed that feminist and other non-mainstream online communities are more likely to be victims of trolling, which means that certain social groups will suffer more **harm** from trolling than others (Herring et al., 2002).

There is also a broader interpretation of trolling as a concept that includes certain forms of information manipulation. One example would be the use of fake accounts to spread disinformation and influence the public debate. This stretches the definition of trolling beyond its normal limits because it goes beyond intentionally disrupting a (specific) community. For example, people who set up fake accounts are more likely to want to manipulate information and the public debate. Trolling is not a criminal offence, unless it is associated with threats, slander, libel, defamation or other offences.

Griefing is a form of trolling that occurs mainly in the gaming community and that involves the deliberate harassment of other players. Griefing has elements of cyberbullying but often does not target a specific person. It is a player's 'avatar' that is the target, and not the person behind the avatar. Unlike in the case of cyberbullying, players can more easily 'escape' griefing by ceasing to play the game or by blocking the griefer, for example (Coyne et al., 2009).

Scale

Online trolling has been a subject of study since the late 1990s. There have also been empirical studies on the nature and scale of certain forms of trolling in the Netherlands. For example, it appears that there is no coordinated troll network in the country, although smaller networks do exist.

An investigation by national newspaper *NRC Handelsblad* in 2017 showed that various internet trolls disseminating disinformation (see also above) had been working for the Dutch political party Denk. As a group, they posted or shared 1,636 messages in support of Denk's campaign and gave it 2,171 likes (Kouwenhoven & Logtenberg, 2017). In February 2021, it was revealed that a member of the Dutch political party GroenLinks had employed trolls to influence social media. He is alleged to have created several fake profiles, using them to engage in discussions and to voice his opinions (Knieriem, 2021). Another study in 2017 found that trolls had tweeted largely negative comments about left-wing party leaders and positive comments about the right (Borra et al., 2017). Pointer's big data study on trolls in elections found 476 Dutch-language trolls at the time of the most recent US presidential elections (Pointer, 2021a). Pointer does not state whether this figure represents an increase or how it compares to other countries.

The University of Amsterdam studied more than five million tweets mentioning Dutch political leaders between 1 January and 31 December 2020 (Sabel & Verhagen, 2021). The researchers also looked at 'troll-like behaviour' and the 'classic bullies', the heavy internet users who sometimes use automated trolling to try to hijack discussions and bombard politicians with tweets. Just under 0.4% of all accounts showed evidence of this type of behaviour (Sabel & Verhagen, 2021). These approximately 1,000 accounts in fact generated 15% of the five million political tweets studied (see Figure 3.5). The vast majority of these accounts are on the right wing of the political spectrum (Sabel & Verhagen, 2021).

One example of a Dutch troll account is @PeterBrekelmans, who has sent almost 125,000 tweets since mid-2020. This troll account tends to tag or comment on tweets by left-wing politicians. Two other troll accounts are @loweshenny (more than 26,500 tweets and retweets between October 2020 and March 2021) and @ChrisHagenviet (16,500 tweets since 2013) (Sabel & Verhagen, 2021). In 2018, it emerged that the Dutch singer Dotan had deployed an army of approximately 140 trolls (fake profiles on Facebook and Twitter) to boost his image and income (Miserus & Van der Noordaa, 2018).

Shame-sexting, sextortion and revenge porn

According to knowledge centre Movisie, **shame-sexting** is 'the uninvited forwarding of sexually oriented images with the aim of pillorying the person depicted' (Movisie, 2019). Sexting in itself is not harmful or illegal and, according to the Dutch expertise organisation Bureau Jeugd & Media, it is part of healthy experimentation by adolescents (Kleijer, 2015). According to Rutgers, the Dutch knowledge centre on

sexuality, sexting only becomes harmful if others make *unwanted* use of sexually-tinged images (Rutgers, 2018). It is then known as 'shame-sexting'.
Combined with other phenomena such as hacking and phishing, and amplified by **online** mechanisms (virality, scalability), shame-sexting can be more **harmful** than would be the case offline. That became clear recently when a 13-year-old girl, Desteny, jumped from a flat after sex videos of her were circulated online. The sharing of sexual images may also lead to sextortion or revenge porn.

Sextortion is a form of extortion in which someone threatens to disseminate images of a sexual nature without the victim's consent in order to procure additional images, money, or sexual acts (Patchin & Hinduja, 2020; Wolak et al., 2018). Sextortion, like blackmail, may be a criminal offence under Section 318 of the Dutch Criminal Code (blackmail and extortion).

Revenge porn is the unauthorised possession, disclosure and distribution of stolen sexual images, for example by hackers, partners, ex-partners, child abusers, rapists and human traffickers (Rijksoverheid, n.d.). Unlike sextortion, revenge porn is not extortion but an attempt to deliberately harm victims by disclosing the images. On 1 January 2020, revenge porn became a criminal offence under Section 139h of the Dutch Criminal Code. It is now an offence both to possess and to distribute sexual images without the subject's consent. In 2020, Dutch Minister of Justice and Security Ferdinand Grapperhaus instructed his Ministry's Research and Documentation Centre (WODC) to investigate whether the creation and distribution of 'deep nudes' – in which artificial intelligence is used to make fake nude images of someone – should also be made a criminal offence.

Girls with eating disorders sometimes fall prey to pro-ana coaches who are looking for sexually explicit images. According to the Dutch center for child and human trafficking (CKM), the coaches are usually men between 20 and 30 years of age who pose as weight-loss advisers but soon start asking for sexually-tinged photographs or videos (Simons et al., 2020). Intentionally harming someone's health is regarded as equivalent to assault under Section 300 of the Dutch Criminal Code, meaning that online pro-ana coaches may be committing a criminal offence.

Scale
Experts are concerned about sexually transgressive behaviour online and the disclosure of private images among minors. Nevertheless, actual figures are few and far between and much remains shrouded in mystery (Wagemakers & Toksöz, 2021). Records only reflect reported cases. The Amsterdam police department dealt with 224 cases of sextortion between 2019 and the end of 2020, and Helpwanted.nl received 2,683 reports of sextortion (Wagemakers & Toksöz, 2021).

A recent study by Statistics Netherlands showed that in 2020, nearly 30% of 16- to 18-year-old women and 23% of 18- to 24-year-old women reported having been

exposed to unwanted sexual behaviour online in the past 12 months (CBS, 2020a). In addition, 9% and 8% of their male peers respectively had also been affected (Wagemakers & Toksöz, 2021). In the study, sexual harassment or unwelcome behaviour covered such matters as sexually tinged remarks, unwanted touching or being coerced into acts of a sexual nature. More than a third (36%) have kept their experience of sexual harassment to themselves (CBS, 2020a).

Helpwanted.nl, the helpline for online sexual abuse victims, saw the number of reports it received double during the first lockdown. The helpline commissioned a study to find out whether children and youths had indeed had more negative experiences online than before the lockdown. They had the organisation Safer Internet Centre Netherland survey 1,164 youths from 12 to 25 years of age. The outcomes revealed a downward trend rather than an upward one. Prior to the pandemic, a third of all youngsters had had a negative sexual experience online, whereas after it was more than a quarter (28%) (ECP, 2021). Before the pandemic, 64% of adolescents said they found online sexting annoying; during the pandemic, this figure dropped to 39%. The share of adolescents between the ages of 12 and 18 who see sexting as something positive has increased during the pandemic from 19% to 28% (ECP, 2021).

Images shared by sexting can easily lead to shame-sexting and revenge porn. Approximately 14% of boys and 9% of girls have seen a nude image or sex video online of someone they know. Other figures are 23% and 24% respectively, depending on educational level: the less educated, the higher the percentage (Rutgers & Soa Aids Nederland, 2019). International research shows that in the six months before they were interviewed, 3.1% of 4,453 children and adolescents between the ages of 11 and 18 had shared (or almost shared) nude photographs of themselves (Lewis, 2018).

Figures provided by the Dutch Ministry of Justice and Security (2021) show that between 2015 and 2019, the police handled nearly 16,500 incidents of sexual violence against children. In 2017, almost 10% of the incidents registered concerned unwanted sexting. By 2019, this figure had risen to almost 14% (Ministerie van Justitie en Veiligheid, 2021).

What is worrying about online sexual incidents and online stalking is that 40% to 50% of the perpetrators are complete strangers to the victim (CBS, 2018). Halt, a Dutch agency that works with juvenile delinquents, has noted a growing demand for its awareness-raising and prevention programmes addressing online safety and sexting. One explanation is that children are using smartphones at an increasingly young age and are informed too late about the risks of mobile phone use. Often this does not occur until group 8 or their first year of secondary school, according to one of the experts interviewed.

Grooming

Grooming is the process whereby an adult develops a sexually abusive relationship with a minor through the use of cyber-technology, such as mobile telephones, internet games and chat rooms (Lorenzo-Dus & Izura, 2017). It is also referred to as ‘Online Grooming’ (EOKM, 2020). The interaction between adult and child by itself can be sexually pleasurable for the adult and therefore constitutes a form of sexual abuse. **Online** grooming differs from offline grooming in that it is much easier for adults to initiate contact with children (anonymously). Research also shows that youngsters are more likely to engage in high-risk behaviour online (Whittle et al., 2013).

Grooming is **harmful** because it can lead to the sexual abuse of minors. Even if the abuse only occurs online, it creates an unsafe situation for minors, regardless of whether they have been forced to commit acts. Grooming has been a criminal offence under Section 248e of the Dutch Criminal Code since 2010. For a crime to have occurred, the offender must have undertaken an ‘action intended to bring about’ a meeting with the victim.

Scale

We have been unable to find figures indicating the overall scale of grooming in the Netherlands. As with many other phenomena, grooming is often not reported to the authorities. EOKM and Helpwanted.nl define the phenomenon of ‘grooming’ as ‘soliciting children online for sex’. Of the total number of times that someone contacted Helpwanted (6,318 times in 2020), about 9% were related to grooming or being approached for sex online (EOKM, 2020, p. 13). In 2019, the total number of reports of child pornography was five times higher than in 2015, rising from 5,534 in 2015 to 25,628 in 2019 (Ministerie van Justitie en Veiligheid, 2021).

Extreme pranks

Extreme pranks are a form of interpersonal humiliation involving a three-way relationship between the one who humiliates, the victim and the witnesses (Hobbs & Grafe, 2015). Often, the purpose of the prank is to provoke an emotional reaction. Such pranks also typically involve people in unequal power relationships, for example parents who play pranks on their children. **Online**, extreme pranks often take the form of videoed ‘offline’ pranks, with the camera zooming in on the victim’s reaction. One example of an extreme online prank that began to emerge in 2002 is the Scary Maze Game. In this game the victim of the prank has to solve a maze puzzle that requires a high level of concentration from them. Suddenly, the game is interrupted by an ear-piercing scream and ghastly images from horror films. Prank

videos showing children reacting to the images by crying or screaming became very popular on YouTube and other sites.

In January 2019, YouTube announced that it would prohibit videos of dangerous pranks and pranks that could lead to serious physical injury (YouTube, 2019). This includes pranks where someone is tricked into thinking they are in severe danger. In practice, it is difficult to ascertain when pranks are **harmful**. Researchers do not always agree about whether a prank is benign or harmful and sadistic in nature. Perpetrators and witnesses of extreme pranks often put their own feelings before those of the victim. A 2017 study found that both perpetrators and witnesses often enjoy extreme pranks, even when they know that the victim is being hurt (Burris & Leitch, 2018).

The shock response of victims is a key element of extreme pranks. This means that victims often do not consent to taking part in the prank. It is also impossible to verify whether victims are aware that they are appearing online in an extreme prank video. The physical and emotional trauma of the prank itself is amplified because the victim's humiliation continues online. Online pranks may therefore be even more harmful than offline ones, which are not shared with a large audience. Extreme online pranks may be a criminal offence if they are accompanied by physical violence.

Scale

We found very few scholarly publications on the subject of extreme online pranks. It is therefore very difficult to gauge the scale of the phenomenon in the Netherlands. In addition, many of the examples of violent pranks cited in the news media are not Dutch. It seems, then, that we do not yet have a clear picture of the scale of extreme pranks in the Netherlands and the number of victims they claim. There is no national helpline or foundation that assists victims of this phenomenon, as there is for racism or shaming.

Minors guilty of committing 'happy-slapping' pranks – physically attacking a person at random and posting images of the incident online – have been arrested in the Netherlands, however (Nu.nl, 2020). The happy-slapping cases took place between 21 August and 7 October 2020 around Osdorpplein and Tussen Meer in the Nieuw-West district of Amsterdam.

Examples of violent and harmful pranks are more common in the US. For example, Heather and Michael Martin played harmful pranks on their children, with their videos gaining hundreds of thousands of views. The couple made money from their videos and eventually lost custody of their children partly for that reason. The couple's YouTube channels were removed in 2017 (RTL Nieuws, 2017).

Popular but controversial English-language prank vloggers are Roman Atwood (with 16.5 million YouTube subscribers), Ken Duchamp (514,000) and Sam Pepper (2 million). They earn or earned huge amounts of money every year by pranking their families or other victims. To put these numbers in perspective: one of the most popular YouTubers PewDiePie had 109 million YouTube subscribers at the time of writing, more than anyone else in the world.

3.6 Cyber fraud

Cyber fraud is the use of high-tech tools to deceive for personal gain. It includes hacking, phishing, catfishing and crypto fraud. It is fraud committed mainly for reasons of *greed* (financial or personal gain) in which the perpetrator often assumes a different identity online by hacking, phishing or catfishing. Many of the phenomena that fall into this category are criminal offences and are sometimes labelled as ‘cybercrime’. None of them would exist without the internet, hence the designation ‘cyber’. There is therefore no need to spell out the **online** nature of the phenomena in this category.

Cyber fraud is used not only for financial gain but also to extract information. Hacking can be used to steal nude photos, for example, which the perpetrator can then use to blackmail the subject. Interestingly, phishing is relatively more common than hacking and can be very harmful. During the pandemic, the number of rogue web shops and their victims has grown considerably.

Ignorance is a huge problem when it comes to cyber fraud. Many victims do not report cyber fraud to the police. ‘Traditional’ types of crime are reported far more often than hacking and phishing in the Netherlands, for example (van de Weijer et al., 2019). Cyber fraud often involves deceiving an individual (except for hacking, where vulnerabilities in computer systems are exploited rather than individuals). This differs from information manipulation, in which the deception targets larger groups of people.

Hacking

Hacking has been defined as activities involved in attempting or gaining unauthorised access to IT systems (Furnell, 2009, p. 173). Not all forms of hacking are harmful. Ethical or ‘white hat’ hackers do not damage systems; they in fact want to make IT systems safer. ‘Black hat’ hackers, on the other hand, are specifically looking to **harm** systems or steal confidential information (Aiken et al., 2016). Both our analysis of the literature and our interviews show that young hackers often see

hacking as a game or a challenge to overcome technical barriers and find flaws in systems. They are often not properly aware of the harm that hacking can cause and the ramifications for themselves if they are discovered.

Some hackers are eager to disclose classified information, such as the ones who leaked confidential information in Hillary Clinton's e-mails to WikiLeaks in 2016. The term 'hacktivist' (hack + activist) refers to hackers who break into IT systems for political or socially motivated purposes. One example is Aaron Swartz, who broke into the digital repository JSTOR and downloaded large numbers of academic articles because he believed in open access (Naughton, 2015). After his trial, he committed suicide while awaiting sentencing.

Hacking is often used as a tool to support other forms of malicious behaviour. It can be used to steal personal data that is then exploited to threaten the victim. Or it can be used to hack someone's Instagram account and post in their name. As in traditional forms of burglary, the motives may differ from one hacker to another and may not always involve financial gain.

Hacking is a criminal offence, known as 'computer trespass', under Section 138ab(1) of the Dutch Criminal Code. To have committed computer trespass, an individual must intentionally and unlawfully gain entry to a computerised device or system. If the hacker also steals data, then Section 138ab(2) also applies, which states that copying, intercepting or recording such data is a punishable offence.

Scale

Hacking appears to be growing slightly more prevalent in the Netherlands, with the victimisation rate rising from 4.9% in 2017 to 5.5% in 2019 (CBS, 2019a). The age group most affected are 25 to 44-year-olds, at 6.4% in 2019. Older people over 65 are the least affected, with 4.1% being victimised in the same year (CBS, 2019a).

Only a small proportion of hacking victims report such incidents. Of all the cases of identity fraud, purchase and sales fraud, hacking and cyberbullying combined, only one out of eight (13%) were reported to the police in 2019 (CBS, 2019a).

Phishing

Phishing is fraudulently acquiring information about persons and organisations by e-mailing users links to fake versions of a popular website to trick them into providing sensitive details (Vayansky & Kumar, 2018). Unlike hacking, phishing makes use of social engineering techniques, i.e. deceptive tactics to obtain data. For example, by tailoring the content to the recipient's personal situation, the phishing email may appear trustworthy. Phishing does not involve breaking into IT systems, but rather exploiting human vulnerabilities.

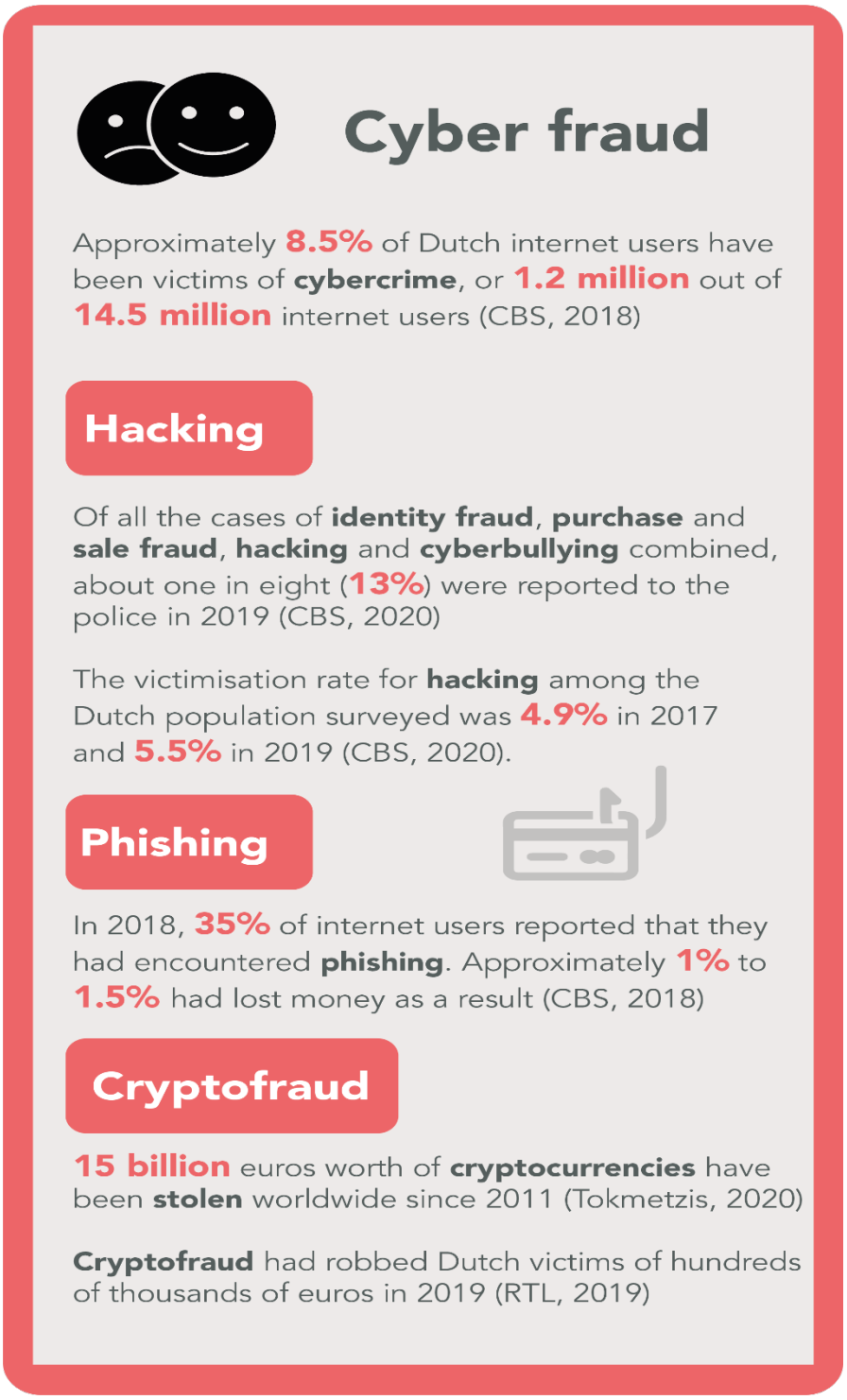
Victims of phishing are not only **harmed** financially but may also feel ashamed or lose faith in people. A well-known example of contemporary phishing is WhatsApp fraud, where criminals pretend to be the recipient's child in acute financial trouble. By exploiting the psychology of victims and by making the situation appear to be urgent, criminals posing as close relatives can get their victims to transfer thousands of euros to their accounts. Victims do not always go to the police because they are embarrassed by what has happened.

Phishing is a criminal offence in the Netherlands under Section 326 (fraud) and/or Section 225 (forgery) of the Dutch Criminal Code.

Scale

According to the police, the rise in cybercrime in recent years can be attributed mainly to phishing (see Figure 3.6). This is confirmed by figures from preceding years (Lastdrager, 2018). Between 2012 and 2014, 0.4% of the Dutch population over the age of 14 had experienced an incident of identity fraud (Lastdrager, 2018, p. 2). According to another survey, in 2015 about 4.5% of the Dutch population had (Paulissen & Van Wilsem, 2015).

The number of phishing reports accounted for 14% of the total number of reported victims in 2014 (Lastdrager, 2018, p. 2). The threshold for phishing is relatively low and perpetrators are difficult to trace. The first time that a software builder responsible for writing phishing programs was arrested in the Netherlands was in late 2020. The culprit was a 19-year-old who had earned around a hundred thousand euros with phishing (Heck, 2020).



Bron: Rathenau Instituut

Figure 3.6 Cyber fraud

Catfishing

Catfishing is an extreme form of online dating deception that involves falsely representing oneself to a potential romantic partner, without the intention of meeting in person (Mosley et al., 2020). By pretending to be someone else, catfishers try to improve their chances on the dating market. We classify catfishing as ‘cyber fraud’ because it involves defrauding individuals for personal gain, whereas information manipulation (including sock puppeting) involves defrauding larger groups of people in society. Catfishing could also fall under our category of ‘online harassment and violence’, depending on the perpetrator’s intentions.

Research indicates that men are more likely to engage in catfishing. Meeting a catfisher in real life can lead to ugly and dangerous situations for the victims, who often do not recognise them. Purposely using outdated photographs or lying about one’s age can also be considered catfishing. The ease of assuming a different identity and operating anonymously **online** makes catfishing a characteristic internet phenomenon. Catfishing can also be used as a tool for perpetrating other forms of online harmful behaviour, such as cyberbullying.

The mental **harm** suffered by victims of catfishing is considerable: their confidence is seriously shaken and their safety may also be compromised. A survey of vulnerable LGBTQ+ men in the USA found that catfishing can make the online environment even less safe for vulnerable groups. One of the respondents gave the example of a victim’s acquaintances pretending to be someone else online and obtaining sensitive personal information as a result (Lauckner et al., 2019). The researchers recommended that social workers be especially alert to catfishing among sexual minorities because of their vulnerability and frequent use of online dating apps.

Catfishers often use images of other people without their consent. That is harmful for the image’s owner because their identity has been stolen and misused. As a result, catfishing can have two sets of victims: those whose identities have been stolen and those who have been duped.

Identity fraud can be a crime in the Netherlands depending on the type of offence, for example under Section 310 of the Dutch Criminal Code (theft of identity). But catfishing victims are not always victims of identity fraud, nor are they always victims of fraud (if the perpetrator’s motive is not financial). That makes catfishing difficult to prosecute.

Scale

National Dutch newspaper *Algemeen Dagblad* reported in February 2021 that the police do not have figures on catfishing as a form of cyber fraud (Quekel, 2021).

The article states that catfishing does seem to be on the rise, however, based on information provided by the Dutch anti-bullying alliance Stop Pesten Nu, Helpwanted.nl and the government's Central Identity Theft and Error Reporting Centre (CMI).

Cryptofraud

Cryptofraud is a form of deception in which people are persuaded to buy cryptocurrency in order to boost its price. The perpetrators then quickly sell their holdings in that currency, leaving the victims with financial **harm**. The growing popularity of cryptocurrencies as a novel form of investment has also given rise to new types of cryptofraud. The Dutch Authority for the Financial Markets (AFM) and the Dutch central bank DNB already warned about the risk of fraud posed by cryptocurrencies back in 2018 in their report *Cryptos: Recommendations for a regulatory framework* (AFM & DNB, 2018). Among the risks that they cite are the highly technical nature of cryptocurrencies, price manipulation, and hacks of online cryptocurrency exchange platforms.

'Pump-and-dump' practices – whereby masses of people artificially inflate the value of a security and then sell all their holdings at a profit – are prohibited on stock exchanges. Dutch public news service NOS devoted an article to this practice on 6 May 2021 in which it noted that cryptofraud was evidently becoming more common (NOS, 2021b). The decentralised nature of cryptocurrencies and the anonymity afforded by the internet make it easy for offenders to evade enforcement. Invitations to join pump-and-dump operations are circulated on Twitter, Discord, Telegram and other social media platforms. Perpetrators tend to focus on smaller and unknown cryptocurrencies, as these are easier to manipulate than Bitcoin, for example.

When cryptofraud manifests itself as embezzlement and fraud, it is regarded as a criminal offence under Sections 321 and 326 of the Dutch Criminal Code. To date, cryptofraud has rarely been prosecuted, partly due to the unregulated nature of the cryptocurrency market. Recent research by the University of Rome shows that cryptoexchanges themselves could take more responsibility to combat cryptofraud, for example by imposing stricter controls and exercising closer monitoring (La Morgia et al., 2021). The European Commission is working on proposals to rein in the cryptomarket as part of its Digital Finance Package.

Scale

The Covid-19 pandemic led to a further rise in the global trade in cryptocurrencies in 2020. Stock exchanges have always been subject to manipulation, but such practices have recently attracted fresh interest due to developments in cryptocurrency markets. For example, in January 2021 a group of activists on

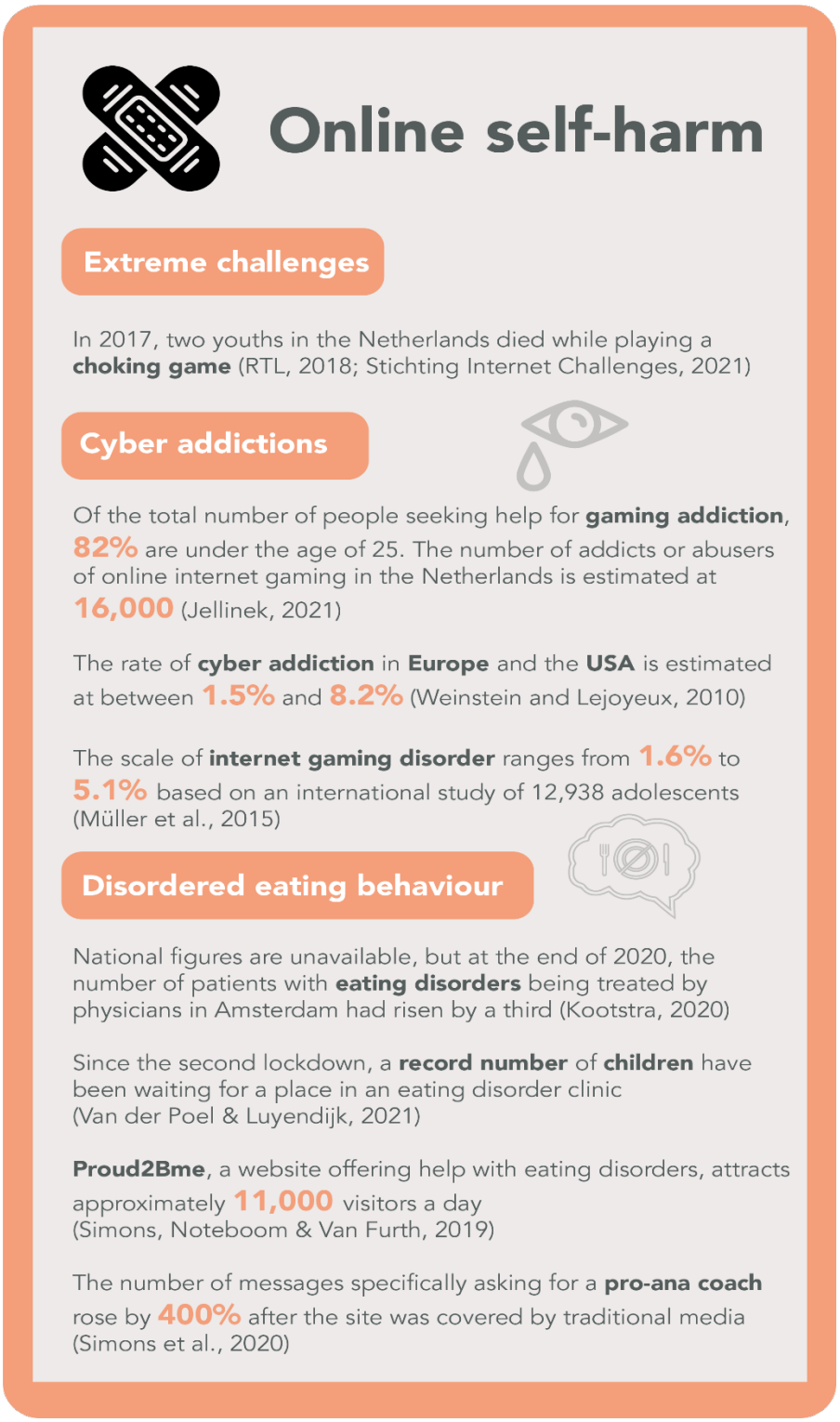
Reddit began an operation against various hedge funds and managed to raise the share price of GameStop by more than 1900% (La Morgia et al., 2021). Although the operation did not involve a cryptocurrency, the mechanisms behind it were similar.

The Dutch investigative website *De Correspondent* reported in 2020 that 15 billion euros worth of cryptocurrencies have been stolen worldwide since 2011. It based this figure on an analysis of legal and media reports (Tokmetzis, 2020). It is unclear just how many people in the Netherlands have fallen victim to cryptofraud. Specific fraud cases sometimes reveal the losses suffered by the Dutch, for example when news broadcasting service RTL Nieuws reported in 2019 that cryptofraud had robbed Dutch victims of hundreds of thousands of euros (RTL, 2019).

3.7 Online self-harm

Online self-harm refers to behaviour in which a person harms themself without a perpetrator necessarily being involved. The phenomena involved are **extreme challenges**, **cyber addictions** and **disturbed eating behaviour**. In all such cases, online mechanisms are contributory factors and are harmful to the person exhibiting the behaviour. For example, social media can be addictive and online pro-anorexia (‘pro-ana’) content can induce disturbed eating behaviour. We purposely do not refer to motives here because the harm is self-inflicted and there is no traditional victim-perpetrator relationship in these phenomena. It is important to note, however, that both disturbed eating behaviour and cyber addictions are mental illnesses requiring treatment. The internet is integral to all the phenomena in this category, so we do not compare their offline and online versions.

Experts claim that the bleakness of the Covid-19 lockdown, which is forcing youths to spend much more time at home than they normally would, is an important factor in the growing number of young people struggling with mental health issues (see Figure 3.7). The crisis has also led to a rise in the number of hours young people spend online. There is a demonstrable correlation between exposure to online content depicting risky behaviour and viewers’ risky behaviour offline. This correlation was found for drug use, excessive alcohol use, disordered eating, self-harm, violence to others, and dangerous pranks (Branley & Covey, 2017).



Bron: Rathenau Instituut

Figure 3.7 Online self-harm

Extreme challenges

In online challenges, people are encouraged to complete certain tasks and then share a video of themselves doing so online. One well-known example is the Ice Bucket Challenge, which went viral in the summer of 2014. People were urged to dump a bucket of ice-cold water over themselves, donate to research on ALS (motor neurone disease), and challenge others to do the same. Online challenges vary considerably in terms of the risk they pose to participants. The Blue Whale Challenge encouraged viewers to engage in self-harm and eventually to kill themselves (Khasawneh et al., 2020). The media often cover such challenges after they have resulted in casualties. We classify extreme challenges as self-harm because they often induce people to engage in dangerous activities that may lead to self-injury.

Unlike pranks, videos of challenges are shared by those who respond to and engage in the challenge themselves, with humiliation not being a major factor. Challenges are an especially popular means of entertainment on TikTok and give its users a means of engaging in a shared experience with their age peers. Clever marketing firms also jump on the online challenge bandwagon by launching challenge-related campaigns, thus drawing attention to their products.

Not all online challenges are **harmful**, but dangerous challenges can have significant consequences. Online, warnings about the risks of online challenges mainly come from educational websites targeting parents and from traditional media. Articles bearing such titles as ‘21 Dangerous TikTok Trends Every Parent Should Be Aware Of’ warn parents to keep an eye on their children when it comes to certain, potentially dangerous, challenges (Morris, 2021). Examples are the Cinnamon Challenge, in which the participant consumes a large quantity of ground cinnamon in a short period of time, or the Choking Game, in which adolescents strangle one another until they lose consciousness.

Extreme challenges may be a criminal offence, depending on what they entail. In 2018, for example, the Keke Challenge became popular; the participant would jump out of a moving car, do a dance and get back in. A spokesperson for the Dutch Public Prosecution Service described the challenge as a criminal offence under Section 5 of the Dutch Road Traffic Act [*Wegenverkeerswet*] because people participating in it were endangering themselves and others in traffic (NOS, 2018). Prohibition of a challenge and prosecution of its ‘instigators’ is difficult because it is often unclear who started the challenge or hashtag, and because every participant in fact encourages others to join in.

Scale

We have not found any Dutch figures on the total number of people who harmed themselves whilst performing extreme online challenges. There are two documented cases in the Netherlands of children dying as the result of an online challenge. The victims were Clay Haimé and Tim Reynders, both of whom died after participating in the Choking Game in 2017 (RTL, 2018; Stichting Internet Challenges, 2021). The Choking Game drew worldwide attention when a ten-year-old girl in Italy died after joining the challenge. The Italian government responded by temporarily blocking access to TikTok for users whose age could not be proved definitively (Agence France-Presse, 2021).The overall scale of harmful online challenges is difficult to ascertain; injuries are often not documented and only generate media attention when they are severe, even though they can also result in less serious injuries.

Cyber addiction

Cyber addiction, also known as internet addiction, is excessive and uncontrolled online activity with prolonged internet use, especially in social networking, online gaming and use of pornography sites (Liu et al., 2020; Müller et al., 2015). Other compulsive and obsessive behaviours that may arise as a result include excessive shopping or gaming or obsessively searching for health-related information (cyberchondria) (Aiken, 2016).

Social media platforms often have addictive features, such as ‘endless scrolling’ on TikTok (Montag et al., 2021). ‘Likes’, personalised content and other mechanisms can also induce users to use social media longer than they intended. In turn, they produce more extreme content. Platforms thus encourage addictive behaviour.

Cyber addictions are harmful because research has shown them to be associated with chronic sleep deprivation, anxiety and emotional problems, among other things (Alimoradi et al., 2019; Cerniglia et al., 2017). Like any psychological issue, cyber addiction is not a criminal offence. Initiatives are underway worldwide to legislate against the addictive properties of social media, however. For example, a bill submitted to the state legislature of Missouri (it had not yet been enacted at the time of writing) prohibits companies from exploiting human psychology and restricting people’s freedom of choice in that way.

Scale

Internet addiction and internet gaming disorders are becoming increasingly common (Chia & Zhang, 2020). It is estimated that the percentage of the population in Europe and the USA suffering from internet addiction is between 1.5% and 8.2% (Weinstein & Lejoyeux, 2010).

The scale of internet gaming disorder varies between 1.6% and 5.1%, based on an international study involving 12,938 adolescent subjects (Müller et al., 2015). Approximately 3% of online gamers can be regarded as ‘game addicts’: they have difficulty quitting, play more than they intend to, get too little sleep and lag behind in their homework (van Rooij et al., 2012). Of the people seeking help for internet gaming disorder, 82% are younger than 25. The number of addicts or abusers of internet gaming in the Netherlands is estimated at 16,000. Of these, 537 are receiving treatment (Jellinek, 2021).

Spending excessive amounts of time online increases the risk of obesity among young people and is also linked to internet addiction. Another risk is paediatric venous thromboembolism (VTE), sometimes referred to as ‘gamer’s thrombosis’, a condition that can be fatal after prolonged gaming. The incidence of this condition among adolescents has increased over the past two decades (Kohorst et al., 2018). We have not found any figures on gamer’s thrombosis in the Netherlands.

Disturbed eating behaviour

Eating disorders are psychological disorders characterised by disordered eating behaviour and/or compensatory behaviour (self-induced vomiting, laxative misuse). People with an eating disorder have a distorted body image, are obsessed with their weight or body shape, and are terrified of gaining weight (NJi, 2019).
Online, people with eating disorders interact in ‘pro-ana’ or ‘pro-mia’ communities. ‘Pro-ana’ and ‘pro-mia’ stand for professional anorexia and professional bulimia, respectively. They are the names of online groups consisting (primarily) of young people with eating disorders who are active on online forums, chat rooms and websites providing information and a space for interaction aimed primarily at promoting, supporting and sustaining eating disorder-related behaviour (van Furth et al., 2011).

In 2020, the online magazine *Wired* reported that TikTok was full of pro-ana content that was made easily accessible to young girls by its recommendation algorithms (Gerrard, 2020). National Dutch newspaper *de Volkskrant* reported on challenges such as ‘Can you wrap the cable of your EarPods around your waist twice and then tie a knot in it?’ and seemingly innocent memes about eating disorders (Bouyeure, 2020). This is how TikTok users make pro-ana content seem whimsical and therefore appealing.

Disturbed eating behaviour can cause even more harm online if the content and relevant communities reinforce people’s body image and eating habits. It is particularly worrying that pro-ana content is now packaged as funny and

identifiable, and that it is not always meant to be malicious. That makes it harder to qualify as harmful content. People who promote pro-ana content are not always committing a crime. In 2019, Dutch Health Minister Hugo de Jonge commented that he did not see any reason to ban pro-ana content online as he thought doing so could in fact hamper people searching for help (Ministerie van Justitie en Veiligheid, 2019). Behaviour linked to pro-ana content may be a criminal offence, however, because intentionally harming someone's health is regarded as equivalent to assault under Section 300 of the Dutch Criminal Code. If minors are involved, it may even constitute child abuse, according to the Dutch Child and Human Trafficking Centre (CKM) and the Ursula clinic for eating disorders in a recent study on pro-ana coaches (Simons et al., 2020).

Scale

'Thinspiration' (content that inspires people who are looking to lose weight, such as diet or exercise tutorials), TikTok challenges and memes about eating disorders are becoming increasingly popular and children are getting smarter about bypassing algorithms by using variations on hashtags (Bouyeure, 2020). Health apps, pedometers and other trackers now found on smartphones and watches are also fuelling eating disorders.

By the end of 2020, the number of patients with eating disorders being treated by paediatricians in Amsterdam had risen by a third. They suspect this is connected in some way with the Covid-19 pandemic (Kootstra, 2020). There are no national figures for the Netherlands (Kootstra, 2020). Since the second lockdown, a record number of children, some as young as ten, have been waiting for a place in an eating disorder clinic, according to national newspaper *NRC Handelsblad* (Van der Poel & Luyendijk, 2021). The waiting time varies from six weeks to as much as six months in the central regions of the Netherlands. The number of underage anorexics admitted involuntarily to a facility is also on the rise. In the first nine months of 2020, 24 young people were subject to involuntary admission orders (Van der Poel & Luyendijk, 2021).

General research (not specific to the online environment) shows that approximately 0.3% of 13 to 18-year-olds in the Netherlands suffer from anorexia nervosa (Verhulst et al., 1997). A similar percentage has bulimia nervosa, according to this study (NJi, 2019). Anorexia is much more common in women than in men (95% are women), usually occurring in adolescence and young adulthood. Proud2Bme, a website offering help to young people with eating disorders, attracts approximately 11,000 visitors a day (Simons et al., 2020). The number of online messages specifically asking for a pro-ana coach rose by 400% after the site was discussed in traditional media (Simons et al., 2020).

3.8 Conclusion

Our taxonomy of harmful and immoral behaviour online differentiates between six categories of behaviour divided into 22 phenomena. This approach has yielded a wide variety of new and older behaviours that any internet user might encounter, ranging from disinformation to online discrimination and shaming.

Rathenau Instituut has thus taken the first step towards describing and identifying the nature and scale of harmful and immoral behaviour online in the Netherlands. The variety and availability of data regarding that scale make clear that all Dutch people run the risk of becoming involved in this behaviour as a victim, perpetrator or bystander. Anyone can be affected by one or more forms of the phenomena described in this chapter. For certain phenomena, some groups are more at risk than others, depending on their age, gender, race, sexual orientation, religious beliefs or level of education. It is difficult to generalise based on the available data, however.

The diverse nature of the phenomena we address and the absence of precise definitions and systematic measures mean it is impossible in the context of this study to determine which phenomenon is growing fastest or is most worrying. The answer also depends on the chosen criteria. For example, are we looking at the number of victims or at the severity, scale or risk of harm in the future? We conclude that all phenomena are worrisome in their own way, for society as a whole, for individuals or for groups of individuals. That is why we refrain here from prioritising the phenomena in our taxonomy.

Case: Disinformation

The case below is fictitious and intended to illustrate possible risks that may arise from the phenomenon of disinformation. It is, however, based in part on a combination of incidents that occurred in the Netherlands and abroad. We conclude the case by discussing the mechanisms and stakeholders involved. Chapter 4 looks at these mechanisms in detail, while Chapter 5 discusses the stakeholders. In Chapter 6, we offer suggestions for preventing and addressing situations such as the one described in the case.

Case
Mirjam, a teenager, has been inspired by a global movement of young people who have committed themselves to halting climate change. She is extraordinarily digital-savvy. She soon begins interacting with like-minded people from all over the world who corroborate her views. Initially, she connects with them in open Facebook groups and on Twitter, but she soon also finds her way into private chat groups on platforms such as Signal and Telegram. The groups use these channels to prepare social media campaigns, for example to promote hashtags as trending topics or to defend their supporters on social media when they are criticised. The channels are also used to plan physical demonstrations. The groups sometimes push the limits of lawfulness and draw inspiration from examples abroad, for example blocking roads unannounced, intimidating others or sabotaging companies. Even though Mirjam has never met the group administrators in person and some of them are using pseudonyms, she trusts them completely. After all, they all support the same cause, no matter where they are in the world.

At a certain point, a message is circulated in an international Telegram group that the Dutch government is deliberately underreporting nitrogen emissions from livestock farming. According to calculations by foreign researchers shared in the chats, things are much worse than presented. There are also suggestions that there may be a conspiracy behind it all. Dutch mainstream media also get wind of these reports, but question the alternative calculations. Mirjam’s international contacts, however, warn her not to fall for the news reports. The media must be in cahoots with the government, they say. Mirjam thinks it is time to expose this alleged malpractice and to take action. She organises a spontaneous demonstration in The Hague on a chat app. The local authority is prepared because it has been tipped off by the national intelligence service, which is monitoring the chat channels. Nevertheless, the demonstration turns into a violent confrontation with livestock farmers who believe the opposite, i.e. that that the government has exaggerated the emissions figures. They too had evidently been mobilised through private channels

but that the Dutch intelligence service had not spotted in advance. What Mirjam does not know is that the chat channel's international coordinators are not climate activists but in fact work for a foreign intelligence service. The same could be true of the channels in which the livestock farmers exchange messages.

Sometime after the violent confrontation, the Dutch intelligence service uncovers the deception. The unrest continues, however. The revelation causes people to be even more sceptical of reporting. They no longer know who they can or cannot trust. Their faith in one another has plummeted.

Reflections

This case shows how disinformation is harmful to society. Climate policy is already a source of social unrest and the public debate often bears the hallmarks of polarisation. It is therefore a subject that lends itself very well to disinformation campaigns. We know that state actors may seek to stir up unrest among their geopolitical opponents and polarise their societies. For example, research into the activities of the 'Russian troll factory' Internet Research Agency (IRA) in St Petersburg has shown that 'trolls' organised anti-racism demonstrations in the USA during the 2016 presidential election. It has also been shown that the IRA coordinated protests by anti-Islam and pro-Islam groups at the same location and at the same time (Bertrand, 2017). Interference of this kind is likely motivated by a desire to gain strategic advantages on the world stage. The democratic societies affected have their attention diverted to internal conflicts, leaving them oblivious or unresponsive to geopolitical events.

The case illustrates various mechanisms and phenomena. The climate activists contribute unwittingly to a **disinformation** campaign by a state actor. Internet use is a **daily habit** for them, but despite their digital skills they are vulnerable to deception. They believe that right is on their side and exhibit the features of **digital vigilantism**. **Syndication** and **hyper-connectivity** help them find like-minded people. They also understand how algorithms work and therefore know how to achieve **virality** and how to influence a public discourse by manipulating algorithms. Young activists are perfectly comfortable operating in **anonymity** because they are aware that having a reputation as an activist could be detrimental to their later careers. What also contributes to their sense of **apparent lawlessness** is that actions that go unpunished in other countries are used as examples.

We also see several stakeholders in this case. The climate activists are not under the sway of the foreign intelligence service alone. The mainstream media try to shed light on the matter, but they have only limited influence over those who get their information from closed channels. In a way, mainstream media reporting can be interpreted as confirming a possible government conspiracy. The platforms that

offer closed channels also play a facilitating role. Chat apps like Signal are known to protect users' privacy, for example by allowing anonymous accounts and offering powerful encryption. Other platforms do moderate, but it is difficult for them to distinguish between channels operated by sincere activists and those run by malicious users.

We also see that disinformation can undermine trust in society even further. The national intelligence service infiltrates the chat channels. This form of surveillance may in itself be alarming for some people. Finally, an as yet unnamed but no less important actor is the general public, which hears about the violent confrontation and is likely to feel less safe in general as a result.

4 Mechanisms of harmful and immoral behaviour online

The online environment features certain mechanisms that influence human behaviour. These mechanisms may cause people to deal with values and rules differently online than offline. Besides the mechanisms of the internet, many other factors influence human behaviour, such as social, psychological, cultural or economic drivers. All these factors play a role in the emergence of harmful and immoral behaviour online. This study focuses on the mechanisms that characterise the internet.

4.1 Preliminary observations

Before we delve into the mechanisms, a few observations are in order. Chapter 3, on our taxonomy, described a very wide range of harmful and immoral phenomena. As the case studies also show, most phenomena involve multiple online mechanisms, and some mechanisms are influenced in turn by other factors. The mechanisms are therefore not the sole determinant.

Moreover, the causes and underlying mechanisms of several phenomena are not fully known. Even in the case of commonplace phenomena such as bullying, it is not possible to say with any certainty how they come about. It is beyond the scope of this study to analyse phenomena in their entirety. We do not claim to describe an exhaustive list of mechanisms here, but focus instead on the relationship between mechanisms that characterise the internet and harmful and immoral behaviour.

The internet consists of different types of environments that a user can enter, different types of social media platforms, websites and forums. This means that the availability and influence of mechanisms can vary greatly from one online environment to the next. For example, the design of a platform such as Instagram is based on very different mechanisms than a shock site. It would be virtually impossible to describe the full spectrum of environments here one by one. In fact, the aim of this chapter is to facilitate a broader understanding of immoral and harmful behaviour online by describing general mechanisms that are truly characteristic of the internet.

It is also important to realise that many of the mechanisms described here also have positive effects. They cannot be unilaterally dismissed as harmful, in other

words. The fact that the internet is accessible to all and that information can be disseminated quickly and widely is, for example, one of its great advantages. The implication is that merely combatting a mechanism may result in new, possibly major negative consequences. Online anonymity, for example, may fuel negative behaviour as a mechanism, but it also offers protection from harm.

Finally, the mechanisms are not natural phenomena but are often the result of design choices made by parties in pursuit of their interests or a particular objective. For example, both users and providers of online services and products often want information to be shared as freely and efficiently as possible. They view the potential adverse effects of this as an unfortunate externality.

Mechanisms

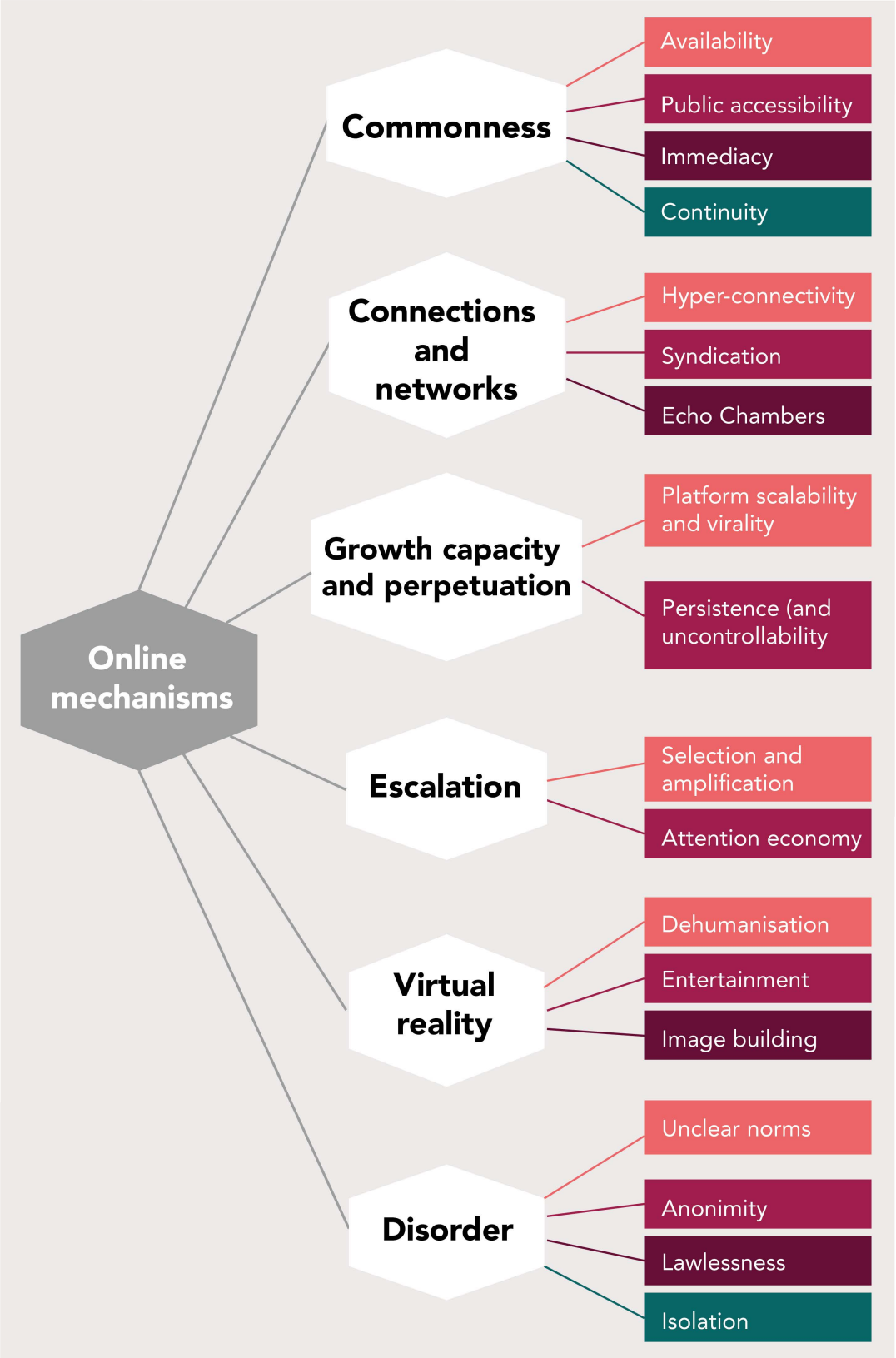
The mechanisms described below go some way to explaining how using the internet can encourage immoral or harmful behaviour. They are also described as factors that contribute to such concepts as ‘moral fog’ (Cocking & van den Hoven, 2018), ‘digital drift’ (Goldsmith & Brewer, 2015) or the ‘online disinhibition effect’ (Suler, 2004). These concepts are often cited in the literature and describe how people online are less able than offline to weigh up the consequences of their actions, less capable of exercising ethical judgement, or less inclined to act ethically.

The mechanisms described in this chapter consist of these concepts, other mechanisms identified by experts in interviews, and additional mechanisms that emerged from our literature review.

For ease of reading, similar mechanisms have been grouped under six descriptive characteristics of the internet:

1. Commonness
2. Connections and networks
3. Growth capacity and perpetuation
4. Escalation
5. Virtual reality
6. Disorder

An overview of all mechanisms and their classification can be found in Figure 4.



Bron: Rathenau Instituut

Figure 4 Overview of online mechanisms

4.2 Commonness

Internet use is an everyday occurrence. Many Dutch people spend most of their day staring at a screen (SCP, 2016). During the Covid-19 lockdown, children's screen time increased from about three hours to more than seven hours a day (Van Baars, 2020). Internet use has thus become routine for many people. This in itself carries the risk of unethical behaviour. It is in fact well known that when people operate by habit, they pay less attention to ethical issues (Vince, 2018). They post photographs without giving it a thought or forward messages without reflecting on the possible consequences.

The commonness of the internet has made its consequences almost inescapable. Even non-users or infrequent users are likely to have information about themselves online and be traceable there. They too can conceivably be harmed.

The following features contribute to the commonness of the internet.

Availability
The internet is available to almost everyone in the Netherlands. According to Statistics Netherlands, 95% of 12 to 55-year-olds use the internet daily, mostly on their smartphones (CBS, 2020b). A logical consequence of this availability is that people use the internet for everything, including immoral or harmful behaviour such as discrimination, racism or other forms of hate speech (Guan & Subrahmanyam, 2009). The accessibility of the internet also makes it very easy to transgress standards (Haspels-Goudriaan, 2020).

The availability of the internet has become crucial for many people and organisations. For example, we store important documents in the cloud and operational processes depend on internet connectivity. As a result, we can no longer always choose to use the internet voluntarily; it has become an inescapable necessity.

Public accessibility
The internet allows us to disclose much of what used to be private. Ninety-seven percent of households are connected to the web (Digitale Overheid, 2020) and no less than 84% of Dutch people use a smartphone to go online (CBS, 2019b). The majority of people now commonly share information on social media platforms such as WhatsApp, Facebook, YouTube and Instagram (Oosterveer, 2021). Although these platforms allow users to restrict access to that information to a select group, once it is shared, it is very easy to make it more widely available. Information may also be made public unintentionally or unwittingly. A photo intended for friends can be misused by malicious parties in a different context. A quarrel or dispute is also

more likely to become public. Such public accessibility also results in greater visibility.

People respond without inhibition on public social media because they feel as if they are in their private world. Technology philosopher Jan Bats describes three experiments in his PhD thesis showing that people ‘feel at home’ on social media because platforms allow them to personalise their environment and make it their own. People feel as if they are in their private domain, even though the messages they share are often publicly available. This causes them to be less inhibited in their responses and to judge others more harshly (Bats, 2019), something also known as the online disinhibition effect (Aiken, 2016; Suler, 2004).

The public nature of the internet is a product of design choices that have remained unchanged since the early days of the World Wide Web. The internet is in fact a network of computers in which each computer is visible to all others under normal circumstances. The advantage of this design is that information can be shared quickly with everyone. The disadvantage is that a poorly secured computer in the Netherlands can be hacked by any internet user anywhere in the world, for example. Internet services such as social media often do allow users to restrict the public nature of their participation. For example, users of LinkedIn can block strangers from tracking down their profile. Alternative designs are also possible in which users have more control over their visibility right from the start (see Chapters 5 and 6).

Immediacy

Interactions on the internet usually have a direct impact. Messages posted on a public platform such as Twitter are visible straight away. Comments and ‘likes’ can also be posted immediately. When people act quickly, they often do so instinctively or emotionally, and any biases they may have are more readily apparent. According to Kahneman (2011), fast thinking of this kind may keep them from the sort of thoughtful responses that take ethical concerns into account.

Experts have also pointed out that the immediacy of the web encourages impulsiveness, a character trait that is particularly prevalent among young people with ADHD, for example (Aiken, 2016, p. 72; Kaakinen et al., 2020). They also suggest a form of ‘tempocracy’, i.e. that the person who generates information quickest or most often sets the tone of the conversation. Such hasty behaviour may explain why people lose sight of standards and rules.

The negative effect of immediacy is reinforced by certain platforms such as 4chan and Snapchat, where content is visible for only a short time and then disappears altogether (unlike many other platforms and websites). The non-permanent nature

of posts seems to encourage users to make their content as provocative as possible, precisely because they know that their post will disappear within hours or minutes (Ludemann, 2018, p. 93).

Internet users as a whole also benefit enormously from its immediacy. The big advantage of chatting, for example, is that communication is much faster than writing letters. Internet service providers and product vendors are therefore keen to prioritise speed. For example, 5G is faster than 4G, games must load as quickly as possible, and fast internet connections allow us to respond immediately to live video streams.

Continuity
The internet is a 24x7 affair. The fact that it never takes a break also means that the risk of harm is ever-present. This may explain why people are constantly going online to check whether anything has happened to them. More than half of smartphone users turn on their device more than 25 times a day, and a quarter of users more than 50 times (Stil, 2020). For some, the smartphone has come to feel like a vital extra limb and they suspect that they are addicted to it (see also section 3.7 on cyber addiction) (*RTL Nieuws*, 2020). One troublesome consequence of the web's continuity is that once someone is harmed, the harm can go on and on. They are drawn into an endless loop with no opportunity to escape.

4.3 Connections and networks

Technically speaking, the internet is primarily a network that connects devices and, consequently, users. The following aspects of this network can encourage harm and immoral behaviour.

Hyper-connectivity
According to the popular theory of 'six degrees of separation', any two people in the world are linked by a string of six acquaintances, on average. In 2016, Facebook found that linking two of the 1.6 billion users on its platform required an average of only 3.5 acquaintances (Edunov et al., 2016). Online, it seems, people are more connected to others. They can also contact others directly, simply by tracking them on a search engine. This means that social network users can also very easily become the target of malicious parties or unintentionally get involved in immoral or harmful activities. In the case of phenomena such as grooming or sextortion, this mechanism makes it easy to target children or other victims.

Internet service providers have a keen interest in maximising connectivity. Usually, the more users they have, the more revenue they generate.

Syndication

The rise of social media and other online platforms has made it easier to find like-minded people and form groups. In the literature, this is referred to as homophily and 'online syndication' or simply 'syndication' (Aiken, 2016, p. 332). This mechanism is reinforced by the appeal of content that has already attracted a great deal of attention. It resembles the Matthew effect (Merton, 1968), a principle that states that a successful or famous individual (for example a scientist) will often get more credit than a comparatively unknown person, even if their work is similar. Many platforms display indicators – for example the number of likes, followers or views – that encourage syndication, and that can have negative consequences. For example, it appears that shamers attract followers on Twitter much faster than non-shamers (Basak et al., 2019), and that individuals in large groups may have a less acute sense of personal responsibility (De Vries, 2021).

Syndication can also have many other consequences. To begin with, an environment of like-minded people can lead to the normalisation ('everyone does it') of certain behaviour (LaFrance, 2020). By forming groups, people are more likely to achieve the sort of critical mass that is necessary to engage in harmful behaviour (Munn, 2021). Then there is the bystander effect, i.e. that no one intervenes when a norm, law or rule is broken. It is more difficult in the online environment than offline to determine whether or not you are a bystander.

In the case of digital vigilantism, the wish to belong is one of the underlying reasons for joining a group. Digital vigilantes may also be motivated by the status that participating bestows (Afuah, 2013). Finally, syndication also makes it easier for malicious parties to seek out vulnerable groups. For example, pro-ana coaches can easily track down people who have posted content showing that they struggle with their self-image.

Echo chambers

Syndication can also reinforce the psychological effect of confirmation bias. This means that people who only encounter information that they believe to be truthful are constantly having their own worldview enforced and legitimised (Sternisko et al., 2020). Groups with like-minded people are more likely to share that information, even if it is disinformation (Marwick, 2018).

In addition, they may end up in 'rabbit holes' (O'Callaghan et al., 2015) or 'echo chambers' (Auxier & Vitak, 2019). Because algorithms feed them more and more of the type of content that holds their interest longest, they no longer view content offering dissenting views (Sternisko et al., 2020). This can cause their thinking to

become narrow and rigid and may encourage radicalisation (Faddoul et al., 2020; *NRC*, 2021).

In general, internet service providers that base their business on an advertising revenue model are inclined to exploit syndication because they want to offer advertisers an online environment that keeps users there for as long as possible (see also section 5.2.2) (Kist, 2020).

4.4 Growth capacity and perpetuation

One of the characteristic features of digital information is the ease with which it can be multiplied. While the replication of immoral or harmful behaviour in the offline world is limited by the effort required of an actor, digital actions can be repeated mechanically, almost effortlessly. To bully someone in the offline world, you have to see them out physically, for example. Online, there is no physical distance to be bridged. The absence of such constraints therefore allows the rapid growth of immoral or harmful behaviour.

The benefits of rapid growth are unevenly divided between those who harm and those who seek to prevent or combat harm. A teasing meme, for example, can very easily be disseminated in chat groups, but the victim has little or no way of knowing where it is being circulated. To combat it, the victim would have to contact every single participant in those chat groups and ask them to remove the meme and not to share it further.

The following mechanisms contribute to the growth capacity and perpetuation effects of the internet.

Platform scalability and virality
Scalability is one of the internet's underlying design principles. This means that there are no intrinsic limits on the number of computers that can be connected. The internet is designed for growth. Any connected computer can basically share information with all the others.

Websites often try to shield information, for example by using a login system, but even then it is often very easy for users to transfer this information and share it with others. In addition, the phenomenon of syndication quickly rallies an interested audience. No matter how niche and peculiar a particular preference may be, it is easy to find an audience online large enough to be worth catering for.

Many platforms are designed specifically to facilitate the rapid dissemination of information. A hashtag can be trending on Instagram, a video can go viral on TikTok, or a website can appear again and again at the top of a search engine list. These are all examples of how algorithms accelerate the dissemination of information.

Persistence (and uncontrollability)
As soon as information is distributed online, any recipient can save that information and pass it on themselves. Combined with the public availability and immediacy of the internet, it may be impossible for the original sender to retract information even if they have only just posted it. So even if someone who has shared a misleading meme about vaccinations on WhatsApp deletes their own message later, chances are one of the recipients has already saved the image to their smartphone or even shared it with others. The persistence of information published online may significantly increase the amount of harm it can cause. Online accusations or shaming, for example, can scar a person for life because removal is practically impossible (Duin, 2020).

Alternative network structures are conceivable that would allow the original authors to retain full control over their data, but they would require a total overhaul of the internet's design. Until then, the best way to retain control is to simply not share information online. After all, even on platforms where content is only visible for a short while, screenshots can be made and continue to circulate online forever (Ludemann, 2018).

4.5 Escalation

The dissemination of information online can have a power escalation effect. Something overwhelming or impressive can quickly attract a lot of attention. The following online mechanisms play a role.

Selection and amplification
People use a variety of tools to navigate the internet, such as search engines and recommendation algorithms. These tools are in fact selection tools. Their use may encourage immoral and harmful behaviour, for example by causing more users to be exposed to immoral or harmful behaviour (amplification).

Some argue that the selectivity of these tools threatens media pluralism and that it traps people in a 'filter bubble' (Pariser, 2012). While there is no evidence for this in research by the Dutch Media Authority (Commissariaat voor de Media, 2019), the findings do acknowledge the influence of platforms on public opinion.

Recommendation algorithms, for example, incorporate editorial choices and thus influence political discourse. Platforms can promote information to a greater or lesser extent or even remove it altogether.

Various experts consulted for this study are worried that selection and amplification can lead to polarisation. They believe that the creators of these tools should take more responsibility and mitigate their adverse effects. At the same time, they are concerned about the tendency of platforms not to modify tools but to remove content and users instead. As a result, these users may seek refuge on other platforms where they are likely to encounter more like-minded people and fewer users who disagree with them, or where different terms of use apply. Parler, for example, has functionalities similar to those of Twitter, but is increasingly described as a 'far-right popular alternative' because of its more lenient terms of use (*Algemeen Dagblad*, 2021).

There are also signs that content creators and moderators are playing a cat-and-mouse game. People who make videos that may violate the terms and conditions of a platform such as YouTube, for example, will only post an introductory video there. Viewers are then given a link directing them to a platform such as Bitchute, where other terms apply.

Attention economy

Users of online platforms can earn money by releasing content on the platforms. For example, video makers can make money per thousand views (CPM, cost per mille views or impressions) by offering increasingly extreme content (see also section 5.2). They can also recommend products or services (branded content) or ask viewers to donate money or to take out a subscription or membership. There are also indirect ways to earn money online. 'Like farming' involves creating eye-catching posts on Facebook that direct users to websites advertising products or services. The creators are not necessarily passionate about the content of their provocative posts. For example, Macedonian teenagers were found to be behind a striking amount of disinformation about the 2016 US presidential elections. They did not post this material because they favoured a particular candidate but because they wanted to generate advertising revenue on their own website (Rathenau Instituut, 2020b). Advertisers often do not know where their communications end up, so they may inadvertently be funding immoral or harmful behaviour (Stop Hate for Profit, 2021).

Since internet users tend to research products and services on only a limited number of websites, providers compete for attention there. This dynamic is described as the attention economy (Davenport & Beck, 2001). The idea is that providers' financial success depends heavily on their ability to attract and hold

people's attention, often by cleverly exploiting selection and amplification mechanisms. The theory is associated with other theories such as surveillance capitalism (Zuboff, 2019), which argues that platforms have a vested interest in knowing as much as possible about users so that they can optimise selection and amplification mechanisms and retain users' attention for as long as possible.

The result of this dynamic is that immoral and harmful content and behaviour are presented in ways that attract attention (Brady et al., 2017). This may explain the extreme forms that all sorts of immoral and harmful phenomena can take (Bishop, 2019). For example, an extreme prank video showing a child playing a computer game and being terrorised by sudden horror-film images has been viewed tens of millions of times (Hobbs & Grafe, 2015).

Ad sales based on user profiles often take the total watch time into account, i.e. how long a visitor to a website watches a video. Many studies have found a correlation between total watch time and mechanisms such as filter bubbles, echo chambers and radicalisation (Auxier & Vitak, 2019; Faddoul et al., 2020; O'Callaghan et al., 2015). A podcast series by *The New York Times* explains how this works (Roose, 2020); it bears the title 'Rabbit Hole' and recounts the radicalisation of a young American after YouTube's recommendation algorithm took him down a 'rabbit hole'.

To some extent, the operators ofonline platforms have an interest in preventing immoral and harmful behaviour and creating a safe and pleasant online environment, for example to placate advertisers or users. The interests of advertisers and users may conflict, however (Gabszewicz et al., 2001). When a platform's or operator's business model depends on advertising revenue, the advertisers' interests will carry a lot of weight, and that has consequences for the platform's content (Sanders, 2021, p. 61). For example, a platform may be inclined to claim more time and attention from users than is in users' best interests, and the calibre of the content may matter less than in a subscription model, where users pay for high-quality content.

4.6 Virtual reality

The online world is intangible and can therefore be perceived as unreal and artificial. Today, however, the internet is an important part of everyday life and actions online have far-reaching consequences in the physical world (Rathenau Instituut, 2020a). Confusion about what is real and what is not can lead to harm, for example when the standards used in fantasy games are applied to reality. The following mechanisms play a role in this context.

Dehumanisation
Is a threat aimed at @minpres on Twitter really a threat against the Prime Minister of the Netherlands or is it merely targeting a virtual Twitter account? The question is relevant because the morality of the matter depends on the target. Threatening a non-living object, such as an online account, is different from threatening a living human being. The confusing thing is that the internet can be very artificial and very personal at the same time. As a medium, the internet brings people together, but also largely strips them of human characteristics (De Vries, 2021). For example, online bullies do not usually see their victim, and in fact do not even need to know their victim. If they did, they might behave differently, i.e. in a more socially acceptable manner.

Entertainment
The entertainment value of the internet also plays a role in various phenomena. More than seven million Dutch people play games on a computer, tablet, smartphone or game console for an average of an hour a day (Multiscope, 2020). These games are often online. The internet is therefore a major source of entertainment. Whether or not something happens within an entertainment context is an important factor when interpreting behaviour. The statement ‘I’ll rip your head off’ means something quite different in a fighting game than it does on a social media platform.

The problem is that games and gaming platforms often have the same features as social media. Gaming platform Steam has the same profiles, friends and chat functions as Facebook. Some social media platforms, for example Discord and Twitch, were originally designed to serve gamers but are now also used for all kinds of other purposes. So it is understandable that game phenomena – the deliberate violation of norms and rules (trolling and griefing) but also innocent threats or expressions of violence – are visible beyond the gaming context. Immoral or harmful behaviour can also result from confusion about the gravity of the context, in other words.

Social media platforms are also designed to be highly entertaining. A platform like TikTok is full of humorous and entertaining videos, but there are also videos that push misleading information or hate speech (Weimann & Masri, 2020). Such content may have been created without any serious or malicious intent. For example, it appears that cyberbullying is often regarded by the perpetrators as a form of entertainment (Raskauskas & Stoltz, 2007), and entertainment is also an important motivation for the vandals who deface Wikipedia (Shachaf & Hara, 2010).

Image building
Information disseminated on the internet has a major impact on people's mental image of reality. Experts consulted for this study point out that in a growing number of professions it is require to have some kind of online presence, to build an online image or reputation. That is why so many people have a LinkedIn or Facebook profile that paints a favourable picture of their careers and lives. It is a way of being part of society.

This trend carries a number of risks, however. The importance of positive image building online may explain why users sometimes respond so sharply to criticism. A damaged reputation can have major consequences. Negative comments in the private domain can easily affect someone's professional life.

Young people no longer distinguish between their online and offline personas and do their best to package a socially desirable, glammed-up version of themselves online (Cocking & van den Hoven, 2018, p. 30). Hardly anyone is boring or unhappy on social media. Feeds are dominated by rose-coloured posts and stunning holiday snaps.

The 'app generation' is more self-focused than youth in decades past. Social media reinforce this tendency because they are organised around user's individual profiles (Gardner & Davis, 2013, pp. 69-71). Social scientists have observed a positive connection between narcissism and the likelihood of posting self-promoting content on social media (Gardner & Davis, 2013, p. 76). About 30% to 40% of ordinary conversation consists of people talking about themselves, whereas around 80% of social media updates are self-focused (Gardner & Davis, 2013, p. 76).

4.7 Disorder

The internet is sometimes referred to as cyberspace, as if it were an environment devoid of borders and national sovereignty. In reality, it is a domain in which parties from many different countries are active. A video originating in China may be on a server in Germany and can be viewed using software under a licence from a party in the United States, for example. This complexity makes it almost impossible to uphold law and order. The following characteristics and mechanisms play a role in this context.

Unclear norms
Many people are still confused about what constitutes 'civilised behaviour' online. How do you deal courteously with others on the internet, for example? Ever since people began interacting online, they have been seeking to identify the rules of

etiquette in cyberspace, the ‘netiquette’ for e-mails, for chats, for games, for forums and for other online environments (Shea, 1994). Every new interactive feature gives rise to the same socialisation process. People behave very differently in the virtual reality environment of VRChat, for example, than in the audio chat service Clubhouse. So they often find themselves in an online environment in which the applicable norms are unclear.

The lack of clarity about norms also means that many people do not know when and how they should call others to account for violations (Movisie, n.d.). Some experiments show that a rebuke can be effective, even if it comes from a bot (Machkovech, 2016) – but only if it is actually posted, of course. Without being corrected, it is difficult for people – and especially young people – to develop a good understanding of what constitutes appropriate behaviour.

Some platforms leave etiquette rule-making entirely up to the users, but others enforce it by technical means. For example, in a bid to prevent harassment, Facebook is actively considering how close social VR avatars in its forthcoming Horizon VR service will be allowed to get to one another (Rabkin, 2021). In many online environments, social norms are not only ambiguous but there is also a lack of guidance when it comes to users learning or monitoring norms. All this can be a source of immoral or harmful behaviour.

Anonymity

It is often unclear who we are communicating with online and whether the other party is human or robot (Christopherson, 2007). Someone can easily conceal their identity or assume the identity of another person. Often, it is very simple to register a free e-mail address and then set up accounts with all kinds of other service providers. Such anonymity is often used by malicious parties to violate norms with impunity. Internet platforms regularly report that they are removing tens or even hundreds of thousands of fake accounts in one go (Van Bemmel, 2020).

Even with sophisticated tools and methods, tracking down the identity of an internet user can be almost impossible. As a result, not only perpetrators but also victims may remain anonymous. The problem is that their anonymity is more likely to encourage immoral behaviour on the part of perpetrators than when victims are known (Yam & Reynolds, 2016). On the other hand, anonymous perpetrators consider the risk of harm to themselves to be negligent and are more likely to make unethical choices (Vince, 2018).

(Apparent) lawlessness

Surfing the internet takes users across a multitude of jurisdictions with the utmost of ease. Unlawful behaviour often goes unpunished because it is difficult to identify the

perpetrators – not only for reasons of anonymity, but also because tracking down and prosecuting them requires complex coordination between international intelligence and law enforcement agencies. Even if an offender can be identified, they may still turn out to reside in a country beyond the Netherlands' sphere of influence. In short, the international nature of the internet complicates law enforcement. For victims of immoral or harmful behaviour, the absence of repercussions translates into powerlessness (see Online Shaming case).

The apparent lawlessness of the internet is also facilitated by commercial parties. Some parties even advertise their willingness not to cooperate with law enforcement. There are also internet services technically designed to ensure that no single party can be held responsible, for example products or services based on blockchain or distributed ledger technology (DLT).[5]

Isolation

Surfing the internet is often a solitary activity. Many people have their own smartphone or computer, and use the internet independently, in isolation from others. This means that there is also limited oversight and non-existent guardianship by parents or others (Peterson & Densley, 2017).

The absence of monitoring and corrective measures can lead to problematic behaviour, as morality has a significant social dimension (Ellemers et al., 2019). This may also be the reason why some people behave so differently online and offline.

5 DLT is a catch-all term for systems in which multiple parties operate in a digital environment that has no central authority or operator. Blockchain is one example. It uses a data structure consisting of a chain of hash-linked blocks of data.

Case: Disturbed eating behaviour

The case below is fictitious and intended to illustrate possible risks that may arise from online challenges that induce disturbed eating behaviour. The case is, however, based in part on a combination of incidents that occurred in the Netherlands and abroad.

Case
During the lockdown, teenager Sam spends days alone in her room with her laptop and smartphone as her only distractions. It all starts when an acquaintance, Kim, posts a challenge on TikTok. Kim is someone Sam remembers from the campsite where she holidayed last summer. Several times a day, Kim posts the number of calories she consumes. She also regularly posts videos of her meals. Sometimes the meal consists merely of a bowl of ice cubes. Kim encourages others to join her extreme challenge and surpass her by consuming even fewer calories a day. Samantha gets hooked on Kim's endless stream of updates. She's bored and thinks it would be fun to track her own calorie intake, so she decides to join in. She is gratified to see her posts getting immediate 'likes' and positive comments. She enjoys interacting with some of her new followers, and that helps her feel less lonely. At least they understand her.

Samantha reads a newspaper article about pro-ana communities and goes online to search for the names and websites it mentions. Pro-ana websites are not banned, so there are plenty for her to explore. Samantha doesn't dare talk about her new obsession offline with her schoolmates or her parents, but online she can assume a new identity and be open about it. That is a relief to her. Samantha creates a new e-mail address under an alias, Rox, and then uses it to set up a sock puppet (a fake account) on TikTok and Instragram. She enters a fake date of birth and a profile picture of a stranger who does not look anything like her. On search engines and social media platforms, she enters misspelled search terms and hashtags, for example anoreixa, anorexiia and annorexia, to circumvent their blockades. It takes her only a few clicks to find a pro-ana group that posts photos more extreme than she has ever seen. Samantha joins a pro-ana community's WhatsApp group, which has a constant stream of messages about weight loss.

She joins a group whose accounts are occasionally removed. They immediately reappear elsewhere and quickly regain their 500+ following, since the members all stay in touch on WhatsApp. The group does not have a moderator, or members who do not have an eating disorder, or adults who would speak up. Some of the girls post photos of their hospital admissions and are very candid about their

experiences. Samantha admires them and would like to watch all their YouTube videos. The content recommendation systems quickly learn which kinds of videos Samantha watches to the end and which images she 'likes'. They recommend more of the same. Recovery accounts by people in her network who have overcome anorexia gradually disappear.

Advertisers promoting weight-loss products are banned on TikTok, but Sam doesn't need banners to find quacks selling laxatives. She receives an expensive watch with a pedometer for her birthday and buys weight-loss apps with calorie counters for her smartphone. She shares her steps and calories on social media every day. She has the backing of a growing number of followers and even earns a bit of money from her posts. Sam reaches her target weight and then goes even lower. With so many followers, she can't simply quit now. No one in her group tells her to stop.

Sam becomes ill, not just physically but also psychologically. Her parents try to get help for her, but there is a waiting list. It is only weeks later, after she faints several times at school, that she is admitted to a clinic. She is allowed to take her smartphone with her. No one there asks her which apps she has on her phone, who she follows and whether she is in touch with people whom she doesn't know offline. She forgets everything her counsellors tell her during the day as soon as she scrolls through her trusted WhatsApp group, Instagram and TikTok feeds. She never really sees images of people who are not super-thin anymore. She finds it impossible to quit the online pro-ana group. They are the only peers she still interacts with. She rarely hears from the few school friends she had, the only ones she had any contact with offline.

Reflection

We see several mechanisms operating simultaneously in this scenario. In this example, they are the physical and emotional **isolation** of a potential victim combined with online **anonymity** and the **continuity** of technology in the private domain. **Syndication** and **echo chambers** also play an important role. A small group of potential victims can cluster easily online and blot out any alternative opinions. Echo chambers, **selectivity** and syndication are facilitated by content recommendation algorithms set up to boost minutes of viewing time, for example. In the **attention economy**, viewing time is one of the benchmarks used by companies that advertise in traditional media and online platforms. Individual users are also rewarded for content that generates large numbers of followers or subscribers. As a result, they seek **amplification** and **virality**, mechanisms best served by sensational content and entertainment. Such quests may encourage minors to engage in harmful behaviour, such as extreme dieting.

There are several stakeholders in this case. Some play a highly active role, such as the adolescent users and followers who urge each other to engage in disturbed eating behaviour. They do this in an environment without guidance and supervision from parents or other adults and without counsellors or moderators. Parents, teachers and school authorities, friends and counsellors in the offline environment play a passive role in this example when it comes to encouraging offline contact and providing help and guidance in dealing with online mechanisms. Finally, quacks and suppliers of weight-loss products and services, online platforms and the traditional media also play significant roles. By focusing on conversion, clicks, viewing time and the collection of individual data profiles, they encourage harmful behaviour in this case.

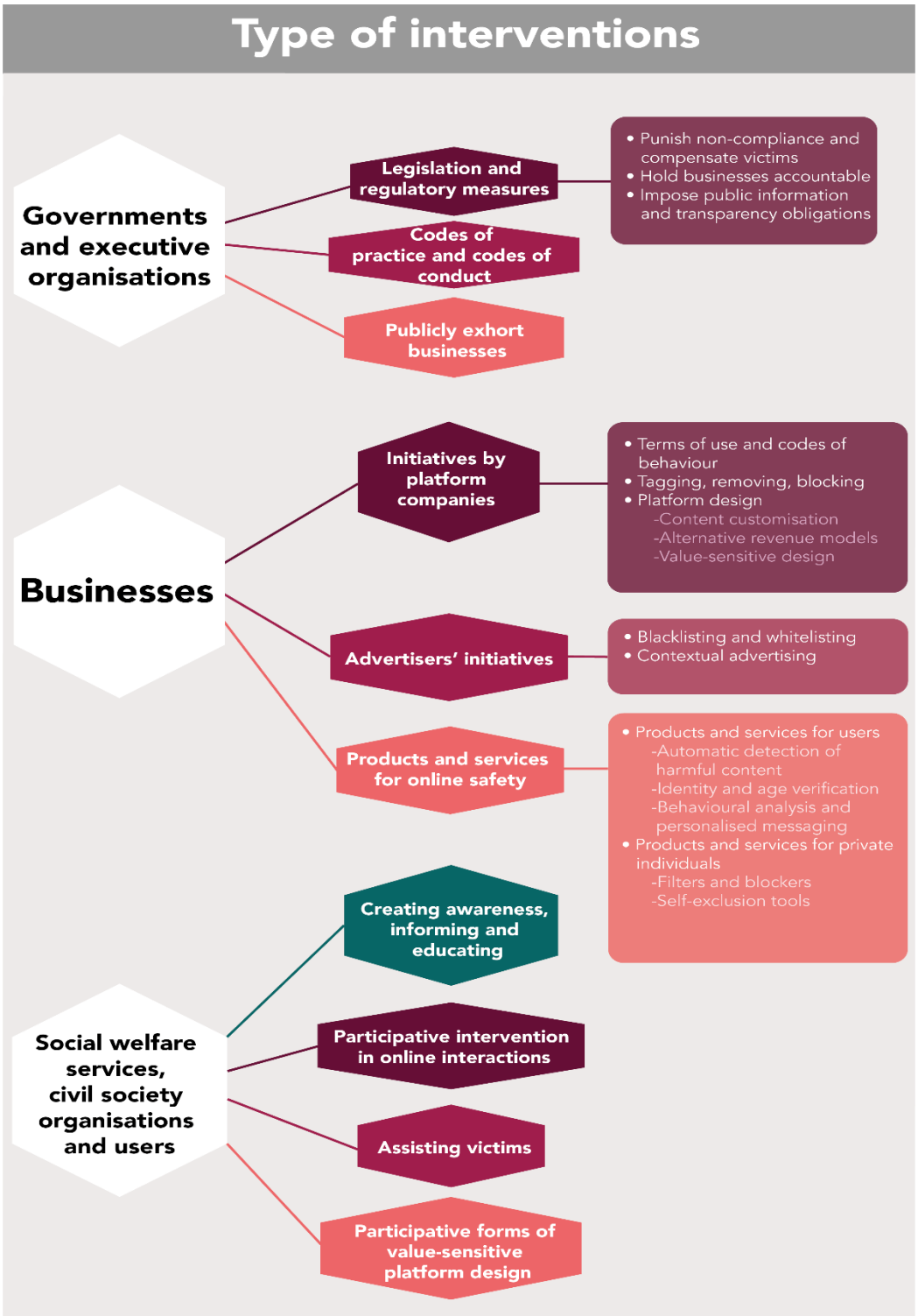
5 Current approach to harmful and immoral behaviour online

Various steps have been taken in recent years to reduce or even prevent harmful and immoral behaviour online. In this chapter we take stock of the existing approach to harmful behaviour by considering the *actors* involved and the *types of interventions* available. Our overview of existing measures sheds light on which interventions work and which show promise. But there are shortcomings in the current approach, and so there is room for additional interventions, inspiring us to present a strategic agenda in Chapter 6.

In the present chapter, we differentiate three groups of stakeholders: governments and their executive agencies; businesses (including platform companies, but also providers of other products and services); and finally, a broad category of social welfare services, civil society organisations and internet users. For each of these groups, we identify the main types of intervention, giving examples (see Figure 5), and then discuss the main lessons that can be learned from each initiative.

5.1 Governments and executive agencies

We begin our overview by looking at initiatives that governments and executive agencies have so far undertaken to counter or prevent harmful and immoral behaviour online. We first discuss legislative and regulatory measures. We then highlight two other alternatives for government control: 1) enter into more voluntary agreements with commercial parties, i.e. codes of practice and codes of conduct, and 2) publicly exhort businesses to take action against problematic behaviour.



Bron: Rathenau Instituut

Figure 5 Taxonomy of interventions

Legislation and regulatory measures

When it comes to regulatory matters, national initiatives have dominated in the EU so far. Several countries have adopted or prepared legislation in recent years to help reduce problematic online behaviour and subsequent harm. Relevant EU legislation is also in the works. The *Digital Services Act* (DSA) and *Digital Markets Act* (DMA) proposed by the European Commission in 2020 should help to combat illegal content and disinformation and rein in large platforms.
In a discussion paper on how to tackle online harm and manipulation, the Behavioural Insights Team, a group of behavioural scientists that advises the UK government, states that lawmakers traditionally have three tools in their toolkit (Costa & Halpern, 2019). First, governments can punish non-compliance and poor behaviour. Second, governments can encourage businesses that own or operate online environments where the behaviour occurs to surpass minimum standards. It can do this by imposing certain responsibilities and obligations on them. Third, governments can encourage businesses to educate consumers or users by exhorting them to be more transparent about what they do. We discuss these options below.

Punish non-compliance and compensate victims
Some of the phenomena discussed in the present report are already criminalised under current legislation. In many cases, these laws were designed to address the 'offline' versions of the relevant behaviour and therefore also apply to online behaviour. For example, many EU member states cover hate speech and hate crimes in criminal codes that also apply to online behaviour (see e.g. Policy Department for Citizens' Rights and Constitutional Affairs, 2020).

In addition, governments may choose to criminalise online behaviour. Many do so with respect to specific phenomena, such as cyberterrorism or child sexual abuse (facilitated by the internet). For example, the Netherlands recently criminalised the unauthorised creation or distribution of sexual images of others, as in revenge porn or sextortion (Ministerie van Justitie en Veiligheid, 2020). Outgoing Justice Minister Grapperhaus wants to extend this to doxing, i.e. sharing people's private data on social media (Bakker, 2021). There is also an EU initiative to improve the prosecution of hate crimes under criminal law (Policy Department for Citizens' Rights and Constitutional Affairs, 2020).

It is also possible to litigate against forms of harmful behaviour online under private law. The point of such civil proceedings is not so much to punish the perpetrators as to hold a party responsible for the harm suffered by the victim. They can help in terms of acknowledging or compensating for the harm done (e.g. damages) or rehabilitating the victim's reputation (e.g. an apology or a rectification). Bureau

Clara Wichmann, a foundation dedicated to women's position under the law, investigated what legal options are available to women to address online hate speech (Bureau Clara Wichmann, 2020).

One concern, both in applying existing laws and designing new ones, is that they are often difficult to enforce. First of all, the democratic rule of law has a weak online presence (Advisory Council on International Relations/AIV, 2020; see also Bantema et al., 2018) and law enforcement agencies often lack the knowledge and technical or other resources to operate efficiently there (e.g. Politie et al., 2020). Second, the legal instruments and procedures available in the offline world are not always suitable for enforcing such laws and regulations. This a problem for the police and the Public Prosecution Service, but also for victims, who face all sorts of obstacles if they want illegal, harmful content removed from a platform, for example (Advisory Council on International Affairs/AIV, 2020; IVIR, 2020). And third, enforcement is tricky because the laws and regulations cover a very wide range of services (IVIR, 2020) provided by businesses that often operate internationally, whereas there is no competent cross-border jurisdiction (Advisory Council on International Affairs/AIV, 2020; Aiken, 2016). Legal experts therefore advocate shifting the focus from national to international regulation (Policy Department for Citizens' Rights and Constitutional Affairs, 2020).

Hold businesses accountable

The internet and internet-related activities have always been largely unregulated. That appears to be changing. Governments appear to be increasingly aware that the properties of the online environment and the parties operating in it are instrumental in inspiring or catalysing harmful behaviour.

The EU's Directive on electronic commerce (2000), which is still in force but will eventually be replaced by the *Digital Services Act* (DSA), basically exempts social media platforms and internet access and web hosting providers from liability for the content uploaded by their users, provided that they act as a 'mere conduit' (European Parliament & Council, 2000). They need only remove unlawful content if they are alerted to its presence (notice-and-take-down procedures). Several European countries have introduced additional rules in recent years, either to force businesses to comply more fully with this obligation or to make them liable for the harmful behaviour occurring through their channels. In many cases, the underlying rationale is that self-regulation has not been sufficiently effective (see e.g. UK Government, 2019).

Below is a table comparing three key national initiatives in Europe: the German *Network Enforcement Act* (2017), the French *Loi Avia* (2020) and the British *Online*

Safety Bill.[6] Although they differ in some respects (e.g. in the type of content they cover), what these laws have in common is that they hold businesses responsible for content removal, complaints procedures, reporting and suchlike. At the bottom of the table we list the most relevant items for the purposes of this study from the *Digital Services Act* as proposed by the European Commission last December. The Act will eventually become effective in all the European Union member states. In the meantime, member states, including the Netherlands, and the European Parliament can still influence the substance of the proposal.

Comments on these initiatives show that the decision to regulate content is a hard one to make. After all, it involves trade-offs between various freedoms and fundamental rights: on the one hand, freedom of expression, the right of access to information, freedom of the press or freedom to engage in business; on the other, the right to personal integrity and safety and a host of democratic principles and principles under the rule of law. Undue friction between these rights and freedoms can cause initiatives to fail. That is what happened in France, for example, where the Constitutional Court struck down certain provisions of the *Loi Avia* that had caused indignation among legal experts and civil society organisations after its adoption by the National Assembly. The law was then watered down significantly (Vie publique, 2020), the argument being that it encroached too much on freedom of expression.

On the following pages:
Table 2 Comparing legislative initiatives in Europe

6 *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG)* and *LOI n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet*, respectively. The UK's Online Safety Bill was published in May 2021 after its underlying principles were explained in two versions of the *Online Harms White Paper* (UK Government 2019 and UK Government 2020). It is also worth mentioning Ireland's *General Scheme for an Online Safety and Media Regulation Bill*, which was in the pre-legislative phase at time of this research (Government of Ireland 2021). The UK's *Online Safety Bill* is likely to be the most ambitious of all the national initiatives, in terms of both content type and scope.

Law	Businesses in scope	Type of content	Obligations (etc.)	Sanctions
Network Enforcement Act (D)	Social networking sites	Unlawful content	- Notice and take-down: duty to remove content or disable access (within a set time frame) after notice filed - Duty to establish accessible and efficient procedures for notice and feedback - Reporting obligation for businesses after a specific number of notices have been filed	- Fines (amount depends on type of infringement and size/reach of business or site)
Loi Avia (F, original version[7])	Web hosting providers	Some unlawful content: hate content, terrorist materials and child pornography	- Duty to remove content (within a set time frame), even without notice - Duty to establish accessible and efficient procedures for notice and feedback	- Fines (+ imprisonment for perpetrators)
Online Safety Bill (UK, draft bill)	All providers of services that host user-generated content; businesses that facilitate public and private online interaction; search engines	Illegal content; content harmful to children; content that is legal but harmful to adults[8]	Statutory duty of care, enshrined in code of practice enforceable by regulator. According to the code, providers of such services have a - duty to take steps to prevent harm to users, for example by removing certain content (within a set time frame) - duty to establish accessible and efficient procedures for notice and feedback - duty to establish clear appeal procedures - reporting duty - duty to protect journalistic content, applicable to providers of Category 1 services (high-risk, high reach)	- Fines (amount depends on type of infringement and size/reach of business or site) - Business disruption measures (e.g. interrupting or disabling services) or criminal action brought against senior managers

7 The table shows the substance of the law as approved by the National Assembly on 13 May 2020. Some time later, it was watered down considerably, after the Constitutional Court struck down certain provisions as unconstitutional (see main text).

8 Frequently cited examples from the 'legal but harmful to adults' category are content promoting self-harm, hate content, online abuse (insofar as not criminalised) or 'content encouraging or promoting eating disorders' (UK Government 2020).

Law	Businesses in scope	Type of content	Obligations (etc.)	Sanctions
Digital Services Act (EU, in preparation)	Web hosting providers, including online platforms; providers of infrastructure intermediary services (e.g., internet access providers, cloud hosting services)	Exclusively illegal content	Depending on the type of service: - statutory notice-and-action obligation; disabling access in the case of repeated infringements - duty to establish accessible and efficient procedures for notice and feedback; including prioritising notices by 'trusted flaggers' (experts in tackling illegal content) - duty to establish clear appeal procedures - reporting duty	Fines and more 'structural' measures (such as the duty to divest parts of the business)

Some experts suggest solving this problem by regulating platform mechanisms instead of content, i.e. the algorithmic principles for ranking messages that ensure that certain content is recommended more than others (Pomerantsev, 2019; Rathenau Instituut, 2021a). This would avoid overly restricting the right to free speech while still containing the reach of harmful utterances (also called freedom of reach, Diresta, 2018). It would mean pruning back an important mechanism behind harmful and immoral behaviour.

Another criticism is that the laws and directives referred to above place considerable decision-making power in the hands of businesses. The risk is that platforms, fearing fines, will remove content proactively and thus engage in a form of censorship (Index on Censorship, 2019). In the case of laws that also cover 'harmful' content (as opposed to only illegal), vagueness of terms and cultural differences may reinforce this risk (Advisory Council on International Affairs/AIV, 2020; Pomerantsev, 2019). In response to such concerns, the United Kingdom's proposed *Online Safety Bill* now includes obligations to safeguard freedom of expression and to protect content relevant to the democratic process as part of the duty of care for businesses. This has not yet reassured critics, however (Hern, 2021). There are no ready solutions to these problems, then, but it is in any event important to monitor content-removal policies. Rathenau Instituut therefore proposed that the Dutch House of Representatives, in the run-up to the *Digital Services Act*, should press for a strong oversight structure and independent public oversight of content moderation (Rathenau Instituut, 2021b).

In response to criticism that the concept of 'harmful content' is unclear, the UK government has decided to define it more precisely in its bill (UK Government,

2020b). But that is not a perfect solution either. After all, one can argue that in a democratic state, such definitions should be the subject of an inclusive social dialogue (cf. e.g. Helberger et al., 2018). Moreover, such methods could lead to an abuse of power, especially if less free states start using them as well. An alternative is to design laws and regulations in such a way that control over content is placed more firmly in the hands of users. They could, for example, be allowed to choose or weight the criteria by which things are ranked or presented, control the level, type or source of advertising, or control how their data are used (Costa & Halpern, 2019). We will discuss these alternatives in more detail later (see section 5.2).

Impose public information and transparency obligations

Some of the laws and directives cited above impose public information and transparency obligations on businesses in addition to their responsibility for content. For example, the UK's *Online Safety Bill* seeks to enforce openness about the prevalence of harmful behaviour online by obliging the industry to make 'transparency reports' with statistical data available. This should help the UK government and its executive agencies get a better handle on the scale and nature of harmful behaviour online. In addition, businesses must provide information explaining what steps they are taking to counter such behaviour and to keep users safe (UK Government, 2020b).

Another form of transparency is that envisaged by the EU's *Digital Services Act*. The EU intends to use the Act to force large platform companies (companies that have a 'gatekeeper function') to disclose how their recommendation systems work, and more generally, the role that data and artificial intelligence play in the services that they offer. The underlying idea is that businesses themselves, but also researchers, will then be able to evaluate, and thus anticipate, the social impact of using specific systems at an earlier stage (Tokmetzis & Bol, 2020). Large platform companies will probably also have to perform risk analysis for this purpose.[9]

Those who criticise regulatory measures that place decision-making firmly in the hands of businesses – for example civil society organisations that defend civil liberties – are generally more inclined to support transparency obligations (see, for example, Pomerantsev 2019). Rathenau Instituut has emphasised that transparency is a multi-faceted matter and that, in the context of regulation, it must be clear exactly what platforms are required to answer for and what factors possible regulators will assess (Rathenau Instituut, 2021a).

9 Transparency obligations are also being introduced at the national level, albeit usually within the context of regulating specific phenomena. One Dutch example is a legislative initiative addressing micro-targeting in political advertising campaigns (see Rathenau Instituut, 2020).

Codes of practice and codes of behaviour

In the aforementioned discussion paper on tackling online harm, the UK Behavioural Insights Team notes that, in addition to the three regulatory tools discussed, governments are increasingly exploring other tools for countering harmful behaviour online (Costa & Halpern, 2019). These are usually voluntary agreements negotiated with businesses in the sector itself. One type of agreement that governments have initiated in recent years consists of codes of practice and codes of conduct. They come in many varieties: national and international, involving different types of businesses, and focusing on different phenomena and types of content. We discuss a few examples below.

A well-known code of conduct in the European Union is the one against online hate speech, which was launched in 2016.[10] The European Commission agreed with Facebook, Microsoft, Twitter and YouTube that they would follow up within 24 hours on at least half of all notifications of content inciting hatred or violence against persons or groups defined by reference to race, colour, religion, descent, or national or ethnic origin and remove any illegal content. The code complemented an existing European decision on racism and xenophobia.[11] Two years later saw the adoption of the Code of Practice on Disinformation (Stolton, 2020), signed not only by large tech companies, but also by industry bodies in the advertising sector and by advertisers themselves. They agreed to provide more transparency about political advertising, close fake accounts, work with fact-checkers and improve the visibility of information that has been fact-checked (Advisory Council on International Affairs/AIV, 2020).

Most national codes of practice and codes of conduct deal specifically with illegal content. For example, pending the introduction of the *Online Safety Bill*, the United Kingdom has a code of practice on terrorism and child sexual abuse (UK Government, 2020a). In the Netherlands, the internet sector has drawn up a code of conduct at the request of the government that should encourage web hosting companies to be quicker about taking down child pornography or other prohibited material after reports of such occurrences (noticeandtakedowncode.nl/, 2018). Most codes of conduct do not address the mechanisms underlying the problematic behaviour, for example the recommendation algorithms.

Government initiatives that rely heavily on the willingness of businesses to self-regulate receive a mixed reception. While many commentators believe that businesses should take responsibility for the behaviour they facilitate, they question

10 In full: *EU Code of conduct on countering illegal hate speech online*.
11 *Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law*.

the effectiveness and reliability of voluntary codes of practice and codes of conduct. The *EU Code of conduct on countering illegal hate speech online*, for example, is seen as ineffective, not least due to its voluntary nature (Rathenau Instituut, 2020b; Stolton, 2020). There is also concern that platform companies are granting themselves undue discretionary power in such codes at the expense of users' freedoms and fundamental rights (Advisory Council on International Affairs/AIV, 2020). Co-regulation, whereby independent regulators assess platform companies' efforts and, in the worst case, impose coercive measures, is therefore seen as more promising (Rathenau Instituut, 2020b).

The Advisory Council on International Affairs (AIV) observes that countries each follow 'their own path' when it comes to making regulatory choices, depending on prevailing views on national security or commercial and individual liberty (Advisory Council on International Affairs/AIV, 2020, p. 6). In the past, the Netherlands' policy has focused on minimal regulation and a free internet market, largely in private hands. Where regulation is needed, the government has traditionally emphasised self-regulation by the tech sector itself (Advisory Council on International Affairs/AIV, 2020). The Council now considers that the time has come to recalibrate this policy; the Rathenau Instituut has endorsed its advice (Rathenau Instituut, 2021a).

Publicly exhort businesses

Governments and executive agencies can also use more informal means to encourage companies to reduce harmful and immoral behaviour online. These can be 'positive' means, by urging them to act responsibly. For example, authorities can invest in developing the technical facilities needed to reduce harmful behaviour. We will discuss this in more detail in the next section. But they can also use 'negative' means, by condemning a failure to act.

A tried and tested strategy is to publicly exhort businesses that are not doing enough to discourage problematic behaviour or to combat relevant content. Following the suicide of a teenager allegedly provoked by the availability of online images of self-harm, for example, the UK's Health Secretary admonished Facebook and Instagram for failing to shield children from such material. The idea behind such 'targeted exhortation' may be that businesses, fearing damage to their reputation, will do more to address the problem raised. Ideally, they would also set an example for other businesses. In the UK case, Instagram immediately

responded to the pressure by admitting it 'was not where we need to be' and committing to taking further action (Costa & Halpern, 2019).[12]

Such examples are rare in the Netherlands. Dutch MPs occasionally ask a minister critical questions, for example about the role of social media in spreading disinformation or about platform companies' lack of transparency about their underlying mechanisms (Facebook, 2021) (Tweede Kamer, 2020). Ministers sometimes threaten platforms with targeted exhortation. For example, last year outgoing Justice Minister Grapperhaus threatened to release a list of web hosting companies that take too little action against the distribution of child pornographic material if they failed to comply with the Dutch notice and take-down code (Houtekamer & Wassens, 2021).

5.2 Businesses

Beyond government intervention, businesses have also made efforts to reduce or even prevent harmful and immoral behaviour online. Large social networking platforms are in the spotlight because their influence puts them under social and political pressure to take action. Other, smaller parties are also taking steps, however. Below, we discuss initiatives by three such stakeholders: platforms that facilitate online interactions, advertisers and their intermediaries, and producers or providers of innovative products or services designed to reduce immoral and harmful behaviour online. In the latter category, we distinguish between products and services aimed at businesses (e.g. web hosting services) and those aimed at individuals.

Initiatives by platform companies

There are roughly three categories of initiatives that platform companies are introducing to reduce or prevent problematic behaviour online. First, they try to encourage desirable behaviour by laying down rules for engaging in online interactions, for example in the form of terms of use or codes of conduct. Second, they take steps to suppress immoral and harmful behaviour, for example by removing or blocking certain content or by sanctioning perpetrators. And third, they make adjustments to their platform design so that users are better protected against harm.

12 'Targeted exhortation' is thus based on the same principle as the phenomenon of shaming, but deploys it for a different purpose. It uses platforms' sensitivity to the power of online image-building.

Terms of use and codes of behaviour

Terms of use establish the rules that users are expected to abide by when accessing a platform. Generally, they address the type of content that may be posted and/or establish guidelines for interacting with other users. Most platforms in any case prohibit the posting of illegal content – even the channels that attract fans of 'extreme' content. They differ, however, in the content or behaviour covered by their terms of use and whether they only discourage or actually prohibit certain behaviour (see Table 3). There are also major differences in the way that they describe potentially harmful content or actions (e.g. in the degree of specificity or detail). The language or tone of the various codes of conduct also varies widely, depending on the intended audience.

In response to social or political pressure, large social networking platforms have tightened up their terms of use, most recently owing to controversies surrounding disinformation (Tumber & Waisbord, 2021). In doing so, they usually define the boundaries between what users are and are not allowed to do, and users effectively agree to this, either explicitly (by agreeing to the terms of use and related definitions when using the service for the first time) or more implicitly (e.g. after the terms are amended). Terms of use define platforms' policies on blocking or removing content.

Terms of use and codes of conduct are generally considered ineffective tools for managing online behaviour. A well-known problem is that users usually simply agree to them without reading them through. And there is good reason for this: the texts are usually not very user-friendly (Costa & Halpern, 2019; UK Government, 2019). There are also recurring weaknesses. For example, many platforms do not yet have robust policies on self-harm and suicide, even though online platforms are increasingly where people search, discuss and seek support for mental health issues (Newton, 2021b).

On the following pages:
Table 3 Platforms' terms of use

Example	Description	Prohibits…	Limits…	Discourages…
Facebook (Facebook, 2021)	Social networking site Broad public More than 2.8 billion active users a month	- Illegal content/ behaviour - Potentially harmful content (e.g. hate speech, violent or sexual content) and the facilitation, organisation or promotion of harmful behaviour		
Twitter (Twitter, 2021)	Microblogging site Broad public More than 330 million active users a month	- Illegal content/ behaviour - Potentially harmful content and promotion of harmful behaviour (e.g. violence and bullying; suicide and self-harm) and 'sensitive' content (e.g. gruesome or sexually explicit material)		
4chan (4chan, 2021)	Discussion forum (for images) Young users, generally anonymous More than 20 million active users a month	- Illegal content/ behaviour - Soliciting or disseminating personal information (doxing) or inciting attacks (raids) - Complaints about 4chan	- Certain potentially harmful content is only permitted on specific channels (e.g. troll posts, racist remarks, pornographic content)	- Spamming or incomprehensible content - Attacking other users (including verbally)
Parler (Parler, 2021)	Microblogging site Appeals to fans of freedom of speech; known for members holding right-wing and ultra-right-wing views Reach varies greatly (recently several million active users a month)	- Illegal content and threatening to commit illegal acts	- Certain content must be tagged: potentially harmful (e.g. violent) or 'sensitive' material (e.g. with nudity)	- Spamming

Reddit (Reddit, 2021a)	Social networking site and discussion platform with 430 million active users a month Consists of more than 100,000 active 'subreddits' focusing on specific interests, from politics to coffee connoisseurs to runners Subreddits can set their own additional rules and have their own moderators	- Illegal content/ behaviour - Harmful content (bullying, violence, hate speech and discrimination) - Manipulation of information (including influencing the electoral process) - Doxing, revenge porn - Sock puppeting (but users may remain anonymous)	- Sexually explicit content must be tagged as such and will not appear in Reddit's general timeline of popular posts	Extreme subreddits that may be harmful are sometimes 'quarantined' by Reddit. They are then hard to find and cannot be viewed by non-subscribers. Examples are communities dedicated to 9/11 conspiracies and pro-ana subreddits.
TikTok (Tik Tok, 2020)	Social networking site meant for sharing short videos; 1.7 million users in the Netherlands Mainly popular with children and adolescents	- Illegal content - Violent extremism, hate speech, suicide, self-harm and dangerous behaviour - Harassment and bullying - Nudity in any form - Grooming, child abuse - Misinformation and sock puppeting		

Tagging, removing, blocking

A more repressive form of intervention is content moderation, i.e. tagging, and sometimes removing, illegal or otherwise harmful material. Some platforms, such as Reddit, have their users voluntarily monitor compliance within the terms of use (Reddit, 2021b). Others, such as Facebook or Twitter, hire professional moderators or use sophisticated automated technology to monitor content (Facebook, 2019). Large platforms in particular call on fact-checkers to combat disinformation. Recent events, for example the Covid-19 pandemic, have lent urgency to this practice. Facebook and YouTube, for example, have tagged or deleted millions of dubious posts in the spring of 2021 (Griffin, 2021; Wagner, 2020). Nowadays, technical tools are also used to partially automate fact-checking (Rathenau Instituut, 2020a).

Content moderation combined with fact-checking has already produced some encouraging results. Facebook's own research shows, for example, that 95% of visitors who saw warning messages on unreliable Covid-19 reports did not click through to the original content (Zuckerberg, 2020). There is a problem, however: moderation by human moderators is difficult to scale up, but algorithmic detection and other technical tools are less reliable than humans and may reproduce or even amplify user bias. What is in any case true is that the criteria applied in content moderation are highly context-specific. What an American company considers harmful is not necessarily regarded as such elsewhere in the world, and vice versa. Knowledge of the local culture is therefore required to moderate content properly. Researchers further point out that traditional media – print and TV – also play a role in encouraging problematic behaviour simply by publicising it (Kaiser et al., 2020). Moderation must therefore always be combined with other measures.

In extreme cases, platforms can also sanction users. For example, they can suspend users by terminating their account (deplatforming) or block access from a certain IP address (blacklisting). Google also blocks parties that are in flagrant violation of its terms of use from its ads network, for example conspiracy-themed websites (Kist, 2020).

To identify perpetrators, businesses sometimes work with developers of specialist technology. Crisp, for example, is an American firm that uses artificial intelligence to track the relationships between various platform users. Based on its findings, it estimates which contacts could be harmful (Crisp Thinking, 2021). Platforms are deploying this type of software in the fight against online child abuse (UK Government, 2019). We look at this more closely in section 5.2.

One of the concerns about content removal, blacklisting and other platform-driven interventions is that they concentrate much of the decision-making power into the platforms' hands. The terms of use that underpin such decisions are rarely grounded in national or international law. Moreover, platforms use unclear definitions and users have little opportunity to object to content removal decisions (Advisory Council on International Affairs/AIV, 2020). In this respect, too, platforms affect people's rights and freedoms. Critics feel that they too often take the place of the courts (e.g. Bureau Clara Wichmann, 2020).

One way in which platforms can mitigate these risks is by working with civil society organisations to define their policy, or by setting up advisory boards or industry regulators for this purpose. One well-known example is the Oversight Board established and funded by Facebook. The Oversight Board is meant to safeguard the rights of users and, in particular, must ensure that Facebook and Instagram respect users' freedom of expression. The Board considers appeals against

decisions on content, such as removal decisions, and its rulings are binding, even if it disagrees with a decision taken by Facebook or Instagram. It also plays an advisory role and makes recommendations concerning the content policy of both platforms.

Commentators have mixed feelings about such oversight boards. On the one hand, it is a good that platforms – which sometimes serve conflicting interests (e.g. those of advertisers and users, or of an authoritarian government and its citizens) – delegate certain decisions to them. On the other hand, some board decisions have met with criticism. One example is the decision of Facebook's Oversight Board to uphold the removal of (then) President Trump from the platform (Paul, 2021). For media and governance researchers, these cases mainly show that the companies behind social media platforms are not regulated strictly enough (e.g. MacCarthy, 2021). Others stress that advisory and oversight boards can only do their job properly if they are transparent about what is done with their input. In reality, such transparency is often lacking (e.g. Helberger et al., 2018; Sánchez Montañés, 2021).

Finally, deplatforming and blacklisting also raise concerns among the experts we consulted, who believe that problematic behaviour is shifting away from the larger platforms with their strict codes of conduct to smaller, alternative platforms where users have more freedom but can also more easily evade social control. Users there are even less likely to hear voices that disagree with the content they seek out or disseminate. The lack of pushback means that mechanisms such as syndication are reinforced and, in turn, feed the problematic behaviour. In addition, there is the risk of social fragmentation: extreme points of view and behaviour live on, but out of the sight of the vast majority.

Platform design

A very different strategy that platforms can deploy is to make design choices that reduce or help prevent immoral or harmful behaviour or the victimisation that results from it. We identify three types of initiatives. The first is the facilitation of content customisation, in which users can themselves choose what content to view, how to do so and when. The second is an alternative revenue model (including an advert-free one). And the third consists of various forms of value-sensitive platform design whereby new platforms are designed to deter mechanisms that promote harmful and immoral behaviour.

Content customisation

Content customisation includes allowing users to control how their data is collected or shared, to exercise more control over the criteria for selecting, ranking or presenting posts on their timelines, or to control the amount or type of advertising

they see (Costa & Halpern, 2019). The assumption is that these kinds of choices can help users arm themselves against immoral or harmful behaviour, or that they can help break down underlying mechanisms such as selectivity and amplification.

Large platform companies are not yet experimenting with this type of solution much. Since they operate in an arena in which both their users and they themselves benefit financially or otherwise from attracting (a lot of) attention (see Chapter 4), their lack of interest is hardly surprising. Nevertheless, initiatives are slowly getting off the ground, in part, perhaps, under public pressure. For example, Facebook wants to give its users more control over how posts are ranked in their News Feeds (Benton, 2021; Newton, 2021a). Twitter is even thinking of a building an app store for social media algorithms where users can choose which ranking algorithms are applied in multiple social networks (Kastrenakes, 2021). In all cases, however, it remains to be seen whether users will know enough about how these systems work to exercise informed control over them (Newton, 2021a).

Commentators believe that such interventions will remain the exception if businesses are not subjected to stricter regulation at the same time. The government could also demand that platforms introduce more content customisation, or that they make existing customisation features easier to find (Costa & Halpern, 2019). In addition, it could do more to create a favourable development climate for new intermediaries, such as developers of innovative services (*ibid*.), whether the platform companies themselves or non-profit organisations. Okuna, for example, is an alternative social network launched in the Netherlands that does not monitor users' activity but leaves it entirely up to them to decide what they see in their news feed, and to some extent how (Okuna, 2021). Such products make it possible to give users more control over their online social networking experience, without the need for regulation.

Alternative revenue models

Okuna is a crowdfunded initiative and based in part on a subscription model, with users paying for extra features (Okuna, 2021). The platform generates the revenue it needs to operate in this way without having to rely on advertising income. This may also be beneficial for the social interactions that take place there. After all, the way in which the online advertising market is currently organised follows the logic of the attention economy, which creates a breeding ground for harmful behaviour. In an advert-free model, reach is in fact less important. This sort of model also appeals to users who want high-quality content, and those willing to pay for quality also tend to behave more decently and politely.

Large tech companies are also starting to see the benefit of subscription models. One example is Facebook's idea of offering fan subscriptions (Ha, 2020) or Apple's

plans for a paid podcast service (Kafka, 2021). A ban on microtargeting (a form of advertising that targets specific groups using data on individual users) could encourage this type of initiative. A political debate is currently underway in Europe about such a ban (see, for example, Vinocur, 2021) and Alphabet (the company behind Google) seems to be anticipating it. In March 2021, it announced that it would no longer be engaging in microtargeting but focus instead on reaching cohorts (groups categorised by their click behaviour) rather than individuals (Morozov, 2021). It is doubtful, however, that this will have a positive effect on privacy and the creation of rabbit holes and echo chambers (Newton, 2021c).

The disadvantage of advert-free revenue models is that they create financial barriers that also exclude certain groups (Chen & Thorson, 2021; Van den Berg, 2021). There is a risk that online environments with less harmful behaviour will become something 'exclusive' in this way. Only a minority of households can afford more than one subscription (Reuters, 2020) and Netflix is often prioritised over a newspaper.

Value-sensitive design

Choosing value-sensitive design means putting public values first when developing online environments. In many cases, those values are derived from existing fundamental rights or human rights, such as the right to privacy or safety. Businesses can implement such values, for example, by carrying out a human rights impact assessment when building a platform or system (Advisory Council on International Affairs, 2020). They would then identify the potentially adverse effects of their project in the development phase and address them during the course of project (Danish Institute for Human Rights, The, 2020).

For example, the designers of a new microblogging service could ask themselves how user safety would be affected if someone could easily link a statement about another user to that person's online profile (Twitter's @mention). On the one hand, this feature encourages lively exchanges between users; on the other, it also facilitates such practices as shaming, or the escalation of online hatred. Can certain design choices create stronger barriers to such phenomena?

The mechanisms underpinning online harm can also inspire value-sensitive platform design. There are already websites and apps that prompt users to read content they wouldn't seek out themselves, exposing them in this way to alternative perspectives (Costa & Halpern, 2019). These sites and apps make very different use of usage patterns analyses than mainstream platforms. There are also initiatives aimed at embedding online social networks into existing, location-specific communities. The Dutch website Gebiedonline, for example, connects subscribers with people and businesses in their neighbourhood (Gebiedonline, 2021). The idea

is that building local connections will help to transfer norms and values from the physical world to online interactions and therefore help to offset the dehumanising effects and apparent 'lawlessness' of the online environment.

Many value-sensitive platforms are small in scale. On the one hand, this is a strength: after all, the scale of large platforms – leading to hyper-connectivity – feeds into many of the mechanisms behind harmful and immoral behaviour. On the other hand, however, it is a weakness because smaller platforms do not offer users a fully-fledged alternative to the dominant networks on which they are active and where they interact with many of their acquaintances. For this reason too, many commentators feel that stricter regulation of platforms is needed. One step that may help in this respect is to shift data ownership to users (e.g. Döpfner, 2021) and enforce interoperability, data standards and data portability (Costa & Halpern, 2019). If people can take their data with them and stay in touch with users on platforms they do not use themselves, it will be easier to switch to another service, thus encouraging the growth of alternative platforms. In 2021, the Rathenau Instituut advised the Dutch House of Representatives to consider additional regulatory measures for gatekeepers in its discussion of the *Digital Services Act*, as interoperability does not eliminate all the network effects that allow new platforms to quickly dominate the market (Rathenau Instituut, 2021b).

Advertisers' initiatives (intermediaries)

In addition to the platform companies, businesses that advertise their goods or services on these platforms can also do their part to limit harmful or immoral behaviour online. We discuss two types of interventions here. None have the reduction of harmful and immoral behaviour as their main objective; in all cases, that is a mere side effect of the advertiser's chosen strategy. Given the role that advertising plays in perpetuating underlying mechanisms (see Chapter 4), however, they are worth mentioning.

Blacklisting and whitelisting

Online behaviour can have harmful consequences for the users of online environments, but by extension, it can also be annoying for companies that advertise there (and contribute to the platforms' revenue model in that way). Businesses that care about their public image (or their 'brand safety') do not wish to be associated with problematic content, for example because their adverts will appear alongside it. To ensure that this does not happen, they can use an intermediary who works with whitelists and blacklists, i.e. lists of 'safe' and 'harmful' content. The intermediary compiles these lists by, for example, scanning transcripts of the audio of online videos for problematic terms (e.g. swear words, or words that

might be associated with sex offences). If such terms occur only rarely, then the videos, the channel where they are distributed, or their creators will be whitelisted.

Whitelisting and blacklisting may be an indirect means of encouraging the creation and sharing of non-harmful content, thereby helping to prevent harmful behaviour online. Content creators who end up on a blacklist run the risk of losing advertising revenue, potentially prompting them to create more content that *does* satisfy the requirements. Whitelisting is a 'positive' incentive to produce or upload 'safe' content.

One caveat about this method is that it makes advertising expensive. Small and medium-sized enterprises often do not have the budget for it and therefore choose quantity over quality (i.e. they aim for a wide reach rather than placing their adverts in the 'right' places). In addition, the algorithms used by intermediaries to scan transcripts for problematic terms are not flawless. They sometimes pick out terms that are not problematic, or overlook ones that are. Moreover, about half of the content that is blacklisted and removed subsequently reappears elsewhere. In that respect, whitelisting and blacklisting are only a temporary solution.

Contextual advertising

Many platforms use programmatic advertising, i.e. an automated placement system for adverts that works with user profiles. Browsing activity is tracked with cookies. The system then uses the data this yields to link a visitor to a specific user profile. The advertiser bids for a certain profile and the system ensures that visitors who fit that profile see its advert. This is also referred to as 'personalised advertising' because it is the user's personal data that determine where adverts are placed.

An alternative to this method is contextual advertising (or contextual targeting). This does not involve cookies; instead, the advertiser combines its advertising message with a certain type of content (as used to be customary in print newspapers). The assumption is that this content will attract an audience that also has an affinity with the product advertised. A forerunner in this respect was the Dutch public broadcasting service NPO, which began experimenting with contextual advertising in 2018 together with STER, the foundation that sells advertising space on its channels (STER, 2020). NPO took this step after the vast majority of its website visitors opted out of cookies. The two largest newspaper publishers in the Netherlands (Mediahuis and DPG Media) are now also planning to transition to contextual advertising (NLProfiel, 2020).

Platforms can use contextual advertising to allow their visitors more privacy, but also to win back advertising revenue that would otherwise go to other platforms. Traditional media offering verified content are competing with newer platforms for

advertisers. By adopting a different advertising strategy, they are also targeting a different clientele: businesses that wish to decide for themselves what kind of content their adverts are shown with. Platforms that use this strategy may see an upgrade in the quality of the content offered there. Contextual advertising can also ensure that less advertising revenue goes to fraudulent websites and more to content creators.

One disadvantage of not having cookies is that there is no information available about what visitors actually do on a website. Advertisers sometimes find this inconvenient, because they would like to gauge the reach of their adverts. It is also thought that contextual advertising works particularly well with the traditional content of newspapers and broadcasters, for example, where the nature of that content is known – unlike on social networks.

Products and services for online safety

A third group of businesses that are already working to combat harmful and immoral behaviour online are producers and providers of ‘online safety tech’, products or services to facilitate safer online experiences, and protect users from (potential) harmful content, contact or conduct (Department for Digital, Culture, Media and Sport, 2020). We distinguish between products and services developed for businesses and those meant for private individuals.

Products and services for businesses

Automatic detection of harmful content

We mentioned the automatic detection of problematic messages or images in the foregoing. Platforms themselves are developing such technology (Rathenau Instituut, 2020a; Sánchez Montañés, 2021), but there are also businesses that specialise in it and offer their products to platform operators, web hosting companies and advertisers (Costa & Halpern, 2019; Department for Digital, Culture, Media and Sport, 2020).

One such automated detection technique is hashing, which works with ‘hash codes’ or ‘hash values’, a short numerical representation of an image that functions as a unique identifier or digital ‘fingerprint’. The technique is already being used in the Netherlands to track down child sexual abuse material. The Dutch Child Pornography Reporting Office, for example, makes a hash check server available to web hosting companies and other businesses free of charge. The Dutch Ministry of Justice and Security encourages businesses to make use of it under the industry’s notice and takedown code of conduct (see section 5.1), and to remove problematic content in the event of a ‘hit’ (Tweede Kamer, 2018).

Identity and age verification

Another relevant area of specialisation is the development of age or identity verification tools. Tools based on attribute-based identity management are particularly promising; they allow users to verify their identity without disclosing too much privacy-sensitive data (using 'attributes'). A Dutch example is the identity platform IRMA. Users create a 'digital passport' that they can then use to log into restricted online environments.

Age verification technology is used to match content to specific groups of users. The international technology company SuperAwesome, for example, helps content owners, platform operators and advertisers to ensure that young users do not come across 'inappropriate' content, such as adverts targeting adults (SuperAwesome, 2021).

Behavioural analysis and personalised messaging

Businesses and organisations also develop methods for analysing user behaviour. They can be used to identify vulnerable groups, for example internet users with a propensity for gambling addiction or self-harm, and then personalise their search results or generate banners referring them to professional help (e.g. Costa & Halpern, 2019). One example of this is the Redirect Method, developed by tech start-up Moonshot and Google incubator Jigsaw, among others. It is an open-source methodology that identifies individuals who are searching for harmful content by analysing their online search terms. They are then sent targeted advertising with constructive alternative messages (Moonshot, 2021). Poland's Samurai Labs recently even built a 'reasoning machine', i.e. a bot that can intervene in online conversations and prevent them from escalating into online hate speech and cyberbullying (Konopka, 2021).

Another way to encourage healthy interaction is to curb users' tendency towards impulsive online behaviour by inviting them to reflect on their own actions at regular intervals. For example, software is under development that automatically detects potentially hurtful or offensive content and buffers it for a short time, subtly giving users an opportunity to change their minds (Costa & Halpern, 2019). Other systems use prompts or reminders that specifically encourage users to reflect before posting. Research shows that this may be a useful approach to tackling harmful behaviour. For example, asking users to review the quality of their posts or messages may help prevent mindless forwarding (Pennycook et al., 2021) and also help curb online mechanisms such as virality.

Commentators stress, however, that online safety tech will only take off in a development climate that fosters new intermediaries, such as commercial software companies (Costa & Halpern, 2019). Government guidance is crucial in this regard.

Research shows that large platforms still make little use of age verification and validation technology (Aiken, 2016), even though it is gradually becoming available. By encouraging or forcing platform or web hosting companies to use them, or by investing in their development themselves, governments can drive research and development (Helberger et al., 2018). The UK government wants to lead the way, by encouraging and supporting the safety tech sector in parallel with the writing of the *Online Safety Bill* (Department for Digital, Culture, Media and Sport, 2020; UK Government, 2020b).

At the same time, using private companies to combat online behaviour or the resulting harm also poses dangers. This is particularly the case when using artificial intelligence, for example to identify vulnerable groups (as Instagram does to ascertain the age of users (Instagram, 2021) or to alter their search results accordingly. Researchers point out that such practices may, for example, put privacy at risk (Costa & Halpern, 2019) and mays also infringe the GDPR, which establishes an interpretability principle. Users are therefore entitled to have an explanation of how the algorithm has 'reasoned'. In the case of self-learning algorithms, however, that reasoning is not always possible to ascertain, not even for IT specialists.

Products and services for private individuals

Another category of online safety tech consists of products and services that allow users to protect themselves or others from harmful or immoral behaviour or its consequences. We distinguish two categories: filters and blocks, and self-exclusion tools.

Filters and blockers

Filters include those that parents install on their computers to prevent children from viewing certain content. Well-known examples are the Dutch Kliksafe or the American Net Nanny (Kliksafe, 2021; Net Nanny, 2021). These products are not always effective, however, as they can both underblock and overblock access to material (Oosterwijk & Fischer, 2017).

Self-exclusion tools

Experts see potential in all kinds of self-exclusion tools to counteract the negative impact of such online mechanisms as availability and continuity. Examples are filters or blockers that people can use to protect themselves against online addictions (e.g. Pluckeye, Cold Turkey or LeechBlock), or software that blocks specific websites or networks (always, or at certain times of the day). Businesses that facilitate certain high-risk transactions also sometimes offer this type of product. Many UK banks, for example, give their customers the option of blocking transactions on their account if the payment request comes from an online

gambling website. They can only lift the block after a specified delay (Costa & Halpern, 2019).

Although this type of option does not eliminate the risk of problematic behaviour, it does encourage users to think about their actions in advance. The filters should be deployable on different platforms simultaneously (Costa & Halpern, 2019), and users must be able to decide for themselves whether to use them. After all, undue control and oversight can be detrimental to people's autonomy and privacy.

5.3 Social welfare services, civil society organisations and users

The fight against harmful and immoral behaviour online is being waged not only by governments and businesses, but also by social workers, civil society organisations, individuals and collectives. To conclude this survey of existing interventions, we highlight four current strategies: creating awareness, informing and educating; participative intervention in online interactions; victim support; and participative forms of value-sensitive platform design.

Creating awareness, informing and educating
Over the past few years, various civil society organisations and citizen groups have campaigned to improve the quality of online interactions. In February 2021, for example, a number of well-known Moroccan-Dutch personalities launched a campaign against online shaming, with politicians, actors and writers starting a petition and speaking out collectively on social media against the online culture of harassment using the hashtag #StopShaming (Redactie NOS, 2021). Citizens' movement DeGoedeZaak also launched an appeal against online hate speech, which included a toolkit with tips and advice about how people could protect themselves or take action on their own (DeGoedeZaak, n.d.).

Commentators stress the importance of such initiatives, in which people speak out publicly about problematic behaviour. After all, all users of online services play a part in the interactions that take place online, and so the way these interactions unfold is a shared responsibility (Rasch, 2021).

'Creating awareness' also covers initiatives in which civil society organisations put pressure on platform companies to take action against harmful behaviour online. Ranking Digital Rights, for example, is a collective of researchers and activists working to advance online civil rights. One of their actions was to rank platforms, from Twitter to Amazon, based on relevant indicators. A salient detail that emerged from their ranking was that platforms still show little willingness be open about how

they collect user data and moderate online interactions, and about the algorithms they use to do so (Brouillette, 2020). Through its research and publications, the collective brings such problems to the attention of the platforms themselves but also of governments (policymakers) and investors.

Specialist media literacy organisations make a more structural contribution to creating awareness of the risks of online interactions. Netwerk Mediawijsheid now has more than 1000 member organisations active in the Netherlands (Netwerk Mediawijsheid, 2021). Many of them have developed projects on cyber security and cyber resilience, usually aimed at children or their carers (Bureau Jeugd en Media, 2021). In addition, such organisations are particularly concerned about fake news and disinformation and about social media etiquette. Projects aimed at children sometimes involve gaming. One example is the game *Slecht Nieuws* (Bad News, about fake news), in which users are asked to imagine themselves as 'bad guys' in order to learn how fake news is created and spread (Slecht Nieuws, 2021). Another interesting project at the interface of education, art and activism is TheirTube, a filter bubble simulator that reveals the role that data and algorithms play in a YouTube user's viewing experience (TheirTube, 2021).

There is a pressing need to take action and invest in media literacy and digital skills not only among children, who are using the internet at an increasingly early age, but also among adults – especially vulnerable ones (Aiken et al., 2016; Rathenau Instituut, 2020a). Unfortunately, there is still little research in this area. For example, it is not clear whether many of the programmes addressing harmful behaviour actually work (see e.g. Oosterwijk & Fischer, 2017). We also know very little about what young people do online and what this means in terms of their physical, cognitive, and emotional development (Aiken, 2016).

Apart from that, experts emphasise that media literacy projects can only succeed if they are designed with a proper feel for the online world of the intended audience. Organisations that have relevant expertise are therefore often better suited to implementing such projects than governments or their executive agencies. Governments can, however, encourage or support these organisations, including financially. Experts also recommend a constructive – rather than repressive – approach, as young people in particular are rarely aware of the consequences of their online behaviour. A conversation about online norms and values should be at the core of such efforts, according to our interviewees.

Finally, it is important to consider the broad social network that young people have. In addition to teachers, others can help to create awareness, inform and educate in the fight against immoral and harmful behaviour online (Oosterwijk & Fischer, 2017). Combining strategies is also seen as conducive to success. An example of a

combined approach is the series of actions taken by the City of Amsterdam to combat sexual harassment and violence. The municipal authorities have allocated funding for research and conducted a simultaneous campaign against shaming (#jijstaatnietalleen, i.e. #youarenotalone) and provided information at schools (Wagemakers & Toksöz, 2021). The City is also investigating the possibility of imposing an 'online restraining order' on perpetrators (Katawazi & Wagemakers, 2021; Wagemakers & Toksöz, 2021).

Participative intervention in online interactions

Alongside organisations, individuals and groups of users are also making an effort to improve the quality of online interactions. Such efforts are known as 'technological placekeeping', the practice of active care and maintenance of digital spaces. People can use it to defend themselves against harmful phenomena but also to promote the health of digital conversations and in doing so make online environments more pleasant for everyone (Wong, 2021).

A recent Dutch initiative in this respect is the hashtag campaign #DatMeenJeNiet (#YouDontSay) by Movisie, a knowledge institute that addresses social issues (Movisie, n.d.). Participating teenagers promise not to behave like bystanders when they come across online discrimination (see Chapter 4), but like *up*standers, i.e. someone who calls others out on their problematic behaviour. Another example, this time regarding disinformation and misinformation, is *Make Media Great Again*. This project gives volunteers tools for annotating online articles and audiovisual productions (Make Media Great Again, 2021). The annotations serve as suggestions to the editor, who can improve the quality of the reporting in this way. Dutch news site NU.nl was the first media partner to join the initiative. The virtual neighbourhood watch could also be seen as a form of technological placekeeping. It is a form of online neighbourhood crime prevention in which tech-savvy internet users cooperate with law enforcement to identify and mitigate cybercrime vulnerabilities in software (see e.g. Oosterwijk & Fischer, 2017).

Like the aforementioned citizens' campaigns, such initiatives also encourage users of online environments to take responsibility for the quality of online interactions. Commentators stress, however, that the responsibility is always a shared one, and that government and businesses must do their part (Helberger et al., 2018). Businesses in particular could do much more to support or reward citizens who actively work to create safe spaces on their platforms (Wong, 2021).

Assisting victims

Victims of harmful or immoral behaviour online are unlikely to receive specialist assistance at present. Children and their carers can go to the portal Meldknop.nl, launched in 2012 by the Dutch Child Pornography Reporting Office and Digibewust

('Digi-aware', a programme run by the then Ministry of Economic Affairs),[13] if they experience violence, bullying, fraud, sexual harassment or something else unpleasant online. The site provides information and advice, along with explanations and tips, on 21 phenomena categorised under the headings bullying, sex, scams and harassment. In addition, they can contact experts at affiliated organisations directly by e-mail, chat or telephone (and/or by downloading an app), including Helpwanted.nl, Vraaghetdepolitie.nl, MiND and Pestweb. Meldknop.nl has approximately 50,000 visitors a year on its homepage, a number that has remained fairly constant in recent years. The site does not provide information on the number of reports submitted to the organisations to which it links (Meldknop.nl, 2021).[14]

Victims of behaviours that have an 'offline' counterpart can sometimes also seek help from organisations that address the broader phenomenon. For example, someone who is bullied online can contact Stichting Stop Pesten Nu (Stop Bullying Now) for information, and someone who struggles with cyber addiction can turn to a mental health authority or other organisation that deals with addiction issues.

Nevertheless, the experts we interviewed for this study believe that victims of online behaviour require a different kind of help. They emphasise, for example, that social workers or others who provide support should also have an online presence. After all, the internet is often the place where the victimisation is being perpetuated. People who are prone to eating disorders, for example, can go online to seek help or inspiration, but the internet can also bring them into contact with those who would exploit their vulnerabilities (see the case on disturbed eating behaviour). To overcome such mechanisms, victims need to be supported in both the online and offline worlds (see Chapter 6).

In addition, victims can only get the help they need if the right people are involved. As in the case of media literacy initiatives, it is considered crucial to be attuned to the victims' world. For example, experts argue that organisations should make more use of former victims of harmful behaviour online who can share their self-protection techniques or recovery strategies from the perspective of a fellow sufferer. Equally important is to build online networks and safe havens to support victims, for example in cases of online hate speech or harassment.

Here too, our interviewees say it is better for government to support existing initiatives and help them to scale up than to take on too many tasks itself. Even so,

13 Since its launch, the portal has been taken over by Veiliginternetten.nl, a joint initiative of the Ministry of Economic Affairs and Climate Change, the Ministry of Justice and Security/National Cyber Security Centre, ECP | Platform voor de InformatieSamenleving, and the business community.

14 Source: e-mail by Meldknop.nl spokesperson at ECP, dated 23 June 2021

they also point out the importance of proper policy. After all, victimisation online often stems from 'offline' vulnerabilities, such as a weak socio-economic position.

Participative forms of value-sensitive platform design
In addition to technological placekeeping, users themselves can influence digital spaces directly or indirectly by contributing, financially or otherwise, to value-sensitive platform design (see section 5.2). We have already mentioned Okuna, an 'alternative' social network that is advert-free, does not track its users, and does not monetise their personal information (Okuna, 2021). Besides giving users more freedom to decide what they see in their feeds (rather than focusing on spectacular, 'viral' content), Okuna also involves them in developing rules of conduct. The social network keeps users engaged in this way and nudges them to display desirable behaviour and encourage it in other users. Another example is the aforementioned Dutch Gebiedonline, a cooperative online environment that tries to connect people in specific offline neighbourhoods or around specific themes, and that aims to maintain or improve social cohesion, with online and offline as extensions of each other (Gebiedonline, 2021).

Researchers who investigate platform design and public values argue that users should in any case play a bigger role in designing the online environments where they spend their time, even if the platforms belong to technology giants. If a platform is clearly taking decisions that go against users' interests or that threaten the health of online interactions, users can exert collective pressure on the company behind it, for example on social media or by raising such issues with the regulator (Dijck et al., 2018; see also Helberger et al., 2018). Governments can help pave the way for users.

5.4 Conclusion

Our overview of existing measures shows that various parties are already taking steps to counter or prevent harmful and immoral behaviour online. These initiatives have offered us inspiration for the strategic agenda that we will discuss in Chapter 6. At the same time, however, there are some striking shortcomings. The most significant one is that many of the current initiatives are quite *reactive* in nature. They are mainly aimed at combating the symptoms of harmful and immoral behaviour, not at the underlying mechanisms. In this respect, we do see differences between various stakeholders. Governments and large platform companies in particular have not been very proactive. In the latter case, this is not surprising. After all, tinkering with mechanisms means choosing an alternative platform design, but that would create uncertainties about revenue models – and when all is said

and done, platforms operate in an established arena. The reality, then, is that it is primarily smaller-scale parties that are experimenting with alternative designs.

Governments mainly take action when behaviours get out of hand and therefore need to be restrained. Clarification of the mechanisms underlying the phenomena of harmful behaviour online, as we have tried to offer in this report, can help governments and other stakeholders to be more pro-active. In Chapter 6, we make further suggestions for boosting the government's knowledge position.

Social workers that we interviewed acknowledge the importance of online mechanisms, but they find it difficult to make the necessary changes in their approach. The existing social welfare organisations tend to focus on phenomena that also have an 'offline' equivalent and are therefore already part of the social welfare landscape. Their professionals struggle with the mechanisms underpinning the online version of such behaviour, such as availability and continuity, syndication or scalability.

Civil society organisations and concerned individuals and collectives tend to function as complements to governments and their executive agencies in terms of their focus on specific phenomena. Governments tend to concentrate on information manipulation and online hate speech, and to a lesser extent on self-harm phenomena. Civil society organisations, on the other hand, take action against bullying and violence or digital vigilantism. In their approach, they also focus more on the underlying mechanisms, such as dehumanisation or unclear norms. That is also the case for some online safety tech firms.

Another trend that has emerged from our overview is that social welfare and civil society organisations play an important role as 'watchmen'. Their knowledge of the online world or their expertise regarding a specific phenomenon means that they can often identify problems arising from online behaviour before they are detected by the government or a platform. This is reason enough for (executive branches of) governments to cherish social workers and civil society organisations and to encourage them in their efforts.

In the next chapter, we propose a strategic agenda for government in cooperation with stakeholders in the private sector and society, based on lessons learned and identified shortcomings.

6 Strategic agenda

This study is the first to map all aspects of harmful and immoral online behaviour in the Netherlands. The Rathenau Instituut has developed a taxonomy of six categories of harmful and immoral conduct online, listing 22 different phenomena that all internet users in the Netherlands may encounter sooner or later (Chapter 3). Our taxonomy is a snapshot; new phenomena will continue to emerge.

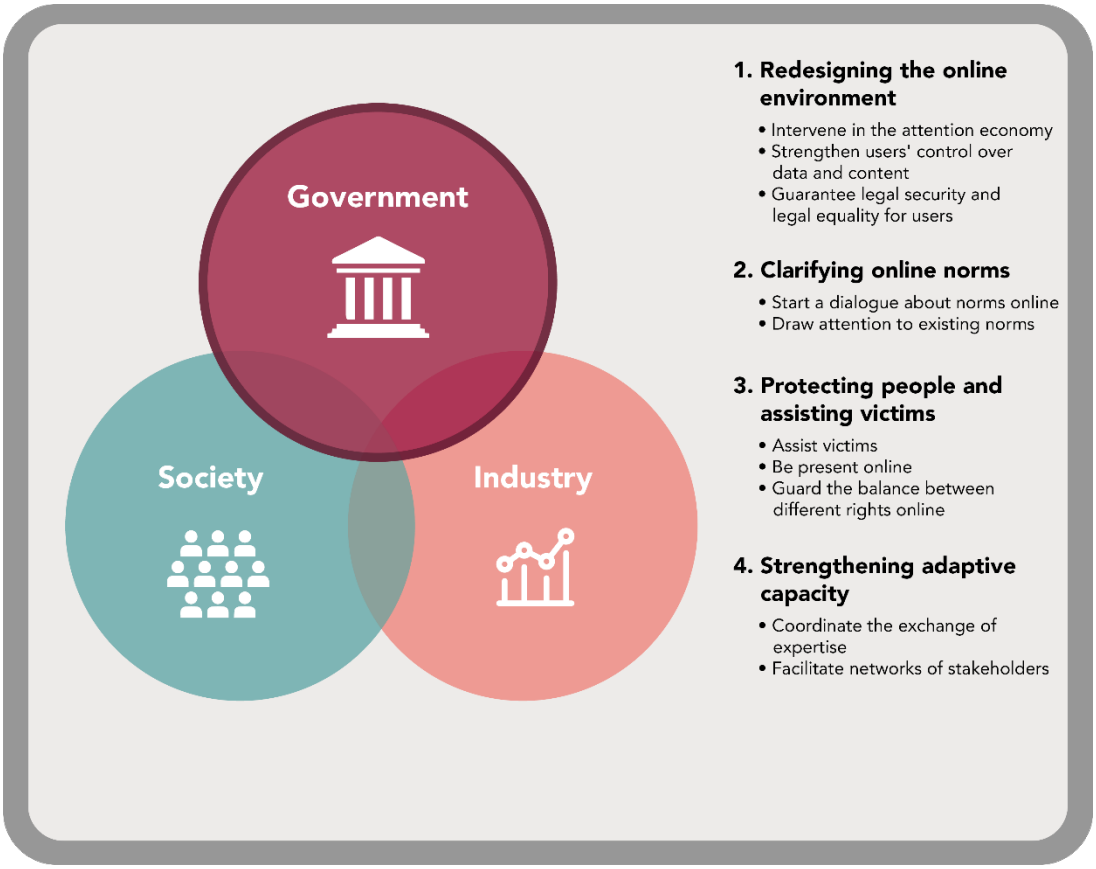
The harm that such behaviour causes can have a heavy impact on individuals, groups and society as a whole. It can range from a teenage girl starving herself because she joins an extreme challenge with other adolescents (see the case on disturbed eating behaviour) or female journalists and scientists being afraid to speak out online (see 'hate speech' in section 3.4) to societal disruption due to the spread of conspiracy theories and disinformation (see the case on disinformation). The scale of this type of phenomenon as described by experts and in the literature shows that Dutch people who go online run a considerable risk of falling victim to harmful behaviour or of slipping into it themselves. Guaranteeing a safe online environment is therefore a major challenge for society, one in which the government can take the lead, together with other stakeholders.

In addition to explaining the nature and scale of harmful and immoral behaviour online, this study also systematically identifies the characteristics and mechanisms of the internet that play a role in initiating, facilitating and amplifying harmful and immoral behaviour online (Chapter 4). For example, we have discussed anonymity, virality, the attention economy and 15 other mechanisms. Nevertheless, very little has been done to reverse these mechanisms (or to limit their adverse effects), and existing interventions are not enough to guarantee online safety (Chapter 5).

The internet has so far been a domain of self-regulation and self-reliance, where the government has taken no oversight role and users were thought to manage themselves. Our research shows that people lack adequate protection on the internet and that fundamental rights are therefore at risk. Businesses, civil society organisations and the public at large need coordinated collective action to counter harmful and immoral behaviour online. Limiting such harm requires intervention in the mechanisms that reinforce harmful and immoral behaviour online. The government has a duty to protect people's fundamental rights, online as well as offline. This agenda offers tools with which to do so.

Based on interviews and discussions with experts in the fields of policymaking, scholarship and professional practice, a review of academic and journalistic

sources and policy documents, combined with the expertise gained by the Rathenau Instituut in previous research and analysis, we identify four strategic themes in which the Dutch national government can play a guiding, coordinating and facilitating role. The agenda is meant to enable the Dutch government to cooperate with stakeholders from industry and society to tackle harmful and immoral online behaviour and to promote a safe online environment.



Bron: Rathenau Instituut

Figure 6 Strategic agenda for tackling harmful and immoral behaviour online

The first theme – *Redesigning the online environment* – contains tools for the Dutch government to reverse the online mechanisms that characterise the internet and contribute to harmful and immoral behaviour online. The second theme – *Clarifying online norms* – deals with the role of the Dutch government, industry and society in renewing the social agreements on norms and values online. The options for action under this theme are intended to bring about broader awareness and understanding of harmful and immoral behaviour online. The third theme – *Protecting people and assisting victims* – contains suggestions for the Dutch government, law enforcement and executive agencies to better respond to the phenomena of harmful and immoral behaviour online and the harm they cause. The fourth theme –

Strengthening adaptive capacity – offers suggestions for the Dutch government to gain and maintain a grip on harmful and immoral online behaviour, which is constantly changing. These suggestions are aimed at future-proofing the four strategic themes that make up the agenda.

Several challenges have been identified under each of the four themes. In each case, we suggest the consideration of a number of options for action that have emerged from our study.

6.1 Theme 1: Redesigning the online environment

We need a government that makes policies to redesign our online environment in a way that addresses the mechanisms contributing to harmful and immoral behaviour and that prevents harm. Stakeholders in the private sector and society are seemingly incapable of preventing online harmful behavior. National and international authorities do not seem to address the underlying mechanisms of online harmful behavior at the time of writing. The Dutch Ministry of Justice and Security could seize upon this realisation to push for action on the mechanisms in the forthcoming government's Digitisation Strategy. The Dutch national government can also use the EU legislative procedures of the Digital Services Act Package, the Artificial Intelligence Framework, the Data Act and the Data Governance Act as an opportunity to revise these mechanisms.

In its Manifesto (2020c), the Rathenau Instituut stated that the government must more effectively counterbalance the power of the big technology companies, which now dominate the design of the online environment. At the same time, these businesses should take more responsibility for protecting the rights of their users. Our discussions with experts and review of many sources have led us to identify the following three tasks for government in designing the online environment:
1) intervene in the attention economy;
2) give users more control over their data and content;
3) guarantee legal security and legal equality for users.

Intervene in the attention economy

The attention economy that has emerged online has a major hand in driving harmful and immoral behaviour online. Users and online platforms compete for the attention of other users, and outrageous or shocking content does this best. Such harmful and immoral content is closely associated with harmful and immoral behaviour online; harmful behaviour leads to harmful content (e.g. hate speech or threats and

harassment), and harmful content may provoke harmful behaviour, for example extreme challenges that encourage young people to harm themselves. Based on our literature review and interviews with experts, we have identified a number of options that the national government could consider in this context:

Get a better grip on the reach of content
It is important to distinguish between regulating content on the one hand and distributing it on the other (Rathenau Instituut, 2021a). Platform revenue models and algorithms play a critical role in the distribution of content. The government could attempt to get a better grip on the reach of that content by requiring tech businesses to be more transparent about how their algorithms select it for users, as is already being discussed within the context of the Digital Services Act. The algorithms used by social media platforms are designed to serve the paying customer (often the advertiser). Imposing transparency requirements on platforms could lead to a better understanding of how these algorithms work and may allow an independent regulator to monitor and control them. The national government can use the policy discussion regarding the DSA and the AI Framework to promote more transparency and stricter oversight of algorithms.

Tackle the market dominance of gatekeepers
The government can tackle the market dominance of gatekeepers to create more room for new parties that may offer a more ethical platform design. It can do this by tightening up the competition rules. The EU is already engaged in relevant political discussions within the context of the draft Digital Markets Act (DMA). Another option is to focus on the gatekeepers' revenue models. For example, the EU can restrict or prohibit the exploitation of personal data and microtargeting, something that can encourage platforms to choose advert-free revenue models. Finally, the government can explore the extent to which forms of ownership other than a stock exchange listing (for example user cooperatives) can ensure that online platforms are better able to fulfil their role of 'critical utility'.

Increase the digital autonomy of public services
The government can ensure that the technical infrastructure of public services, for example the media, education and health care, is more firmly under public ownership. What we consider to be public space offline (for example public squares, schools or roads) is, in the online world, in the hands of private American and Chinese tech companies. That means that the Netherlands cannot set its own rules. A counter-manoeuvre would be to work towards national or European digital autonomy, for example by tightening up the terms and conditions of purchase for suppliers of digital products and services, and by building more Dutch and European internet businesses. It remains to be seen what options for boosting digital autonomy are realistic for a small country like the Netherlands, but they may

well include stricter terms and conditions for digital infrastructure and autonomous oversight. Europe views ‘strategic autonomy’ as increasingly important (Vanheste, 2021). To finance its investment in domestic technology, the national government can claim funding from the EU’s Recovery and Resilience Facility, which has reserved billions for the digital transition.

Encourage value-sensitive technical design
The government can do more to encourage the value-sensitive design of digital infrastructure so that mechanisms such as scalability, hyper-connectivity and virality can be addressed. One example of value-sensitive design cited in the literature and our discussions with experts is content customisation, which can help in developing online solutions to protect victims. At present, such solutions are mainly provided by small companies and platforms that are unable to grow in a market dominated by large ecosystems. Governments could support their growth by subsidising national or European tech companies. Giving users more value-sensitive technology would empower them to create their own online spaces with their own selected content. Online harm cannot be prevented by technical means alone, however.

Promote other forms of advertising
The government can help combat online harm by encouraging advertisers to look less at quantity and more at quality when measuring their reach. That could go some way towards tackling such mechanisms as scalability, virality and the attention economy. Advertisers could focus more on the content itself, rather than on reach and data. Examples suggested by the experts consulted for this study are whitelisting and contextual advertising. The government could use its influence to make these (still) relatively expensive forms of advertising more accessible and affordable for small advertisers. When major advertisers publicly opposed hate speech and racism last year after George Floyd’s death, the effect was limited. Large advertisers represent only a small share of the total revenue of social media platforms.

Give users more control over their data and content

The government has various options for giving users more control over their data and content. Users are now often locked into a particular platform because they built their own online networks on it with friends and family. They cannot simply migrate this network to another online environment where they have opened a new account. This limits their autonomy and therefore their ability to choose an online environment that suits their values and needs and where online mechanisms have been adapted to encourage socially desirable behaviour. Users also have relatively little influence over how their personal data is used and the content

recommendations that platforms make to them and other people in their network. The government can explore the following options to give users more autonomy and control:

Data portability and interoperability
The government can give users more control over data, and encourage data portability and interoperability. At time of writing, data portability and interoperability are being discussed within the context of the EU's draft Digital Markets Act. This is one way of curbing the market dominance of large platforms, creating more room for new providers that give users more access to their personal data or data profiles and more control over how they are used. The caveat, however, is that protection of privacy must always be guaranteed.

Subsidies and public procurement
The government could encourage the development of ethical tools that prevent problematic behaviour from arising or escalating by supporting them financially and in its procurement procedures. Examples are tools for age verification, (upload) content filters or detection software. The government can also encourage revenue models that work on a subscription or membership basis. Users tend to behave with greater civility in an environment in which they are obliged to pay for editorial content. Users are also less anonymous in such environments, as they can be traced through payments. If users are the main source of income for online platforms, their interests automatically take precedence over those of advertisers and their safety and wellbeing will also have greater priority.

Guarantee legal security and legal equality for users

To address the perception – created by unclear norms, anonymity and apparent lawlessness – that disorder reigns in the online environment, online users must be afforded greater legal certainty and legal equality. This study has identified several options that the national government can use to improve online users' legal certainty and equality.

Examine the advantages and disadvantages of online identification
The government can examine the advantages and disadvantages of some form of online identification. An online surfing licence or admission ticket could work as a preventative by stripping perpetrators of their online anonymity. Online identification would make it easier to track down, punish or retaliate against them and can therefore act as a deterrent, but at the same time it would expose to greater risks groups in need of protection, such as victims, journalists or whistle-blowers. A

measure of this kind requires careful consideration of the pros and cons (see also theme 2).

Make international agreements about online jurisdiction
The government can push for international agreements on the enforcement of laws and regulations online. At present, national jurisdictions and the global internet are difficult to reconcile. For example, it is difficult to remove a website with disinformation or other harmful content if the organisation behind it is not located in the territory of the Netherlands. International agreements seem needed to tackle such transnational issues.

6.2 Theme 2: Clarify online norms

The government, the private sector and society all have a role to play and a responsibility to bear in clarifying and monitoring online norms. One of the biggest challenges when it comes to preventing and combating harmful and immoral behaviour is the lack of clear norms in the online environment. As a group, the mechanisms we describe in Chapter 4 produce a sort of moral fog that makes it much harder to recognise behaviour online that we would label unacceptable in the offline world. In addition, digital environments are also giving rise to new behaviour (for example grooming in virtual reality settings). In such cases, the social process that creates norms has yet to kick in, or existing boundaries need to be renegotiated for the online environment.

Morality is the outcome of a social contract. Neither government nor commercial parties such as platform companies have the authority to determine on their own how people should behave towards one another. At the same time, each of these parties is responsible for behaving in a socially desirable way and for respecting the boundaries of the law. A wide-ranging public debate is needed to ensure that online norms are clear and explicit. Our discussions with experts and specialists and our review of the literature have led us to identify the following two tasks for government, the private sector and society when it comes to clarifying online norms:
1) engage in a dialogue about norms and values online;
2) draw attention to existing norms.

Engage in a dialogue about norms and values online

The sort of wide-ranging debate that is needed will not happen spontaneously. Society needs support and encouragement to have a conversation about online

norms. The government can play a facilitative and supportive role here, but businesses and civil society organisations can also ensure that all those concerned about the liveability of online environments join the conversation. Our study produced the following options, involving different stakeholders.

Deliberative processes
Governments play a facilitative role in the conversation about online norms. One option is for the government to initiate and organise the dialogue itself. This can be useful when preparing new policy, for example about anonymity on online platforms. Anonymity is an important mechanism behind harmful and immoral behaviour because it encourages users to throw off all restraints. Removing or restricting anonymity could have a preventive effect, then, but anonymity also offers victims, journalists or whistle-blowers protection in the online environment. Additionally, there are many different ways to restrict user anonymity (using personal data or only using identity attributes – publishing them online or using them solely for verification purposes). Society and the political world will therefore first have to consider which interactions require openness about a person's identity and under what conditions is a degree of anonymity is acceptable, or even necessary.

Bottom-up rules of behaviour
Platform companies often work with rules of behaviour, for example cast in the form of terms of use. But when they devise these rules themselves, they claim a great deal of power for themselves in defining social norms. Governments could encourage or oblige platform companies to involve their users much more actively in drafting rules of behaviour. Users will 'own' the rules this way and be more inclined to follow them. To foster the conversation about norms, platform companies can also facilitate online deliberations, for example by creating highly visible and easily accessible online spaces for it. This approach can also be attractive for the companies themselves, because they can then tailor their services more effectively to meet the needs articulated by society.

Dialogue with young people
Media literacy training today focuses mainly on teaching children and young people the skills they need to protect themselves from harmful behaviour online. They can, however, also become more actively involved in the conversation about online morality. Youngsters growing up today become socialised online to some extent, where authority figures are less visible than in the offline world (see e.g. Aiken, 2016; Cocking & van den Hoven, 2018). Their carers should therefore talk to them about desirable and undesirable behaviour in cyberspace.

It would be advisable to make victim blaming – blaming the victim instead of the perpetrator – part of that conversation. Victims of harmful behaviour online (and in particular catfishing, grooming and shame-sexting) are often the object of victim blaming, and it is an obvious symptom of the normative chaos that can arise in online environments. Young internet users should not have to change their behaviour in advance for fear of being condemned by others, while perpetrators often get off scot-free. The experience of victim blaming often causes victims to feel deeply ashamed. Carers of young victims should focus on removing that shame by talking to them about how norms have been transgressed in their case.

Parties that understand how young people perceive the world would be the appropriate choice for initiating this conversation. Here, too, it is best for the government to support existing expertise in executive agencies and civil society organisations and to empower them by publicising their initiatives or supporting them financially.

Draw attention to existing norms

Once a society has ’agreed’ to a set of norms, they must be given the necessary attention, both online and offline. But that won’t happen by itself. The government and its executive agencies play a role in forcing this attention, but so do the private sector and civil society organisations. We see considerable potential in this respect, as many more parties could leverage their position than they do at present.

Condemn bad behaviour, encourage good behaviour
An executive agency such as the police now operates online, for example to take action in environments where criminal offences are committed (see theme 3). But in addition to enforcing laws and regulations, the police also have another task: spotting and preventing problems. In the online environment, the police can do this by publicly condemning undesirable behaviour. In specific communities, such as the gaming community (where some police units are already active), they can also encourage desirable behaviour in situations that are about to escalate, for one thing by setting a good example. To ensure that such interventions are scalable, the police must have more human resources at their disposal.

Make rules of behaviour more visible
Platforms often have terms of use that are difficult to find and, generally, to understand. The government can force businesses to make their rules of behaviour more accessible in a variety of ways. It would be advisable to involve different groups in drafting these rules, so that all target groups, including vulnerable ones, can understand them easily. In addition, the government can exert pressure more

informally, for example by condemning a failure to act. We concluded in Chapter 5 that this is uncommon in the Netherlands; government ministers rarely admonish businesses publicly. Civil society organisations do so more regularly, and the government could build on their efforts. Advertisers can also help, for example by making content creators aware of existing rules, such as the Dutch Advertising Code.

Mobilise institutions
A wide range of institutions are already involved in adopting or developing social norms, even if their own mission does not cause them to identify with an online problem. Political parties, but also churches, youth movements or even sports associations could broaden their view to online morality and do their part in getting it on the agenda. One good example is the recent campaign in which a number of British football clubs and UEFA and FIFA boycotted social media for three days to draw attention to racial verbal abuse. They already appeal to a specific demographic and can therefore tailor their strategy to what works for that group. The government can enter into a dialogue with such institutions and mobilise them to draw attention to online norms.

Engage in technological placekeeping
Internet users already engage in all kinds of technological placekeeping, i.e. the ‘maintenance’ of online environments, yet few people speak out publicly about others’ online behaviour. Platforms can encourage this by rewarding existing efforts. For example, they could highlight posts by groups actively engaged in the fight against disinformation. In so doing, they would also be serving their own interests; after all, platforms themselves benefit from a healthy online environment. Governments or regulators can, if necessary, oblige companies to adopt initiatives of this kind.

6.3 Theme 3: Protecting people and assisting victims

The government has a primary duty to protect fundamental rights. It must do so online as well as offline. Whenever people become victims, they should receive assistance. Specific groups that are vulnerable online, such as minorities and children, have the most to gain from government protection online.

The phenomena and case studies described in this report show that people are often unprotected in the online environment. Victims do not feel as if either the government or social welfare organisations are there for them or that they have the same level of protection as in the physical domain.

Although the online environment is not actually lawless, it is often perceived as such. For purposes of redress, it is important to punish those who behave unlawfully or who commit crimes, but that does not happen very often online (Van De Weijer et al., 2020). The scale of the phenomena does not allow for it. Even in the case of phenomena that are or may be punishable, then, a legal strategy based on punishment will not do enough to protect people.

There are various reasons for the absence of protection online. One is that people use services based in another country, where malicious parties may have free rein. Another reason is that it is difficult to reach online platforms for assistance. Yet another factor is the extent to which private parties or the government have already organised assistance for victims. For example, in the case of online hate speech, we see a growing level of involvement on the part of civil society organisations, internet service providers and the government. That is much less the case when it comes to extreme challenges, disinformation and shaming, for example.

Here, too, internet service providers have a responsibility to do what they can to prevent harm and to assist in recovery. They must take this into account when designing their services, and when harm occurs, take action (see theme 1). In addition to the need for a conversation about online norms (see theme 2), the government has a clear duty to protect people and to assist victims. Based on its position, it can also encourage other stakeholders to afford people better protection and to assist victims.

Based on our study, we have identified three tasks for government in this context:
1) assist victims;
2) be present online;
3) guard the balance between different rights online.

Assist victims

The government can invest in assisting the victims of harmful and immoral behaviour online. At present, victims often find it difficult to tell others what they have been through, especially if it is not obvious that a crime has been committed. While it is important to prosecute possible perpetrators, it is also crucial that victims feel they are being taken seriously and listened to. This study shows that private helplines, such as those run by Dutch foundations tackling online shaming or child abuse, are in close contact with the platforms. The police can learn from such initiatives. The government should also consider the following two ways of assisting victims.

Set up a national helpline for reporting online abuse
The government can itself set up a national helpline for victims of immoral and harmful behaviour online. The internet discrimination helpline MiND emerged in this study as a useful model for helping victims and curbing harmful and immoral behaviour online. MiND brings in public prosecutors to assess whether a report of discrimination involves unlawful behaviour. In many cases, the platform concerned is then asked to remove the reported content. If it does not do so, the courts may be asked to issue a takedown order. Meldknop.nl (see Chapter 5) already operates as a portal where children and their carers can contact professional organisations that specialise in various phenomena listed in the taxonomy. The models provided by MiND and Meldknop.nl could be extended and scaled up to also cover other phenomena identified in this study and to make the resulting helpline more widely available. A national helpline can also be useful for registering harmful and immoral behaviour online, giving society a better idea of the nature and scale of this problem in the Netherlands. A helpline would also benefit from effective cooperation with online platforms, as they are in a position to limit the impact of harmful and immoral behaviour online. That is also in the best interest of the 'author' of content that is removed. The procedures must therefore also include appeal and remedy mechanisms.

Listen to victims and register all reports
As we saw in the Online Shaming case, it is crucial for victims' sense of justice that they feel the authorities are listening to them. We have noted that victims of immoral and harmful behaviour online often do not report it. This only adds to the impression of lawlessness on the internet. Listening to victims and registering their complaints can improve the quality of the assistance they receive.

Be present online

If people come across an unsafe situation on the street, they generally know what to expect from the authorities. Police and other emergency services will assess the situation and take action after receiving a report. But when people suffer harm on the internet, they do not always know what the authorities can do to help them. This adds to the victims' sense of online lawlessness and perceived failure of the justice system. Having a government with a more pronounced online presence may help. Our literature review and interviews with experts yielded the following options for action in this context:

Online youth social work and community policing
Youth social workers and community police officers who are active online understand what children and teenagers do there and can act to prevent harmful

and immoral behaviour, both offline and online. During the Covid-19 crisis, local officers built positive relationships with youngsters through online gaming. This allowed them to detect escalations and prevent them from getting out of hand. The government can encourage civil society and social welfare organisations to become more active online and to target groups other than youths. They could make their presence felt in social media platforms to let victims know they are there for them. It is, of course, important that social workers and police officers act transparently and that their authority to do so is clear.

Disruptive action by the police
The police can take disruptive action to minimise the impact of harmful and immoral behaviour. They are already empowered to do so in the event of possible criminal offences, including cybercrime. They can frustrate criminals or shut down platforms. Interventions of this kind could also be effective in combating other phenomena, such as sock puppeting or online discrimination. The government must, however, be well aware of the tension between different rights online.

Visibility of the police in the online environment
Increasing the visibility of the police online could help to reduce online harm. Our interviews with experts have shown that the mere presence online of police adverts can make cybercriminals reconsider their behaviour. The physical environment does not seem lawless because of the police presence on the streets. By claiming space online, for example with adverts or arrangements with platforms, the police can remind online users that the internet is not a lawless environment.

Accessibility of social welfare services in the online environment
Victims of harmful behaviour online should be able to find help easily, including online. The website Meldknop.nl, in which government also participates, is a low threshold ‘gateway’ to organisations that can offer such help. It is, however, aimed exclusively at children, and does not cover information manipulation or self-harm. A portal that offers access to experts on a broad range of other phenomena might also be useful for adult victims, if properly promoted.

Guard the balance between different rights online

The friction between different rights online has become a growing part of the public debate in recent years. The major online platforms are doing more content moderation than ever and governments worldwide are struggling with the power platforms have over online content. At the same time, human rights organisations are voicing their concerns about the growing influence of authoritarian regimes on internet freedom. The government must balance different rights online to protect the

public from harmful and immoral behaviour there while also safeguarding their freedoms. Platforms themselves also ask governments to regulate and take a stand in that respect. In practical terms, this means, above all else, that the government must not distance itself from the public debate about human rights online and content moderation.

Content moderation is meant to protect users from harmful and immoral content online, but it can also involve a serious violation of fundamental rights. Users often have no say in online platforms' 'community standards' and cannot simply switch to an alternative because of the platforms' market dominance. Content moderation differs from platform to platform and is not always in line with prevailing cultural and societal norms, since many platform companies are not based in the Netherlands. Through regulatory measures and dialogue, the government can ensure that platforms take users' rights seriously in their content moderation decisions. Based on our study, we can suggest a number of ways that the national government can guard the balance between different human rights online.

Assert democratic control over content moderation
At present, large online platforms determine what is and is not considered harmful and immoral on the internet. American norms concerning harmfulness and immorality thus largely determine what Dutch users get to see in online environments. Platforms are themselves growing uneasy with this role. In May 2021, the Rathenau Instituut wrote to the Dutch House of Representatives that independent public oversight would be one way of preventing content moderation from resting unilaterally on the platforms' shoulders. The government can raise this point during the EU policy discussion concerning the Digital Services Act.

Do not focus solely on illegal content
Focusing solely on illegal content does not protect the public adequately against the various forms of harmful and immoral behaviour that have been identified in this study. New EU legislation such as the Digital Services Act appears to address only the removal of illegal content, not harmful content. Our study shows that although many online behaviours may have an illicit component, this is not always clear to either the victims or the perpetrators. When does online shaming become defamation? And when is quackery prohibited? The lack of clarity about the boundaries of online behaviour means that victims do not always feel protected by existing legal frameworks. There is no ready-made solution because we, as a society, have only recently started grappling with the question of how to properly protect people's rights online. Freedom of expression can clash with the right of people, and minorities in particular, to move about freely and safely online. The government must not shy away from this dilemma and must continue to examine

how best to address it. A narrow focus on removing illegal online content leaves the public with too little protection.

Improve complaint procedures for online platforms to ensure legal certainty and equality
As our study shows, by no means all platforms offer transparent and consistent complaints procedures, with small platforms in particular falling short. The government can use the DSA negotiations to impose more requirements on platforms' terms of use and complaint and redress procedures. Platforms should have expeditious and transparent mechanisms that allow for collective complaint procedures by larger groups. Similar measures have been incorporated into the draft Online Safety Bill being discussed in the UK, which could serve as inspiration.

Exercise restraint in content moderation
Exercise restraint when it comes to technical content moderation systems. Do not simply leave moderation of the online environment to artificial intelligence alone, but do it in dialogue with society. Algorithmic content moderation can only serve as a stand-alone method in a limited number of cases, such as the removal of child abuse images. Human rights organisations often consider such upload filters as an undue encroachment on freedom of expression. The government should therefore be wary of promoting technical 'quick fixes' for content moderation, especially since it does nothing to address the mechanisms behind harmful and immoral behaviour online, but also because we, as a society, have yet to have a conversation about online norms.

6.4 Theme 4: Strengthening the adaptive capacity of society

Stakeholders in the government, the private sector and civil society do react to harmful and immoral behaviour, but in reacting, they run the risk of being constantly overtaken by events. The internet has only been around for a few decades, but in that short time it has facilitated many new forms of behaviour. However, harmful and immoral online behaviour often only becomes visible when it reaches a critical mass. Tackling such behaviour requires a more proactive, adaptive and preventive process design to ensure a future-proof strategy.

Based on the outcomes of our study and the observation that phenomena and mechanisms behind harmful and immoral behaviour are in a constant state of flux, the Rathenau Instituut has identified two ways in which the government can tackle harmful and immoral behaviour online in the long term:

1) coordinate the exchange of expertise about harmful and immoral behaviour online;
2) facilitate networks of stakeholders.

Coordinate the exchange of expertise

This study has shown that our knowledge of the phenomena in Chapter 3 is still incomplete and that the stakeholders are highly diverse. It is the first to develop a taxonomy of harmful and immoral behaviour online. Now that this taxonomy is available, it has become possible to create networks of experts and social welfare organisations and to continuously monitor, add to and tweak the phenomena that the taxonomy contains.

The government can play a coordinating role in bringing together and organising the relevant knowledge and expertise. Options for doing so are as follows.

Appoint a knowledge coordinator
The task of the ‘knowledge coordinator for harmful and immoral conduct online’ will be to continuously concentrate expertise or gather information on the nature and scale of online harm, to identify important trends and developments, and to advise policymakers accordingly. Systematically collecting data on the various phenomena will also make it easier to identify phenomena as urgent and prioritise them at specific times. It is important, however, that limits are placed on collecting behavioural data online to avoid unnecessary surveillance or intrusion into people’s privacy.

Promote research programmes and cooperation
The government can promote the establishment of research programmes and cooperation with research institutes so as to better understand the relevant mechanisms and phenomena and to shed more light on who the victims and perpetrators are. Working with specific target groups can help to clarify what motivates them, what renders them vulnerable and what would support them, and to subsequently develop customised programmes for them.

Using our taxonomy of immoral and harmful behaviour online, the government could draw up a knowledge-building agenda to gain a better understanding of who the perpetrators and victims of each phenomenon are, and of the underlying social factors. Many of the phenomena identified in this study have a disproportionate effect on certain groups in society. This applies, for example, to women and minorities in the case of hate speech and threats, and to young people in the case of cyberbullying and shame-sexting. It must be said, however, that there is already

an abundance of research focusing specifically on children and adolescents, and that researchers may risk losing sight of adults. It is therefore important for the government to understand which groups in society suffer disproportionately from online harmful and immoral behaviour. To this end, it could collaborate with platforms and civil society organisations to obtain these data and to adapt policy and victim assistance accordingly.

Build capacity and invest in professional expertise
The government can invest in capacity-building and the expertise of policy makers, law enforcement, social workers and other professionals. They need tools to understand and respond to the phenomena and mechanisms of harmful and immoral behaviour online. Investing in human resources, specialisation and cooperation may help to better address the challenges associated with harmful and immoral behaviour online. By allowing professionals to determine for themselves which approach works and to discuss this with one another, we can capitalise on their experience and professionalism.

Facilitate networks of stakeholders

The taxonomy of phenomena (Chapter 3) covers a multitude of domains that intersect with the government's responsibility, and therefore with the work of many policymakers and staff spread across different ministries. For example, officials from the Ministry of Justice and Security, Economic Affairs, the Interior, and Education, Culture and Science were involved in this study because their work involves them in such related topics as disinformation, media literacy or the protection of public space.

In addition, numerous organisations active in the Netherlands are building expertise on immoral and harmful behaviour online or the underlying mechanisms. We encountered many of these stakeholders in Chapter 5, including organisations that advise internet users and help them arm themselves against harmful behaviour. In addition, there are organisations that mobilise users to promote desirable behaviour, for example by designing the internet to encourage and reward desirable behaviour (for example of *upstanders*). There are also collectives that stand up for the victims of certain phenomena, organisations that put pressure on platforms to take action against problematic behaviour or to protect their users from it, and lobbyists that pressure the authorities for stricter regulation. In addition, online platforms naturally play a pivotal role in disseminating harmful content, and commercial products are being developed to protect users and businesses from harmful phenomena.

The stakeholders involved in harmful and immoral online behaviour are highly diverse. They could all cooperate much more closely and in so doing, improve their capacity to adapt and respond quickly to new developments. The government can play a coordinating and supportive role in this regard. Specific options for doing so are the following.

Invest in a network of civil servants
The government could invest in a network of civil servants who deal with specific aspects of harmful and immoral behaviour online. This study has already set such a process in motion by involving officials from various ministries. Connecting them allows them to learn from one another and integrate their policies more effectively.

Encourage cooperation between platforms and civil society
The government could encourage cooperation between platforms and civil society through funding or other programmes. Social welfare organisations can actively provide assistance on social media platforms if the content gives cause to do so. In the Netherlands, for example, content related to suicide appearing in the traditional media and or on social media is accompanied by a reference to the national suicide prevention hotline. Having platforms cooperate more with civil society institutions allows them to lower the threshold to online assistance. Examples include platforms providing information on LGBTQ+ organisations alongside homophobic content or collaborating with mental health advocacy organisations. Doing this may help victims who have encountered immoral and harmful behaviour online.

Coordinate cooperation between regulators
The government could coordinate cooperation between regulators to tighten up and improve oversight of the phenomena and mechanisms of harmful and immoral behaviour online. Oversight of harmful and immoral conduct online falls under the purview of various regulators, such as the Dutch Authority for Consumers and Markets, the Dutch Advertising Code Committee and the Dutch Data Protection Authority. Cooperation between regulators appears to be vital for coordinating and intensifying oversight of the various aspects of harmful and immoral behaviour.

6.5 Conclusion

This report proposes a taxonomy of harmful and immoral behaviour online showing 22 behavioural phenomena in context (Chapter 3). It also systematically identifies which characteristics and mechanisms of the internet play a role in initiating, facilitating and amplifying harmful and immoral online behaviour (Chapter 4). Our review of existing interventions suggests that little has been done to change these mechanisms for the better (Chapter 5). That is why in the present chapter, the

Rathenau Instituut identified four strategic themes and tasks that will enable the government, in cooperation with the private sector and society, to take more targeted action against derailment and to promote moral and desirable behaviour online.

The first theme – Redesigning the online environment – contains tools for the Dutch national government to change the online mechanisms that characterise the internet for the better. The second theme – Clarifying online norms – makes recommendations for updating social agreements on norms and values online. The third theme – Protecting people and assisting victims – offers suggestions for the Dutch national government and its executive agencies to better respond to the phenomena of harmful and immoral behaviour online and the harm they cause. The fourth theme – Strengthening adaptive capacity – offers suggestions for the Dutch national government to gain and maintain a grip on harmful and immoral online behaviour, which is constantly changing. These options for action are aimed at future-proofing the strategic agenda.

Though the internet was once a self-regulating domain, it now requires a more active role on the part of the government. The problems online are urgent and the harm is real, platforms ask for clear regulation, and society deserves support and protection. It is up to various ministries, law enforcement and executive agencies to take action. Opportunities to do so are opening up at the time of writing, for example in the government's digitalisation strategy that is set for 2022 and as part of the ongoing discussions of the EU's policy frameworks. The Rathenau Instituut hopes that this study will help the Netherlands to formulate a future-proof approach to harmful and immoral behaviour online.

Bibliography

4chan. (2021). *Rules*. Https://www.4chan.org/rules

Advisory Council on International Affairs/AIV. (2020). *Regulating online content Towards a Recalibration of the Netherlands' Internet Policy*. Advisory report 113, 24 June 2020. AIV.

AFM & DNB. (2018). *Cryptos: Recommendations for a regulatory framework*. Autoriteit Financiële Markten and De Nederlandsche Bank

Afuah, A. (2013). Are network effects really all about size? The role of structure and conduct. *Strategic Management Journal*, *34*(3), 257–273. https://doi.org/10.1002/smj.2013

Aiken, M. (2016). *The Cyber Effect: A Pioneering Cyberpsychologist Explains How Human Behaviour Changes Online*. John Murray Press.

Aiken, M., Mc Mahon, C., Haughton, C., O'Neill, L., & O'Carroll, E. (2016). A consideration of the social impact of cybercrime: examples from hacking, piracy, and child abuse material online. *Contemporary Social Science*, *11*(4), 373–391. https://doi.org/10.1080/21582041.2015.1117648

Algemeen Dagblad. (2021, 20 April). *Omstreden Twitter-alternatief Parler keert terug in appwinkel*. Algemeen Dagblad. https://www.ad.nl/tech/omstreden-twitter-alternatief-parler-keert-terug-in-appwinkel~a169770b/

Alimoradi, Z., Lin, C.-Y., Broström, A., Bülow, P. H., Bajalan, Z., Griffiths, M. D., Ohayon, M. M., & Pakpour, A. H. (2019). Internet addiction and sleep problems: A systematic review and meta-analysis. *Sleep Medicine Reviews*, *47*, 51–61. https://doi.org/10.1016/j.smrv.2019.06.004

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539. https://doi.org/10.1126/sciadv.aay3539

Amnesty International. (2017, 20 November). *Amnesty reveals alarming impact of online abuse against women*. Amnesty International. https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/

Auxier, B. E., & Vitak, J. (2019). Factors Motivating Customization and Echo Chamber Creation Within Digital News Environments. *Social Media + Society*, *5*(2), 2056305119847506. https://doi.org/10.1177/2056305119847506

Bakker, A. (2021, 9 April). *Grapperhaus wil retweeten van privégegevens van agenten strafbaar stellen*. https://www.limburger.nl/cnt/dmf20210409_97571764

Bantema, W., Twickler, S. M. A., Munneke, S. A. J., Duchateau, M., & Stol, W. Ph. (2018). *Handhaving van de openbare orde door bestuurlijke maatregelen in een digitale wereld*. Sdu. https://www.politieenwetenschap.nl/publicatie/politiewetenschap/2018/burgemeesters-in-cyberspace-313/

Basak, R., Surah, S., Ganguly, N., & Ghosh, S. (2019). Online Public Shaming on Twitter: Detection, Analysis, and Mitigation. *IEEE Transactions on Computational Social Systems*, *6*(2), 208–220. https://doi.org/10.1109/TCSS.2019.2895734

Bats, J. (2019). *The moral matter of an interactive online domain: a philosophical and empirical exploration of how our relation with the online domain mediates online morality*. University of Twente.

Bellingcat. (2021, 7 January). *The Making of QAnon: A Crowdsourced Conspiracy*. Bellingcat. https://www.bellingcat.com/news/americas/2021/01/07/the-making-of-qanon-a-crowdsourced-conspiracy/

Benton, J. (2021, 22 April). Facebook is going to ask you more often what you want in your News Feed. *Nieman Lab*. https://www.niemanlab.org/2021/04/facebook-is-going-to-ask-you-more-often-what-you-want-in-your-news-feed/

Bertrand, N. (2017, 1 November). *Russia organized 2 sides of a Texas protest and encouraged 'both sides to battle in the streets'*. Business Insider. https://www.businessinsider.nl/russia-trolls-senate-intelligence-committee-hearing-2017-11

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs Conspiracy: Collective Narratives in the Age of Misinformation. *PLOS ONE*, *10*(2), e0118093. https://doi.org/10.1371/journal.pone.0118093

Bhikhie, A. (2020, 29 juni). *CU en GroenLinks dienen hatecrimewet in om discriminatie harder te straffen*. NU.nl. https://www.nu.nl/politiek/6061053/cu-en-groenlinks-dienen-hatecrimewet-in-om-discriminatie-harder-te-straffen.html

Billingham, P., & Parr, T. (2019). Online Public Shaming: Virtues and Vices. *Journal of Social Philosophy*, *51*(3), 371–390. https://doi.org/10.1111/josp.12308

Billingham, P., & Parr, T. (2020). Enforcing social norms: The morality of public shaming. *European Journal of Philosophy*, *28*(4), 997–1016. https://doi.org/10.1111/ejop.12543

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New Media & Society*, *21*(11–12), 2589–2606. https://doi.org/10.1177/1461444819854731

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. In *Proc. ACM Hum.-Comput. Interact.* (Vol. 1, Issue CSCW, Nov. 2017, Article No. 24, pp 1-19). https://doi.org/10.1145/3134659

Bliuc, A.-M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, *87*, 75–86. https://doi.org/10.1016/j.chb.2018.05.026

Bond, S. (2021, 14 May). *Just 12 People Are Behind Most Vaccine Hoaxes On Social Media, Research Shows*. NPR. https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes

Borra, E., Niederer, S., Preuß, J., & Weltevrede, E. (2017). *Mapping troll-like practices on twitter*. https://dare.uva.nl/search?identifier=c9824b9e-e0e5-4342-83c7-9f5237727f16

Bouma, R. (2020, 25 September). *Amerikaanse complottheorie QAnon ook in Nederland in opkomst*. NOS. https://nos.nl/l/2349814

Bouyeure, L. (2020, 29 May). *Op TikTok ligt de pro-anacontent voor het oprapen*. de Volkskrant. https://www.volkskrant.nl/gs-b00bd886

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Branley, D. B., & Covey, J. (2017). Is exposure to online content depicting risky behavior related to viewers' own risky behavior offline? *Computers in Human Behavior*, *75*, 283–287. https://doi.org/10.1016/j.chb.2017.05.023

Brouillette, A. (2020). *Key findings from the 2020 RDR Corporate Accountability Index*. Ranking Digital Rights. https://rankingdigitalrights.org/index2020/key-findings

Brumfiel, G. (2021, 12 May). *For Some Anti-Vaccine Advocates, Misinformation Is Part Of A Business*. NPR. https://www.npr.org/sections/health-shots/2021/05/12/993615185/for-some-anti-vaccine-advocates-misinformation-is-part-of-a-business

Bureau Clara Wichmann. (2020). *Onderzoeksrapport Online Gendered Hate Speech: Civiel procederen tegen online hate speech*. Bureau Clara Wichmann. https://clara-wichmann.nl/content/uploads/2021/02/Onderzoeksrapport-Online-Gendered-Hate-Speech.pdf [LINK IS A 404!]

Bureau Jeugd en Media. (2021). *Wij zijn we*. https://www.bureaujeugdenmedia.nl/. https://www.bureaujeugdenmedia.nl/project/deinternethelden

Burris, C. T., & Leitch, R. (2018). Harmful fun: Pranks and sadistic motivation. *Motivation and Emotion*, *42*(1), 90–102. https://doi.org/10.1007/s11031-017-9651-5

CBS News. (2017, 17 October). *More than 12M 'Me Too' Facebook posts, comments, reactions in 24 hours*. CBS News. https://www.cbsnews.com/news/metoo-more-than-12-million-facebook-posts-comments-reactions-24-hours/

CBS. (2018). *Digitale Veiligheid & Criminaliteit 2018* [Webpagina]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/publicatie/2019/29/digitale-veiligheid-criminaliteit-2018

CBS. (2019a). *1,2 miljoen slachtoffers van digitale criminaliteit* [Webpagina]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/nieuws/2019/29/1-2-miljoen-slachtoffers-van-digitale-criminaliteit

CBS. (2019b). *Internet. Nederland langs de Europese meetlat.* Centraal Bureau voor de Statistiek. https://longreads.cbs.nl/europese-meetlat-2019/internet/

CBS. (2019c). *Veiligheidsmonitor 2019* [Webpagina]. Centraal Bureau voor de Statistiek. https://doi.org/10/veiligheidsmonitor-2019

CBS. (2020a). *Online seksuele intimidatie - Prevalentiemonitor Huiselijk Geweld en Seksueel Geweld 2020* [Webpagina]. Centraal Bureau voor de Statistiek. https://longreads.cbs.nl/phgsg-2020/online-seksuele-intimidatie

CBS. (2020b, 1 April). *453 duizend Nederlanders hadden in 2019 thuis geen internet*. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/nieuws/2020/14/453-duizend-nederlanders-hadden-in-2019-thuis-geen-internet

CBS. (2021). *CBS Statline*. Centraal Bureau voor de Statistiek. https://opendata.cbs.nl/#/CBS/nl/dataset/83095NED/table

Cerniglia, L., Zoratto, F., Cimino, S., Laviola, G., Ammaniti, M., & Adriani, W. (2017). Internet Addiction in adolescence: Neurobiological, psychosocial and clinical issues. *Neuroscience & Biobehavioral Reviews*, *76*, 174–184. https://doi.org/10.1016/j.neubiorev.2016.12.024

Champlin, E. (1998). Nero Reconsidered. *New England Review*, *19*(2), 97–108.

Chen, W., & Thorson, E. (2021). Perceived individual and societal values of news and paying for subscriptions. *Journalism*, *22*(6), 1296–1316. https://doi.org/10.1177/1464884919847792

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1217–1230. https://doi.org/10.1145/2998181.2998213

Chia, D. X. Y., & Zhang, M. W. B. (2020). A Scoping Review of Cognitive Bias in Internet Addiction and Internet Gaming Disorders. *International Journal of Environmental Research and Public Health*, *17*(1), 373. https://doi.org/10.3390/ijerph17010373

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions: 'On the Internet, Nobody Knows You’re a Dog'. *Computers in Human Behavior*, *23*(6), 3038–3056. https://doi.org/10.1016/j.chb.2006.09.001

Cocking, D., & van den Hoven, J. (2018). *Evil Online*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119471219

Commissariaat voor de media. (2019). *Filterbubbels in Nederland*. Commissariaat voor de media.

Common, M., & Kleis Nielsen, R. (2021, 19 February). *How to respond to disinformation while protecting free speech*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/risj-review/how-respond-disinformation-while-protecting-free-speech

COMPACT Education Group. (2020). *Guide to Conspiracy Theorists*. https://conspiracytheories.eu/_wpx/wp-content/uploads/2020/03/COMPACT_Guide-2.pdf

Costa, E., & Halpern, D. (2019). *The behavioural science of online harm and manipulation, and what to do about it: An exploratory paper to spark ideas and debate*. Behavioural Insights Team. https://www.bi.team/publications/the-behavioural-science-of-online-harm-and-manipulation-and-what-to-do-about-it/

Council of Europe. (2021). *Hate Speech*. Council of Europe. https://www.coe.int/en/web/freedom-expression/hate-speech

Coyne, I., Chesney, T., Logan, B., & Madden, N. (2009). Griefing in a Virtual Community: An Exploratory Survey of Second Life Residents. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(4), 214–221. https://doi.org/10.1027/0044-3409.217.4.214

Crisp Thinking. (2021). *We understand and identify emerging threats from online groups*. Crisp Thinking. https://www.crispthinking.com/our-approach/

Danish Institute for Human Rights. (2020). *Introduction to human rights impact assessment*. The Danish Institute for Human Rights. https://www.humanrights.dk/tools/human-rights-impact-assessment-guidance-toolbox/introduction-human-rights-impact-assessment

Davenport, T. H., & Beck, J. C. (2001). *The attention economy: understanding the new currency of business*. Harvard Business School Press.

De Vries, A. (2018, 23 April). Opsporen? Doe het zelf! *Social Media DNA*. https://socialmediadna.nl/opsporen-doe-het-zelf/

De Vries, N. (2021, 22 February). *Digital corpses: waarom mensen online naar lijken kijken*. Trouw. https://www.trouw.nl/religie-filosofie/digital-corpses-waarom-mensen-online-naar-lijken-kijken~b13b49a3/

DeGoedeZaak. (n.d.). *Stop Shaming!* DeGoedeZaak. https://campagnes.degoedezaak.org/campaigns/stopshaming

Dehue, F., Bolman, C., Vollink, T., & Pouwelse, M. (2012). Cyberbullying and traditional bullying in relation to adolescents' perception of parenting. *Journal of CyberTherapy and Rehabilitation*, *5*(1), 25–34.

Delgado-López, P. D., & Corrales-García, E. M. (2018). Influence of Internet and Social Media in the Promotion of Alternative Oncology, Cancer Quackery, and the Predatory Publishing Phenomenon. *Cureus*, 1–11. https://doi.org/10.7759/cureus.2617

Denef, S., De Vries, A., Hadjimatheou, K., & Roosendaal, A. (2017). *DIY policing*. Fraunhofer IAO.

Department for Digital, Culture, Media and Sport. (2020). *Safer technology, safer users: The UK as a world-leader in Safety Tech; A Sectoral Analysis of UK Online Safety Technology*. UK Government. https://www.gov.uk/government/publications/safer-technology-safer-users-the-uk-as-a-world-leader-in-safety-tech

Digan, K. (2021, 23 March). *President KNAW: Universiteiten, bescherm je medewerkers*. ScienceGuide. https://www.scienceguide.nl/2021/03/president-knaw-universiteiten-bescherm-je-medewerkers/

Digitale Overheid. (2020, 14 April). *97 procent Nederlanders heeft thuis internet*. Rijksoverheid.nl. https://www.digitaleoverheid.nl/nieuws/97-procent-nederlanders-had-in-2019-thuis-internet/

Dijck, J. van, Poell, T., & Waal, M. de. (2018). *The Platform Society: Public Values in a Connective World*. Oxford University Press.

Diresta, R. (2018, 30 augustus). Free Speech Is Not the Same As Free Reach. *Wired*. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/

Döpfner, M. (2021, 27 January). *It's time for Europe to take private data from the hands of powerful tech monopolies and give it back to the people*. *Business Insider*. https://www.businessinsider.com/big-tech-private-data-facebook-google-apple-europe-eu-2021-1

Duin, R. J. (2020, 9 March). *Eenmaal online gaat de foto van een verdachte nooit meer weg*. *Het Parool*. https://www.parool.nl/nieuws/eenmaal-online-gaat-de-foto-van-een-verdachte-nooit-meer-weg~ba172078/

ECP. (2021, 9 February). Internet blijkt lichtpuntje in coronajaar: geen toename van negatieve online ervaringen onder jongeren. *ECP | Platform voor de InformatieSamenleving*. https://ecp.nl/actueel/internet-blijkt-voor-jongeren-lichtpuntje-in-coronajaar/

ECRI. (2019). *ECRI report on the Netherlands* (p. 65). Council of Europe.

Edunov, S., Bhagat, S., Burke, M., Diuk, C., & Onur Filiz, I. (2016, 4 February). Three and a half degrees of separation. *Facebook Research*. https://research.fb.com/blog/2016/02/three-and-a-half-degrees-of-separation/

Eindhovens Dagblad. (2019, 17 December). *Boete voor arts uit Eindhoven vanwege reclame*. ed.nl. https://www.ed.nl/eindhoven/boete-voor-arts-uit-eindhoven-vanwege-reclame~ab38a423/

Ellemers, N., Van Der Toorn, J., & Paunov, Y. (2019). The Psychology of Morality: A Review and Analysis of Empirical Studies Published From 1940 Through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. https://doi.org/10.1177/1088868318811759

EOKM. (2020). *Factsheet Grooming*. Expertisebureau Online Kindermisbruik. https://www.eokm.nl/wp-content/uploads/2020/08/EOKM-Factsheet-Grooming_update_aug_2020v2.pdf

Espinoza, J. (2020, 11 november). *Former White Congressional Candidate Tweets 'I'm a Black Gay Guy,' Explanation Leads to Bizarre Series of Events*. Complex. https://www.complex.com/life/2020/11/white-gop-candidate-dean-browning-tweets-im-a-gay-black-guy-bizarre-explanation

European Institute for Gender Equality. (2016). *sexism*. European Institute for Gender Equality. https://eige.europa.eu/thesaurus/terms/1367

European Parliament and Council. (2000, 8 June). *Directive on electronic commerce*. Official Journal L 178. 17/07/2000. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0031

Eurostat. (2021). *Individuals - internet activities*. https://ec.europa.eu/eurostat/databrowser/product/page/ISOC_CI_AC_I

Facebook. (2019). *Rules Enforcement*. https://transparency.twitter.com/en/reports/rules-enforcement.html#2019-jan-jun

Facebook. (2021). *Community Standards*. Facebook. Https://www.facebook.com/communitystandards/.

Faddoul, M., Chaslot, G., & Farid, H. (2020). A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. *arXiv:2003.03318 [cs]*. http://arxiv.org/abs/2003.03318

Fox, J., Cruz, C., & Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, *52*, 436–442. https://doi.org/10.1016/j.chb.2015.06.024

Agence France-Presse. (2021, 23 January). *Italy blocks TikTok for certain users after death of girl allegedly playing 'choking' game*. The Guardian. http://www.theguardian.com/world/2021/jan/23/italy-blocks-tiktok-for-certain-users-after-death-of-girl-allegedly-playing-choking-game

Freckelton QC, I. (2020). COVID-19: Fear, quackery, false representations and the law. *International Journal of Law and Psychiatry*, *72*, 101611. https://doi.org/10.1016/j.ijlp.2020.101611

Furnell, S. (2009). Hackers, viruses and malicious software. In *Handbook of internet crime* (pp. 173–193). Willan.

Gabszewicz, J. J., Laussel, D., & Sonnac, N. (2001). Press advertising and the ascent of the 'Pensée Unique'. *European Economic Review*, *45*(4–6), 641–651. https://doi.org/10.1016/S0014-2921(01)00139-8

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering Online Hate Speech* (p. 71). UNESCO. http://en.unesco.kz/countering-online-hate-speech

Gardner, H., & Davis, K. (2013). *The App Generation: How Today's Youth Navigate Identity, Intimacy, and Imagination in a Digital World*. Yale University Press. https://www.jstor.org/stable/j.ctt5vm7dh

Gebiedonline. (2021). *Ons platform*. Gebiedonline. Https://gebiedonline.nl/ons-platform

Geerts, G., & Den Boon, C. A. (1999). *Van Dale Groot woordenboek van de Nederlandse taal* (13de dr.). Van Dale Uitgevers.

Gelfert, A. (2018). Fake News: A Definition. *Informal Logic*, *38*(1), 84–117. https://doi.org/10.22329/il.v38i1.5068

Gerrard, Y. (2020, 3 September). TikTok Has a Pro-Anorexia Problem. *Wired*. https://www.wired.com/story/opinion-tiktok-has-a-pro-anorexia-problem/

Ghaffary, S. (2021, 13 May). *How angry Apple employees' petition led to a controversial new hire's departure*. Vox. https://www.vox.com/recode/2021/5/13/22435266/apple-employees-petition-controversial-antonio-garcia-martinez-new-hire-departure

Goldsmith, A., & Brewer, R. (2015). Digital drift and the criminal interaction order. *Theoretical Criminology*, *19*(1), 112–130. https://doi.org/10.1177/1362480614538645

Grant, H. (2020a, 10 December). *Pornhub to ban unverified uploads after child abuse content claims*. The Guardian. http://www.theguardian.com/global-development/2020/dec/10/pornhub-to-ban-unverified-uploads-after-child-abuse-content-claims

Grant, H. (2020b, 15 December). *How extreme porn has become a gateway drug into child abuse*. The Guardian. http://www.theguardian.com/global-development/2020/dec/15/how-extreme-porn-has-become-a-gateway-drug-into-child-abuse

Griffin, A. (2021, 27 January). *YouTube reveals full scale of coronavirus misinformation on its platform*. The Independent. https://www.independent.co.uk/life-style/gadgets-and-tech/youtube-covid-19-coronavirus-misinformation-b1793492.html

Guan, S. A., & Subrahmanyam, K. (2009). Youth Internet use: risks and opportunities. *Curr Opin Psychiatry*, *22*(4), 351–356. https://doi.org/10.1097/YCO.0b013e32832bd7e0

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), 1–8. https://doi.org/10.1126/sciadv.aau4586

Ha, A. (2020, 29 June). Facebook expands its fan subscription program. *TechCrunch*. https://social.techcrunch.com/2020/06/29/facebook-expands-fan-subscriptions/

Hadjimatheou, K. (2019). Citizen-led digital policing and democratic norms: The case of self-styled paedophile hunters. *Criminology & Criminal Justice*, 1748895819880956. https://doi.org/10.1177/1748895819880956

Harambam, J. (2017). De politisering van de Waarheid. *Sociologie*, *13*(1), 73–92. https://doi.org/10.5117/SOC2017.1.HARA

Haspels-Goudriaan, J. (2020, 16 March). *Geen overval, winkelinbraak of diefstal: Jongeren beginnen criminele loopbaan vaker in cyberspace*. Algemeen Dagblad. https://www.ad.nl/den-haag/geen-overval-winkelinbraak-of-diefstal-jongeren-beginnen-criminele-loopbaan-vaker-in-cyberspace~a61e4bd5/

Heck, W. (2020, 5 November). *Maker aangehouden van'phishing panels'voor grootschalige oplichting*. NRC. https://www.nrc.nl/nieuws/2020/11/05/maker-van-phishing-panels-aanghouden-voor-grootschalige-oplichting-a4018890

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, *34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Hern, A. (2021, 12 May). *Online safety bill 'a recipe for censorship', say campaigners*. The Guardian. https://www.theguardian.com/media/2021/may/12/uk-to-require-social-media-to-protect-democratically-important-content

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for Safety Online: Managing 'Trolling' in a Feminist Forum. *The Information Society*, *18*(5), 371–384. https://doi.org/10.1080/01972240290108186

Herweijer, K., & Ververs, C. (2020, 12 November). *Fenomeen pedojagen groeit: 'We komen even wat vragen stellen vriend'*. Telegraaf. https://www.telegraaf.nl/nieuws/1340573470/fenomeen-pedojagen-groeit-we-komen-even-wat-vragen-stellen-vriend

Hinson, L., Mueller, J., O'Brien-Milne, L., & Wandera, N. (2018). Technology-facilitated gender-based violence: What is it, and how do we measure it? *International Center for Research on Women*, 8.

Hobbs, R., & Grafe, S. (2015). YouTube pranking across cultures. *First Monday*, *20*(7). https://doi.org/10.5210/fm.v20i7.5981

Houtekamer, C., & Wassens, R. (2021, 2 April). *Het afvoerputje van het internet zit in een Noord-Hollands dorp*. NRC. https://www.nrc.nl/nieuws/2021/04/02/het-afvoerputje-van-het-internet-zit-in-een-noord-hollands-dorp-a4038329

Husting, G., & Orr, M. (2007). Dangerous Machinery: 'Conspiracy Theorist' as a Transpersonal Strategy of Exclusion. *Symbolic Interaction*, *30*(2), 127–150. https://doi.org/10.1525/si.2007.30.2.127

Index on Censorship. (2019, 5 April). The UK government's online harms white paper shows disregard for freedom of expression. *Index on Censorship*. https://www.indexoncensorship.org/2019/04/uk-government-online-harms-white-paper-shows-disregard-freedom-expression/

Instagram. (n.d.). *@pedohunterznl*. Instagram. https://www.instagram.com/hunterzprotectnl/

Instagram. (2021). *Continuing to Make Instagram Safer for the Youngest Members of Our Community*. Instagram. Https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community

Ipsos. (2020). *Trust misplaced?* https://www.ipsos.com/en/trust-misplaced

IVIR. (2020). *WODC-onderzoek: Voorziening voor verzoeken tot snelle verwijdering van onrechtmatige online content*. Instituut voor Informatierecht. https://www.ivir.nl/publicaties/download/WODC_voorziening_onrechtmatige_content.pdf

Jellinek. (2021, 21 April). *Hoeveel mensen in Nederland zijn verslaafd en hoeveel zijn er in behandeling?* Jellinek. https://www.jellinek.nl/vraag-antwoord/hoeveel-mensen-zijn-verslaafd-en-hoeveel-zijn-er-in-behandeling/

Juvonen, J., & Gross, E. V. (2008). Extending the School Grounds?—Bullying Experiences in Cyberspace. *Journal of School Health*, *78*(9), 496–505. https://doi.org/10.1111/j.1746-1561.2008.00335.x

Kaakinen, M., Sirola, A., Savolainen, I., & Oksanen, A. (2020). Impulsivity,

internalizing symptoms, and online group behavior as determinants of online hate. *PLOS ONE*, 1-17. https://doi.org/10.1371/journal.pone.0231052

Kafka, P. (2021, 20 April). *Apple will let podcasters sell subscriptions and keep a cut for itself*. Vox. https://www.vox.com/recode/2021/4/20/22394032/apple-podcast-subscription-plans

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin Books.

Kaiser, J., Schmidt, C., Benkler, Y., Tilton, C., Etling, B., Roberts, H., Clark, J., & Faris, R. (2020, 21 oktober). *Mail-In Voter Fraud: Anatomy of a Disinformation Campaign | Berkman Klein Center*. https://cyber.harvard.edu/publication/2020/Mail-in-Voter-Fraud-Disinformation-2020

Kastrenakes, J. (2021, 9 February). *Twitter's Jack Dorsey wants to build an app store for social media algorithms*. The Verge. https://www.theverge.com/2021/2/9/22275441/jack-dorsey-decentralized-app-store-algorithms

Katawazi, G., & Wagemakers, T. (2021, 11 May). *Met 'online straatverbod' hoopt burgemeester Halsema online shamers harder aan te pakken*. AT5. https://www.at5.nl/artikelen/208616/met-online-straatverbod-hoopt-burgemeester-halsema-online-shamers-harder-aan-te-pakken

Khasawneh, A., Madathil, K. C., Dixon, E., Wiśniewski, P., Zinzow, H., & Roth, R. (2020). Examining the Self-Harm and Suicide Contagion Effects of the Blue Whale Challenge on YouTube and Twitter: Qualitative Study. *JMIR Mental Health*, *7*(6), e15973. https://doi.org/10.2196/15973

Kist, R. (2020, 25 September). *Adverteerders financieren valse informatie over corona*. NRC. https://www.nrc.nl/nieuws/2020/07/06/adverteerders-financieren-valse-info-corona-a4005111

Kist, R., & Van den Bos, M. (2021, 8 March). *Hoe Nederlandse complotdenkers en virussceptici sociale media telkens te slim af zijn*. NRC. https://www.nrc.nl/nieuws/2021/03/08/onderzoek-valse-informatie-van-sociale-media-weren-lukt-lang-niet-altijd-a4034651

Kleijer, J. (2015, 16 April). *Shame sexting is een groepsding*. Bureau Jeugd & Media. https://www.bureaujeugdenmedia.nl/shame-sexting-is-een-groepsding/

Kliksafe. (2021). *homepage*. Kliksafe. https://www.kliksafe.nl/

Knieriem, P. (2021, 2 februari). *Politicus in Zeist zou trollen gebruiken om sociale media te beïnvloeden: 'Die Noortje bestaat helemaal niet'*. RTV Utrecht. https://www.rtvutrecht.nl/nieuws/2133571/?fb=true

Kohorst, M. A., Warad, D. M., Nageswara Rao, A. A., & Rodriguez, V. (2018). Obesity, sedentary lifestyle, and video games: The new thrombophilia cocktail in adolescents. *Pediatric Blood & Cancer*, *65*(7), e27041. https://doi.org/10.1002/pbc.27041

Konopka, B. (2021, 15 April). *Gdynia firm's 'Cyber Guardian' leading the way in combatting online violence*. https://www.thefirstnews.com/article/gdynia-firms-cyber-guardian-leading-the-way-in-combatting-hate-speech-online-21268

Kootstra, J. (2020, 10 December). *Sterke stijging anorexiapatiënten die helemaal stoppen met eten en drinken*. https://nos.nl/l/2360082

Kouwenhoven, A., & Logtenberg, H. (2017, 10 February). *Hoe Denk met 'trollen' politieke tegenstanders monddood probeert te maken*. NRC. https://www.nrc.nl/nieuws/2017/02/10/de-trollen-van-denk-6641045-a1545547

Kraak, H. (2020, 22 October). *Hoe een 27-jarige rapper op pedofielen jaagt vanuit zijn huiskamer in Deventer*. de Volkskrant. https://www.volkskrant.nl/columns-opinie/hoe-een-27-jarige-rapper-op-pedofielen-jaagt-vanuit-zijn-huiskamer-in-deventer~bac60f27/

La Morgia, M., Mei, A., Sassi, F., & Stefa, J. (2021). The Doge of Wall Street: Analysis and Detection of Pump and Dump Cryptocurrency Manipulations. *ArXiv:2105.00733 [Cs]*. http://arxiv.org/abs/2105.00733

Laato, S., Islam, A. K. M. N., Islam, M. N., & Whelan, E. (2020). What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems*, *29*(3), 288–305. https://doi.org/10.1080/0960085X.2020.1770632

LaFrance, S. by A. (2020, 15 December). Facebook Is a Doomsday Machine. *The Atlantic*. https://www.theatlantic.com/technology/archive/2020/12/facebook-doomsday-machine/617384/

Lastdrager, E. (2018). *From Fishing to Phishing*. Gildeprint Drukkerijen. https://www.fraudehelpdesk.nl/vergroot-uw-kennis/from-fishing-to-phishing/

Lauckner, C., Truszczynski, N., Lambert, D., Kottamasu, V., Meherally, S., Schipani-McLaughlin, A. M., Taylor, E., & Hansen, N. (2019). 'Catfishing,' cyberbullying, and coercion: An exploration of the risks associated with dating app use among rural sexual minority males. *Journal of Gay & Lesbian Mental Health*, *23*(3), 289–306. https://doi.org/10.1080/19359705.2019.1587729

Levey, T. G. (2018). *Sexual harassment online: shaming and silencing women in the digital age*. Lynne Rienner Publishers, Inc.

Lewis, R. (2018). Literature review on children and young people demonstrating technology-assisted harmful sexual behavior. *Aggression and Violent Behavior*, *40*, 1–11. https://doi.org/10.1016/j.avb.2018.02.011

Linnemann, E., & Melchior, M. (2017, 3 March). *Zo gaan vrouwelijke opiniemakers om met online haat en discriminatie*. De Volkskrant. https://www.volkskrant.nl/wetenschap/zo-gaan-vrouwelijke-opiniemakers-om-met-online-haat-en-intimidatie~b1764a77/

Liu, W., Mirza, F., Narayanan, A., & Souligna, S. (2020). Is it possible to cure Internet addiction with the Internet? *AI & SOCIETY*, *35*(1), 245–255. https://doi.org/10.1007/s00146-018-0858-0

Lorenzo-Dus, N. & Izura, C. (2017). 'cause ur special': Understanding trust and complimenting behaviour in online grooming discourse. *Journal of Pragmatics*, *112*, 67–82. https://isiarticles.com/bundles/Article/pre/pdf/159239.pdf

Lubach, A. (2020). *De online fabeltjesfuik | Zondag met Lubach (S12)*. https://www.youtube.com/watch?v=FLoR2Spftwg

Ludemann, D. (2018). /pol/emics: Ambiguity, scales, and digital discourse on 4chan. *Discourse, Context & Media*, *24*, 92–98. https://doi.org/10.1016/j.dcm.2018.01.010

MacAllister, J. M. (2016). The Doxing Dilemma: Seeking a Remedy for the Malicious Publication of Personal Information. *Fordham Law Review*, *85*, 2451–2483.

MacCarthy, M. (2021, 11 May). The Facebook Oversight Board’s failed decision distracts from lasting social media regulation. *Brookings*. https://www.brookings.edu/blog/techtank/2021/05/11/the-facebook-oversight-boards-failed-decision-distracts-from-lasting-social-media-regulation/

Machkovech, S. (2016, 15 November). *Twitter bots can reduce racist slurs—if people think the bots are white*. Arstechnica. https://arstechnica.com/science/2016/11/twitter-bots-can-reduce-racist-slurs-if-people-think-the-bots-are-white/

Make Media Great Again. (2021). *Doe mee aan onze collaboratieve media beweging!* Make Media Great Again. https://www.mmga.io/

Marwick, A. (2018). Why do people share fake news? A sociotechnical model of media effects. *Georgetown Law Technology Review*, *2*(2), 474–512.

Mendes, K., Ringrose, J., & Keller, J. (2018). #MeToo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies*, *25*(2), 236–246. https://doi.org/10.1177/1350506818765318

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

MiND. (2020). *Discriminatiecijfers in 2019* (p. 88). Art.1. https://www.mindnederland.nl/wp-content/uploads/2020/04/Discriminatiecijfers-in-2019-1.pdf

Ministerie van Justitie en Veiligheid. (2019). *Kamerbrief 2018D24515*. https://zoek.officielebekendmakingen.nl/kst-31015-177.html

Ministerie van Justitie en Veiligheid. (2020, 3 December). *Strafvorderingsrichtlijn misbruik seksueel beeldmateriaal - Nieuwsbericht - Openbaar Ministerie* [Nieuwsbericht]. Openbaar Ministerie. https://www.om.nl/actueel/nieuws/2020/12/03/strafvorderingsrichtlijn-misbruik-seksueel-beeldmateriaal

Ministerie van Justitie en Veiligheid. (2021). *Dadermonitor seksueel geweld tegen kinderen 2015-2019 - Rapport - Nationaal Rapporteur* [Rapport]. https://www.nationaalrapporteur.nl/publicaties/rapporten/2021/06/08/dadermonitor-seksueel-geweld-tegen-kinderen-2015-2019

Mink, I., & Van Bon, S. (2017). Discriminatiecijfers 2016. *Artikel 1*. https://www.art1.nl/publicaties/landelijke-meldingen-discriminatie-2016/

Miserus, M., & Van der Noordaa, R. (2018). *Het trollenleger van Dotan*. De Volkskrant. https://www.volkskrant.nl/kijkverder/2018/dotan/#/

Möller, J., Helberger, N., & Makhortykh, M. (2019). *Filter bubbles in the Netherlands?* https://dare.uva.nl/search?identifier=2d8db249-cb3a-4eae-b514-56897c08a2d6

Montag, C., Yang, H., & Elhai, J. D. (2021). On the Psychology of TikTok Use: A First Glimpse From Empirical Findings. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.641673

Moonshot. (2021). *Redirect Method*. Moonshot. Https://moonshotcve.com/redirect-method/

Morozov, E. (2021, 7 May). *Spoiler: Zo innovatief zijn Apple en Google helemaal niet*. De Correspondent. https://decorrespondent.nl/12349/spoiler-zo-innovatief-zijn-apple-en-google-helemaal-niet/18470148096642-19468a5f

Morris, S. (2021, 3 June). *21 Dangerous TikTok Trends Every Parent Should Be Aware of*. Newsweek. https://www.newsweek.com/21-dangerous-tiktok-trends-that-have-gone-viral-1573734

Mortimer, K. (2017). Understanding Conspiracy Online: Social Media and the Spread of Suspicious Thinking. *Dalhousie Journal of Interdisciplinary Management*, *13*(1). https://doi.org/10.5931/djim.v13i1.6928

Mosley, M. A., Lancaster, M., Parker, M. L., & Campbell, K. (2020). Adult attachment and online dating deception: a theory modernized. *Sexual and Relationship Therapy*, *35*(2), 227–243. https://doi.org/10.1080/14681994.2020.1714577

Motivaction. (2021, 2 September). *Ondanks discussie over nepnieuws: groeiende meerderheid vertrouwt journalistiek wél*. https://www.motivaction.nl/kennisplatform/nieuws-en-persberichten/ondanks-discussie-over-nepnieuws-groeiende-meerderheid-vertrouwt-journalistiek-wel

Movisie. (n.d.). *Campagnepagina #DatMeenJeNiet*. Movisie. Accessed 14 June 2021, https://www.movisie.nl/campagnepagina-datmeenjeniet

Movisie. (2019). *Shame sexting bij tienermeiden met een Marokkaans-islamitische*

achtergrond. Movisie. https://www.movisie.nl/artikel/shame-sexting-tienermeiden-marokkaans-islamitische-achtergrond

Müller, K. W., Janikian, M., Dreier, M., Wölfling, K., Beutel, M. E., Tzavara, C., Richardson, C., & Tsitsika, A. (2015). Regular gaming behavior and internet gaming disorder in European adolescents: results from a cross-national representative survey of prevalence, predictors, and psychopathological correlates. *European Child & Adolescent Psychiatry*, *24*(5), 565–574. https://doi.org/10.1007/s00787-014-0611-2

Multiscope. (2020, 13 February). *Nederlanders gamen dagelijks half miljard minuten*. Multiscope. http://www.multiscope.nl/persberichten/nederlanders-gamen-dagelijks-half-miljard-minuten.html

Munn, L. (2021). More than a Mob: Parler as Preparatory Media for the Capitol Storming. *First Monday*, *26*(3). https://doi.org/10.5210/fm.v26i3.11574

Naughton, J. (2015, 7 February). *Aaron Swartz stood up for freedom and fairness – and was hounded to his death*. The Guardian. http://www.theguardian.com/commentisfree/2015/feb/07/aaron-swartz-suicide-internets-own-boy

NCTV, Ministerie van Justitie en Veiligheid. (2021, 14 April). *Fenomeenanalyse 'De verschillende gezichten van de coronaprotesten' - Publicatie - Nationaal Coördinator Terrorismebestrijding en Veiligheid* [Publicatie]. NCTV. https://www.nctv.nl/documenten/publicaties/2021/04/14/fenomeenanalyse-de-verschillende-gezichten-van-de-coronaprotesten

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society*, *20*(10), 3720–3737. https://doi.org/10.1177/1461444818758715

Net Nanny. (2021). *homepage*. Net Nanny. https://www.netnanny.com/.

Netwerk Mediawijsheid. (2021). *Netwerk Mediawijsheid*. Netwerk Mediawijsheid. https://www.netwerkmediawijsheid.nl

Newton, C. (2021a, 1 April). *Nick Clegg tries to reset the conversation*. Platformer. https://www.platformer.news/p/nick-clegg-tries-to-reset-the-conversation

Newton, C. (2021b, 9 April). *The case of the missing platform policies*. https://www.platformer.news/p/the-case-of-the-missing-platform

Newton, C. (2021c, 16 June). *Why Google's FLoC flopped*. https://www.platformer.news/p/why-googles-floc-flopped

Nikolaou, D. (2017). Does cyberbullying impact youth suicidal behaviors? *Journal of Health Economics*, *56*, 30–46. https://doi.org/10.1016/j.jhealeco.2017.09.009

NJi. (2019). *Eetstoornissen - Cijfers | NJi*. Nederlands Jeugdinstituut. https://www.nji.nl/nl/Databank/Cijfers-over-Jeugd-en-Opvoeding/Cijfers-per-onderwerp/Eetstoornissen

NLProfiel. (2020, 9 September). *NLProfiel – samenwerkende Nederlandse uitgevers op gebied van targeting in digitale media*. https://nlprofiel.nl/

NOS. (2018, 29 July). *Challenge met rijdende auto: 'Volslagen idioot en strafbaar'*. https://nos.nl/l/2243726

NOS. (2021a, 11 January). *Zwarte lijst met namen van artsen verboden door rechter*. https://nos.nl/l/2363920

NOS. (2021b, 6 August). *Hoe de cryptohandel gemanipuleerd wordt: 'Het werkt het best bij onervaren mensen'*. https://nos.nl/l/2379563

noticeandtakedowncode.nl/. (2018). *Gedragscode Notice-and-Take-Down*.

noticeandtakedowncode.nl/. https://noticeandtakedowncode.nl/wp-content/uploads/2018/12/ECP_01054-Gedragscode-notice-and-takedown-pdf-2.pdf

NRC. (2021, 14 January). *Waarom radicaliseren mensen?* NRC. https://www.nrc.nl/nieuws/2021/01/14/gevaar-vernauwt-het-denken-a4027567

Nu.nl. (2020, 18 November). *Vijf minderjarigen aangehouden voor 'happy slapping' in Amsterdam*. NU. https://www.nu.nl/amsterdam/6091203/vijf-minderjarigen-aangehouden-voor-happy-slapping-in-amsterdam.html

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Okuna. (2021). *Home*. Okuna. https://about.okuna.io/en/home.

Oleshchuk, P. (2020). The Instruments of Modern Media Lobbying. *Future Human Image*, *14*, 48–55. https://doi.org/10.29202/fhi/14/6

Oosterveer, D. (2021, 23 January). *Social media in Nederland 2021: TikTok-gebruik door jongeren stijgt explosief en passeert Facebook*. Marketingfacts. https://www.marketingfacts.nl/berichten/social-media-in-nederland-2021

Oosterwijk, K., & Fischer, T. F. C. (2017). *Interventies jeugdige daders cybercrime*. WODC. https://veiligheidscoalitie.nl/action/?action=download&id=2386

Ortiz, S. M. (2019). 'You Can Say I Got Desensitized to It': How Men of Color Cope with Everyday Racism in Online Gaming. *Sociological Perspectives*, *62*(4), 572–588. https://doi.org/10.1177/0731121419837588

Pariser, E. (2012). *The filter bubble: what the Internet is hiding from you*. Penguin Books.

Parler. (2021). *Community guidelines*. https://legal.parler.com/documents/guidelines.pdf

Paul, K. (2021, 5 May). *Facebook ruling on Trump renews criticism of oversight board*. The Guardian. http://www.theguardian.com/technology/2021/may/05/facebook-oversight-board-donald-trump

Paulissen, & Van Wilsem. (2015). *Politie en Wetenschap*. https://www.politieenwetenschap.nl/publicatie/politiewetenschap/2015/dat-heeft-iemand-anders-gedaan-259/

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Peterson, J., & Densley, J. (2017). Cyber violence: What do we know and where do we go from here? *Aggression and Violent Behavior*, *34*, 193–200. https://doi.org/10.1016/j.avb.2017.01.012

Petities.nl. (2021, 31 March). *Stop chemtrails nu en weermanipulatie nu*. Petities.nl. https://petities.nl/petitions/stop-chemtrails-nu-en-weermanipulatie-nu?locale=nl

Phillips, W. (2015). *This is why we can't have nice things: mapping the relationship between online trolling and mainstream culture*. The MIT Press.

Plan International. (2020, 5 October). *Wereldwijd 58 procent van de meisjes slachtoffer van online intimidatie*. Plan International. https://www.planinternational.nl/actueel/wereldwijd-58-procent-van-de-meisjes-slachtoffer-van-online-intimidatie

Pointer. (2021a, 21 March). *De invloed van sociale media op de verkiezingscampagne*. KRO-NCRV. https://pointer.kro-ncrv.nl/de-invloed-van-sociale-media-op-de-verkiezingscampagne

Pointer. (2021b, 20 May). *Nederlands trollenleger verspreidt en coördineert desinformatie over vaccin*. KRO-NCRV. https://pointer.kro-ncrv.nl/nederlands-trollenleger-verspreidt-en-coordineert-desinformatie-over-vaccin

Policy Department for Citizens' Rights and Constitutional Affairs. (2020). *Hate speech and hate crime in the EU and the evaluation of online content regulation approaches*. European Parliament.

Politie, Openbaar ministerie, Regioburgemeesters, & Ministerie van Justitie & Veiligheid. (2020). *Position paper: de politie van morgen en overmorgen - Publicatie - Openbaar Ministerie* [Publicatie]. https://www.om.nl/documenten/publicaties/om-onderdelen/pag-om/map/position-paper-de-politie-van-morgen-en-overmorgen

Pomerantsev, P. (2019). *A Cycle of Censorship: The UK White Paper on Online: Harms and the Dangers of Regulating Disinformation†* (Working papers of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression). Instituut voor Informatierecht. https://www.ivir.nl/publicaties/download/Cycle_Censorship_Pomerantsev_Oct_2019.pdf

Powell, A. (2015). Seeking rape justice: Formal and informal responses to sexual violence through technosocial counter-publics. *Theoretical Criminology*, *19*(4), 571–588. https://doi.org/10.1177/1362480615576271

Prij, J., & Janssens, M. (2020, 12 June). *Nepnieuws: graag een beetje bezonnenheid!* Christen Democratische Verkenningen. https://www.tijdschriftcdv.nl/inhoud/tijdschrift_artikel

Quekel, S. (2021, 15 February). *Nieuwe vorm van identiteitsfraude rukt op: 'Foto's*

verschijnen op pornosite'. Algemeen Dagblad. https://www.ad.nl/tech/nieuwe-vorm-van-identiteitsfraude-rukt-op-foto-s-verschijnen-op-pornosite~aafb7609/

Rabkin, M. (2021, 18 March). *Facebook shows its upcoming social VR avatars for Horizon at SXSW* [CNET]. https://www.cnet.com/news/facebook-shows-its-upcoming-social-vr-avatars-for-horizon-at-sxsw/

Rasch, M. (2021, 18 April). *Zelfs na je dood laat Big Tech je niet met rust*. Follow the Money - Platform voor onderzoeksjournalistiek. https://www.ftm.nl/artikelen/dood-big-tech-deepfake

Raskauskas, J., & Stoltz, A. D. (2007). Involvement in traditional and electronic bullying among adolescents. *Developmental Psychology*, *43*(3), 564–575. https://doi.org/10.1037/0012-1649.43.3.564

Rathenau Instituut. (2018a). *Digitalisering van het nieuws: online nieuwsgedrag en personalisatie in Nederland.* https://www.rathenau.nl/nl/digitale-samenleving/digitalisering-van-het-nieuws

Rathenau Instituut. (2018b). *Public trust in science | Rathenau Instituut*. https://www.rathenau.nl/en/science-figures/impact/trust-science/public-trust-science

Rathenau Instituut. (2020a). *Cyber resilience with new technology*. Rathenau Instituut (authors: Boheemen, P. van, G. Munnichs, L. Kool, G. Diercks, J. Hamer & A. Vos).

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

Rathenau Instituut. (2020c). *Rathenau Manifesto: Set 10 design requirements for tomorrow's digital society now*.

Rathenau Instituut. (2021a). *Reactie Rathenau Instituut op Consultatie AIV advies Regulering van online content*. https://www.rathenau.nl/sites/default/files/2021-03/Reactie_Rathenau_Instituut_Consultatie_AIV_advies_Regulering_online_content.pdf

Rathenau Instituut. (2021b). *De toekomst van online platformen: Twee Europese wetsvoorstellen onder de loep*. https://www.rathenau.nl/sites/default/files/2021-05/Rathenau_Instituut_Bericht_aan_parlement_De_toekomst_van_online_platformen.pdf

Redactie NOS. (2021, 25 February). *Strijden tegen online shaming en expose accounts*. NPO Radio 1. https://www.nporadio1.nl/binnenland/29898-strijden-tegen-online-shaming-en-expose-accounts

Redactie ScienceGuide. (2021, 28 May). *Nog voor zomerreces wetsvoorstel tegen "walgelijke" doxing Vizier op Links*. ScienceGuide. https://www.scienceguide.nl/2021/05/nog-voor-zomerreces-wetsvoorstel-tegen-walgelijke-doxing-vizier-op-links/

Reddit. (2021a). *Reddit Content Policy*. Reddit. https://www.redditinc.com/policies/content-policy

Reddit. (2021b). *Ways to become a moderator*. Https://mods.reddithelp.com/hc/en-us/articles/360001745332-Ways-to-become-a-moderator

Reuters. (2020). *Overview and Key Findings of the 2020 Digital News Report*.

Reuters Institute Digital News Report. https://www.digitalnewsreport.org/survey/2020/overview-key-findings-2020/

Rijksoverheid. (n.d.). *Wraakporno*. Rijksoverheid.nl. Accessed 18 June 2021, https://www.rijksoverheid.nl/onderwerpen/seksuele-misdrijven/wraakporno

ROB. (2019). *Adviesrapport Zoeken naar waarheid - Publicatie - Raad voor het Openbaar Bestuur* [Publicatie]. https://www.raadopenbaarbestuur.nl/documenten/publicaties/2019/05/09/zoeken-naar-waarheid

Rogers, R., & Niederer, S. (2019). *The Politics of Social Media Manipulation*. The Hague: Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

Roose, K. (2020). Rabbit Hole. The New York Times. https://www.nytimes.com/column/rabbit-hole

RTL Nieuws. (2018, 22 May). *Clay (15) overleden door 'onnozele challenge': 'Zijn verlies is afgrijselijk'*. RTL Nieuws. https://www.rtlnieuws.nl/nieuws/nederland/artikel/4200111/clay-15-overleden-door-onnozele-challenge-zijn-verlies

RTL Nieuws. (2019, 28 February). *Bitcoin-oplichters verdienen tonnen aan Nederlandse slachtoffers*. RTL Nieuws.

https://www.rtlnieuws.nl/tech/artikel/4625991/bitcoin-scam-oplichting-broker-trading-investering-4-procent-rendement

Mediacourant. (2021, 19 March). *RTL Nieuws sluit reacties over Sylvana Simons na baggertsunami*. Mediacourant.nl. https://www.mediacourant.nl/2021/03/rtl-nieuws-sluit-reacties-over-sylvana-simons-na-baggertsunami/

RTL Nieuws. (2017, 4 May). *Stel raakt voogdij over kinderen kwijt na schokkende prank-video's op YouTube*. RTL Nieuws. https://www.rtlnieuws.nl/editienl/artikel/100621/stel-raakt-voogdij-over-kinderen-kwijt-na-schokkende-prank-videos-op

RTL Nieuws. (2020, 24 August). *Verslaafd aan je smartphone: 'Het is een ledemaat geworden'*. RTL Nieuws. https://www.rtlnieuws.nl/editienl/artikel/5179060/verslaafd-aan-smartphone-nomofobie-ledemaat-telefoon

RTL Nieuws. (2021, 4 May). *Tot 12 maanden cel voor 'pedojagers' na fatale mishandeling 73-jarige man*. RTL Nieuws. https://www.rtlnieuws.nl/nieuws/nederland/artikel/5228920/pedojagen-arnhem-mishandeling-dood-man-73-uitspraak

Rutgers. (2018, 4 September). *Sexting: Praat met jongeren over de gevaren en risico's*. Sexting: Praat met jongeren over de gevaren en risico's. https://www.rutgers.nl/nieuws-opinie/nieuwsarchief/sexting-praat-met-jongeren-over-de-gevaren-en-risico%E2%80%99s

Rutgers, & Soa Aids Nederland. (2019). *Seks onder je 25e VSO 2019*. Seks onder je 25e. https://seksonderje25e.nl/vso

Sabel, P., & Verhagen, L. (2021, 5 March). *Politici zijn het doelwit van tientallen haattweets per dag – wie zit erachter?* De Volkskrant. https://www.volkskrant.nl/nieuws-achtergrond/politici-zijn-het-doelwit-van-tientallen-haattweets-per-dag-wie-zit-erachter~b6e2744e/

Sánchez Montañés, M. (2021, 14 April). *Big tech cannot crack down on online hate alone*. World Economic Forum. https://www.weforum.org/agenda/2021/04/big-tech-cannot-crack-down-on-online-hate-alone/

Sanders, M. (2021). *Owner Identity and Interdependent Markets: an examination of ownership filters of institutional complexity, coalitional change and value creation in disrupted two sided market categories*. https://repub.eur.nl/pub/135457

Saris, K., & Van de Ven, C. (2021, 3 March). *De online haat dreigt vrouwen uit de politieke arena te verdrijven*. De Groene Amsterdammer. https://www.groene.nl/artikel/misogynie-als-politiek-wapen

Schildkamp, V., & Rodenburg, F. (2021, 21 February). *Begraafplaats Bodegraven doelwit van complotdenkers over pedo-netwerk, ook RIVM deed al aangifte*. Algemeen Dagblad.

SCP. (2016). *Resultaten*. https://www.mediatijd.nl/tijdsbesteding/resultaten

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, *36*(3), 357–370. https://doi.org/10.1177/0165551510365390

Shaikh, F. B., Rehman, M., & Amin, A. (2020). Cyberbullying: A Systematic Literature Review to Identify the Factors Impelling University Students Towards Cyberbullying. *IEEE Access*, *8*, 148031–148051. https://doi.org/10.1109/ACCESS.2020.3015669

Shanahan, J. (2021, 5 March). Support for QAnon is hard to measure — and polls may overestimate it. *Nieman Lab*. https://www.niemanlab.org/2021/03/support-for-qanon-is-hard-to-measure-and-polls-may-overestimate-it/

Shea, V. (1994). *Netiquette* (Ed. 1.0). Albion Books.

Simons, E. I., Nooteboom, F., & Van Furth, E. F. (2020). *De Wereld van Pro-ana Coaches*. https://www.hetckm.nl/nieuws-en-publicaties/pro-ana-coaches-maken-bewust-misbruik-van-meisjes-met-eetstoornis/1

Sipma, T., & Leijsen, E. M. C. van. (2019). Slachtofferschap van online criminaliteit. *Den Haag*. https://repository.wodc.nl/handle/20.500.12832/236

Slecht Nieuws. (2021). *intro*. Slecht Nieuws. https://www.slechtnieuws.nl/#intro

SOS. (2021). *Rechter verbiedt zwarte lijst artsen – Stop Online Shaming* [Stichting Online Shaming]. https://www.stoponlineshaming.org/rechter-verbiedt-zwarte-lijst-artsen/

Ster. (2020, 8 December). *Online adverteren 2,5 jaar na de verscherping van de privacywet: wat zijn de lessen? - Ster reclame* [Blog]. Ster.nl. https://www.ster.nl/nieuws/online-adverteren-2-5-jaar-na-de-verscherping-van-de-privacywet-wat-zijn-de-lessen/

Sternisko, A., Cichocka, A., & Van Bavel, J. J. (2020). The dark side of social movements: social identity, non-conformity, and the lure of conspiracy theories. *Current Opinion in Psychology*, *35*, 1–6. https://doi.org/10.1016/j.copsyc.2020.02.007

Stichting Internet Challenges, W. (2021). *https://www.internetchallenges.nl/de-stichting*. https://www.internetchallenges.nl/de-stichting

Stil, H. (2020, 19 January). *Hoe de smartphone ons leven in kroop*. Het Parool. https://www.parool.nl/nieuws/hoe-de-smartphone-ons-leven-in-kroop~b35ae634/

Stolton, S. (2020, 5 June). EU code of practice on disinformation 'insufficient and unsuitable,' member states say. *Www.Euractiv.Com*. https://www.euractiv.com/section/digital/news/eu-code-of-practice-on-disinformation-insufficient-and-unsuitable-member-states-say/

Stop Hate for For Profit. (2021). *#StopHateForProfit*. Stop Hate For Profit. https://www.stophateforprofit.org/

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Superawesome. (2021). *Homepage*. Superawesome. Https://www.superawesome.com/.

SVDJ. (2021, 18 March). Nick Waters (Bellingcat): 'Geweld tegen journalisten neemt toe'. *SVDJ*. https://www.svdj.nl/nick-waters-bellingcat-geweld-tegen-journalisten-neemt-toe/

The Guardian. (2020, 16 October). *QAnon: a timeline of violence linked to the conspiracy theory*. The Guardian. http://www.theguardian.com/us-news/2020/oct/15/qanon-violence-crimes-timeline

The Independent. (2019, 16 March). *How nonsensical white genocide conspiracy theory cited by alleged gunman is spreading poison around the world*. The Independent. https://www.independent.co.uk/news/world/australasia/new-zealand-christchurch-mosque-attack-white-genocide-conspiracy-theory-a8824671.html

TheirTube. (2021). *homepage*. TheirTube. Http://www.their.tube/.

Tik Tok. (2020 December). *Community Guidelines*. Tik Tok. https://www.tiktok.com/community-guidelines

Tokmetzis, D. (2020, 7 January). *De schaduwzijde van cryptovaluta: er is al voor 15 miljard euro opgelicht en gestolen*. De Correspondent. https://decorrespondent.nl/10826/de-schaduwzijde-van-cryptovaluta-er-is-al-voor-15-miljard-euro-opgelicht-en-gestolen/527193722-4276042b

Tokmetzis, D., & Bol, R. (2020, 3 November). *De macht van bedrijven als Google en Apple is gigantisch. Zo trekken Europa en de VS de teugels aan*. De Correspondent. https://decorrespondent.nl/11732/de-macht-van-bedrijven-als-google-en-apple-is-gigantisch-zo-trekken-europa-en-de-vs-de-teugels-aan/571313204-b1c40942

Tumber, H., & Waisbord, S. (2021). *The Routledge Companion to Media Disinformation and Populism*. Routlegde.

Tweede Kamer. (2018). *Kamerstuk II, 31015, nr. 175*. https://zoek.officielebekendmakingen.nl/kst-31015-175.html

Tweede Kamer. (2020). *Kamerstuk II, 30821, nr. 120*. https://zoek.officielebekendmakingen.nl/kst-30821-120.html

Twitter. (2021). *Rules and policies*. Https://help.twitter.com/en/rules-and-policies#twitter-rules

UK Government. (2019). *Online Harms White Paper*. GOV.UK. https://www.gov.uk/government/consultations/online-harms-white-paper

UK Government. (2020a). *Interim code of practice on online child sexual exploitation and abuse*. UK Government. https://www.gov.uk/government/publications/online-harms-interim-codes-of-practice/interim-code-of-practice-on-online-child-sexual-exploitation-and-abuse-accessible-version

UK Government. (2020b). *Online Harms White Paper: Full government response to the consultation*. GOV.UK. https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response

University of Oxford. (2020, 22 May). *Conspiracy beliefs reduce the following of government coronavirus guidance*. University of Oxford.

https://www.ox.ac.uk/news/2020-05-22-conspiracy-beliefs-reduces-following-government-coronavirus-guidance

Van Baars, L. (2020, 25 September). *De tijd die kinderen achter een scherm doorbrengen, is verdubbeld naar ruim 7 uur per dag*. Trouw. https://myprivacy.dpgmedia.nl/consent?siteKey=w38GrtRHtDg4T8xq&callbackUrl=https%3a%2f%2fwww.trouw.nl%2fprivacy-wall%2faccept%3fredirectUri%3d%252fbinnenland%252fde-tijd-die-kinderen-achter-een-scherm-doorbrengen-is-verdubbeld-naar-ruim-7-uur-per-dag%257ebe89f15d%252f

Van Bemmel, N. (2020, 12 June). *Twitter verwijdert tienduizenden accounts wegens Chinese, Turkse en Russische staatspropaganda*. De Volkskrant. https://www.volkskrant.nl/nieuws-achtergrond/twitter-verwijdert-tienduizenden-accounts-wegens-chinese-turkse-en-russische-staatspropaganda~bedd1f53/

Van de Weijer, S. G. A., Leukfeldt, E. R., & Van Der Zee, S. (2020). *Slachtoffer van onlinecriminaliteit, wat nu? Een onderzoek naar aangiftebereidheid onder burgers en ondernemers*. https://www.politieenwetenschap.nl/publicatie/politiewetenschap/2020/slachtoffer-van-onlinecriminaliteit-wat-nu-356/

Van de Weijer, S. G. A., Leukfeldt, R., & Bernasco, W. (2019). Determinants of reporting cybercrime: A comparison between identity theft, consumer fraud, and hacking. *European Journal of Criminology*, *16*(4), 486–508. https://doi.org/10.1177/1477370818773610

Van den Berg, I. (2021, 5 February). *Journalistiek kost geld. Wie betaalt?* OneWorld. https://www.oneworld.nl/lezen/achtergrond/journalistiek-kost-geld-wie-betaalt/

Van der Poel, R., & Luyendijk, W. (2021, 24 March). *Vader van Nora vraagt zich af: wat als je begint met luisteren naar een meisje met anorexia?* NRC. https://www.nrc.nl/nieuws/2021/03/24/ik-ging-stapje-voor-stapje-mee-en-werd-zo-medeplichtig-a4037178

Van Furth, E., Hemkes, S., & Dingemans, A. (2011). Het fenomeen Pro-ana. *Psychopraktijk*, *3*(5), 35–37. https://doi.org/10.1007/s13170-011-0075-8

Van Houwelingen, K. (2017, 17 August). *'Iedereen zal weten wie deze types zijn'*. De Gelderlander. https://advance-lexis-com.proxy.uba.uva.nl:2443/document/?pdmfid=1516831&crid=6022c8b6-8318-

4a6d-888f-410788641f18&pddocfullpath=%2Fshared%2Fdocument%2Fnews%2Furn%3AcontentItem%3A5P8C-0CK1-DYRY-X16T-00000-00&pdcontentcomponentid=149018&pdteaserkey=sr43&pditab=allpods&ecomp=5bq2k&earg=sr43&prid=8786c04d-cd46-428e-a032-84325da6df62

Van Noort, W. (2020, 6 August). *Je mening is niet slecht, jij bent slecht, waarom online shamen niet werkt*. NRC. https://www.nrc.nl/nieuws/2020/08/06/je-mening-is-niet-slecht-jij-bent-slecht-waarom-online-shamen-niet-werkt-a4008036

Van Rooij, A. J., Schoenmakers, T. M., van den Eijnden, R. J. J. M., & van de Mheen, D. (2012). Online video gameverslaving: verkenning van een nieuw fenomeen. *Tijdschrift voor gezondheidswetenschappen*, *90*(7), 420–426. https://doi.org/10.1007/s12508-012-0146-1

Vanheste, T. (2021, 30 September). *Hoe vult Europa het verlangen naar technologische soevereiniteit in?* Rathenau Instituut. https://www.rathenau.nl/nl/vitale-kennisecosystemen/hoe-vult-europa-het-verlangen-naar-technologische-soevereiniteit

Vayansky, I., & Kumar, S. (2018). Phishing – challenges and solutions. *Computer Fraud & Security*, *2018*(1), 15–20. https://doi.org/10.1016/S1361-3723(18)30007-1

Veldhuis, P., & Ingabire, S. (2021, 2 May). *Het was 'sensatiezucht' en 'dom kuddegedrag', maar de pedojacht had een fatale afloop*. NRC. https://www.nrc.nl/nieuws/2021/05/02/het-was-sensatiezucht-en-dom-kuddegedrag-maar-de-pedojacht-had-een-fatale-afloop-a4042134

Vie publique. (2020, 29 June). *Loi du 24 juin 2020 visant à lutter contre les contenus haineux sur internet*. Vie publique.fr. https://www.vie-publique.fr/loi/268070-loi-avia-lutte-contre-les-contenus-haineux-sur-internet

Vince, G. (2018, 3 April). *Evolution explains why we act differently online*. BBC. https://www.bbc.com/future/article/20180403-why-do-people-become-trolls-online-and-in-social-media

Vinocur, N. (2021, 2 April). *The movement to end targeted internet ads*. POLITICO. https://www.politico.eu/article/targeted-advertising-tech-privacy/

Visser, M. (2020, 15 August). *Een op de tien Nederlanders gelooft dat rond corona vieze spelletjes worden gespeeld*. Trouw. https://www.trouw.nl/binnenland/een-op-

de-tien-nederlanders-gelooft-dat-er-rond-corona-vieze-spelletjes-worden-gespeeld~bd98ce41/

Vogels, E. A. (2021, 13 January). The State of Online Harassment. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Völlink, T., Dehue, F., & Guckin, C. M. (2016). *Cyberbullying: From Theory to Intervention*. Routledge.

Von Piekartz, H. (2020, 15 October). *NCTV: boosheid en ongenoegen over coronamaatregelen leiden vaker tot radicalisering*. De Volkskrant. https://www.volkskrant.nl/nieuws-achtergrond/nctv-boosheid-en-ongenoegen-over-coronamaatregelen-leiden-vaker-tot-radicalisering~baa3e15a/

Von Piekartz, H., & Bahara, H. (2021, 26 March). *Doxing: hoe online dreigementen hun weg vinden naar de fysieke wereld*. https://www.volkskrant.nl/nieuws-achtergrond/doxing-hoe-online-dreigementen-hun-weg-vinden-naar-de-fysieke-wereld~be39da12/

VRT. (2020, 18 August). *Chatgroep die holebi's viseert, is niet enige die oproept tot haat en geweld: 'Er bestaan er tientallen bij ons'*. vrtnws.be. https://www.vrt.be/vrtnws/nl/2020/08/17/chatgroepen-die-aanzetten-tot-homohaat-en-geweld-tegen-lgbt-s/

Wagemakers, T., & Toksöz, Z. (2021, 13 May). *Als die ene seksfoto van lang geleden je blijft achtervolgen*. NRC. https://www.nrc.nl/nieuws/2021/05/13/als-die-ene-seksfoto-van-lang-geleden-je-nog-steeds-achtervolgt-a4043259

Wagner, K. (2020, 9 November). *Facebook Labeled 167 Million User Posts for Covid Misinformation*. Bloomberg. https://www.bloomberg.com/tosv2.html?vid=&uuid=a2f841f0-a987-11eb-8bca-4308fa876329&url=L25ld3MvYXJ0aWNsZXMvMjAyMC0xMS0xOS9mYWNlYm9vay1sYWJlbGVkLTE2Ny1taWxsaW9uLXVzZXItcG9zdHMtZm9yLWNvdmlkLW1pc2luZm9ybWF0aW9u

Weimann, G., & Masri, N. (2020). Research Note: Spreading Hate on TikTok. *Studies in Conflict & Terrorism*. https://doi.org/10.1080/1057610X.2020.1780027

Weinstein, A., & Lejoyeux, M. (2010). Internet Addiction or Excessive Internet Use. *The American Journal of Drug and Alcohol Abuse*, *36*(5), 277–283. https://doi.org/10.3109/00952990.2010.491880

Whittle, H., Hamilton-Giachritsis, C., Beech, A., & Collings, G. (2013). A review of online grooming: Characteristics and concerns. *Aggression and Violent Behavior*, *18*(1), 62–70. https://doi.org/10.1016/j.avb.2012.09.003

Wiegman, M. (2016, 23 May). *Alledaags racisme vervuilt het debat*. Het Parool. https://www.parool.nl/nieuws/alledaags-racisme-vervuilt-het-debat~ba98ca17/

Wingfield, N. (2014, 15 October). *Feminist Critics of Video Games Facing Threats in "Gamergate" Campaign*. The New York TImes. https://www.nytimes.com/2014/10/16/technology/gamergate-women-video-game-threats-anita-sarkeesian.html

Wong, J. C. (2021, 16 January). *Banning Trump won't fix social media: 10 ideas to rebuild our broken internet – by experts*. The Guardian. http://www.theguardian.com/media/2021/jan/16/how-to-fix-social-media-trump-ban-free-speech

Yam, K. C., & Reynolds, S. J. (2016). The Effects of Victim Anonymity on Unethical Behavior. *Journal of Business Ethics*, *136*(1), 13–22. https://doi.org/s10551-014-2367-5

YouTube. (2019, 16 January). *Announcement: Strengthening enforcement of our Community Guidelines - YouTube Community*. Announcement: Strengthening enforcement of our Community Guidelines. https://support.google.com/youtube/thread/1063296/%F0%9F%9A%A9-announcement-strengthening-enforcement-of-our-community-guidelines?hl=en

Zheng, H., Sin, S.-C. J., Kim, H. K., & Theng, Y.-L. (2020). Cyberchondria: a systematic review. *Internet Research*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/INTR-03-2020-0148

Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power* (First edition). PublicAffairs.

Zuckerberg, M. (2020, 16 April). Facebook. Https://www.facebook.com/zuck/posts/10111806366438811.

Appendix 1: Advisory Committee

1. Prof. D.R. Veenstra (chairperson) – Professor of Sociology, University of Groningen
2. Dr T. Völlink – Assistant Professor of Psychology, Open University of the Netherlands
3. Drs. S. van der Waal – Research Director, Waag Technology & Society
4. Dr J.B. de Jong (initiating party) – Senior Strategy Advisor, Ministry of Justice and Security
5. Drs. T.L. van Mullekom (commissioning party) – Project Manager, Research and Documentation Centre (WODC)

Appendix 2: Explanatory workshop

On 21 January 2021, the Rathenau Instituut research team organised an exploratory workshop to map out what ministries, law enforcement and social welfare organisations already knew about harmful and immoral behaviour online, and what relevant initiatives were already underway at that time. A further aim was to identify research gaps to help guide the research. Fifteen participants attended the workshop.

Participant	Organisation
Mirjam Buisman	Ministry of Education, Culture and Science
Hidde Brugmans	Ministry of Economic Affairs and Climate Policy
Franca van der Laan	Dutch Police
Janet Lambeck	Ministry of Justice and Security
Joyce de Leij	Dutch Police
Ymke Lugten	Public Prosecution Service
Maarten Glorie	Ministry of Education, Culture and Science
Puck Gorrissen	Ministry of the Interior and Kingdom Relations
Pieter van Koetsveld	Ministry of Education, Culture and Science
[identity known to researchers]	National Coordinator for Security and Counterterrorism
[identity known to researchers]	National Coordinator for Security and Counterterrorism
Jolise Stol	Victim Support Netherlands
Paul Thewissen	Ministry of Education, Culture and Science
Bastiaan Winkel	Ministry of Justice and Security
Marcel Woltjes	Ministry of Education, Culture and Science

Appendix 3: Respondents

In February and March 2021, the research team interviewed 15 experts on harmful and immoral behaviour online. The interviews, combined with scholarly publications and grey literature, served as our sources for Chapter 3 (taxonomy), Chapter 4 (mechanisms), Chapter 5 (existing initiatives) and Chapter 6 (strategic agenda).

The respondents included researchers, business owners, social workers and experts by experience. Some of the respondents had broad expertise (e.g. on the mechanisms behind harmful and immoral behaviour online), while others were experts on a specific phenomenon or approach. We list them below.

Respondent	Role and organisation
Emine Uğur	Expert by experience on online hate speech
Jan Bats	Lecturer in Sociology and Philosophy of Technology, Open University of the Netherlands and The Hague University of Applied Sciences
Nick Beentjes	Managing Director Benelux, Channel Factory
Claudia van Diessen	Policy advisor, Halt
Eric van Furth	Director, GGZ Rivierduinen; Professor of Eating Disorders, Leiden University Medical Centre; Member of the K-EET steering committee (Chain Analysis for Eating Disorders)
Scarlet Hemkes	Press officer and communications advisor, 113 National Suicide Prevention; founder and former editor-in-chief of Proud2Bme.nl
Nina Hoek van Dijke	Owner, Jong & Je Wil Wat
Jeroen van den Hoven	Professor of Technology and Philosophy, Delft University of Technology
David Nieborg	Assistant Professor of Media Studies, University of Toronto
Richard Rogers	Professor of New Media and Digital Culture, University of Amsterdam
Emma Simons	Policy officer and researcher, Dutch Child and Human Trafficking Centre

Kees Teszelszky	Curator of Digital Collections, National Library of the Netherlands
Daniel Trottier	Associate Professor, Department of Media and Communication, Erasmus University Rotterdam
Patti Valkenburg	Professor of Media, Youth and Society, University of Amsterdam
[identity known to researchers]	National Coordinator for Security and Counterterrorism

Appendix 4: Interview guide

The study included semi-structured interviews based on the following interview guide. The questions are divided into twelve themes.

1. How does the respondent view the problem of immoral or harmful behaviour online, and what examples of such behaviour are known to them?
2. What is the scale of the problem (in the Netherlands), in the respondent's estimation? Is the respondent familiar with any statistics in that regard? Is the problem growing or diminishing, and what changes can be observed?
3. What differences does the respondent perceive between online and offline environments in terms of a specific phenomenon? What new developments can be expected, in the light of current technological advances and trends?
4. What are the harmful effects of the phenomenon, in the respondent's estimation?
5. Which online mechanisms inspire, facilitate or catalyse the phenomenon/behaviour?
6. Which groups are affected by the phenomenon/behaviour?
7. Which aspects of the phenomenon does the respondent feel are neglected at the moment?
8. Can the respondent describe any best practices in relation to the phenomenon?
9. What policy recommendations would the respondent like to make?
10. What sources or other experts can the respondent suggest for the researchers?
11. How might the Rathenau Instituut's research help the respondent?
12. Comments or suggestions by the respondent: What tips or pointers would the respondent like to pass on to the researchers?

Appendix 5: Workshop

On 13 April 2021, the Rathenau Instituut research team organised a workshop on options for tackling harmful and immoral behaviour online. The literature review and the interviews led to five solution categories that received considerable support but had not been developed into actual initiatives. The aim of the workshop was to flesh out these solution categories by encouraging a dialogue between staff members from different ministries, law enforcement and social welfare organisations, researchers, representatives of civil society organisations and others with relevant expertise. There were 22 participants in the workshop, divided into the five solution categories:

1. Online monitoring and assistance
2. Conversation about norms online
3. Value-sensitive platform design
4. Technical solutions
5. Enforcement of laws and rules online.

Participants	Organisation
1. Online monitoring and assistance	
[identity known to researchers]	OSINT Team, Central Unit, Dutch Police
Irene van Aarle	Proud2Bme
Willem Bantema	Thorbecke Academie, NHL Stenden University of Applied Sciences
Mirjam Buisman	Ministry of Education, Culture and Science
Jolise Stol	Victim Support Netherlands
2. Conversation about norms online	
Nick Felix	Public broadcasting company KRO-NCRV
Maarten Glorie	Ministry of Education, Culture and Science
Puck Gorrissen	Ministry of the Interior and Kingdom Relations
Fleur Jongepier	Radboud University
3. Value-sensitive platform design	

[identity known to researchers]	National Coordinator for Security and Counterterrorism
Blanca Harms	University of Groningen
Edo Haveman	Facebook
Pieter van Koetsveld	Ministry of Education, Culture and Science
Stefan Oude Wesselink	Opt Out Advertising
4. Technical solutions	
[identity known to researchers]	National Coordinator for Security and Counterterrorism
Michiel Leenaars	NLnet Foundation
Mieke van Heesewijk	SIDN Fund
Roelof Muis	Dutch Police
5. Enforcement of laws and rules online	
Nicole Lieve	Dutch Police
Jaqueline de Jong	Ministry of Justice and Security
Rolf van Wegberg	Delft University of Technology
Inge Welbergen	Ministry of Education, Culture and Science

Appendix 6: Validation meeting

On 26 May 2021, the Rathenau Instituut research team organised an expert meeting to validate the research results. The meeting was attended by the researchers and the staff of various executive agencies and civil society organisations.

Prior to the meeting, the participants were sent a summary of the study (approximately 20 pages). They were given the opportunity to comment on the research results during the meeting. The focus was mainly on the options for action arising from the analysis given in the report. The purpose of this exercise was to work on prioritising options for action and to reflect on the role that different parties can play in tackling harmful and immoral behaviour online.

Seven people attended the validation meeting.

Participant	Organisation
[identity known to researchers]	National Coordinator for Security and Counterterrorism
Linda Hell	Association of Dutch Advertisers
Heleen Janssen	University of Amsterdam
Franca van der Laan	Dutch Police
Willem van Lynden	Owner, Mediamaze; Board member of Stop Online Shaming
Jan-Willem van Prooijen	VU Amsterdam, Netherlands Institute for the Study of Crime and Law Enforcement
Arnout de Vries	TNO, Netherlands Organisation for Applied Scientific Research

Open Access
The Rathenau Instituut has an Open Access policy. Reports and background studies, scientific articles and software are published publicly and free of charge. Research data are made freely available, while respecting laws and ethical norms, copyrights, privacy and the rights of third parties.

Contact
Rathenau Instituut
Anna van Saksenlaan 51
P.O. Box 95366
2509 CJ The Hague
The Netherlands
+31 70 342 15 42
info@ Rathenau.nl
www.Rathenau.nl
Publisher: Rathenau Instituut

The Rathenau Instituut stimulates public and political opinion forming on social aspects of science and technology. We perform research and organise debate relating to science, innovation and new technologies.

Rathenau Instituut