ISD | Powering solutions to extremism and polarisation

ONLINE CIVIL COURAGE INITIATIVE von facebook

Hate Speech and Radicalisation Online

The OCCI Research Report

Johannes Baldauf, Julia Ebner and Jakob Guhl (Eds.)

Foreword by Prof. Dr. Peter Neumann

With contributions by

Simone Rafael
Alexander Ritzmann
Daniel Köhler
Prof. Dr. Christian Montag
Karolin Schwarz
Josef Holnburger
Dr. Matthias Quent
Sina Laubenstein
Alexander Urban

About this study

The research series Hate Speech and Radicalisation on the Internet provides interdisciplinary insights into the current developments of extremist activities on the internet. With the aid of expert contributions from all over Germany, the psychological, political, anthropological and technological aspects of online hate speech and radicalisation will be considered and recommendations will be made for political leaders, social media platforms as well as NGOs and activists.

About the publishing organisations

In 2016, together with the International Centre for Radicalisation and Political Violence (ICSR), the Institute for Strategic Dialogue (ISD) and the Amadeu Antonio Foundation, Facebook launched the Online Civil Courage Initiative (OCCI). The objective is to combat extremism and hate speech on the internet.

The ISD is an independent 'think and do tank', which co-operates with leading figures in politics, business, civil society and science, in order to find transnational answers to the challenges of our age regarding geostrategy, society and security. The objective of the ISD is to combat extremism worldwide and to bridge intercommunal rifts.

About the authors

Johannes Baldauf is the Head of OCCI Germany for Facebook.

Julia Ebner is a researcher at the Institute for Strategic Dialogue in London.

Jakob Guhl is a project associate at the ISD in London.

Prof. Dr. Peter Neumann is a senior fellow and founding director of ICSR at King's College, London.

Simone Rafael is the editor-in-chief of Belltower.News, the watchblog of the Amadeu Antonio Foundation.

Alexander Ritzmann is a member of the steering committee of the Radicalisation Awareness Network of the European Commission (DG HOME) and a scientific assistant and senior research fellow at the Brandenburg Institute for Society and Security.

Daniel Köhler is the Director of the German Institute of Radicalisation and De-radicalisation Studies.

Prof. Dr. Christian Montag is a professor at the Institute for Psychology and Education Theory (Molecular Psychology) at Ulm University.

Karolin Schwarz is a freelance journalist, fact-checker and founder of hoaxmap.org.

Josef Holnburger is a political scientist and blogger.

Dr. Matthias Quent is the Director of the Institute for Democracy and Civil Society in Jena.

Sina Laubenstein is Project Manager for the New German Media Makers and responsible for the No Hate Speech campaign there.

Alexander Urban is Chief Administrator of the Facebook group Ich Bin Hier [I Am Here] and the moderation team.

Designed by forster.co.uk. Typeset by Janina Neumann Design.

Contents

Greeting

By Marie-Teresa Weber

Public Policy Manager, Facebook Germany GmbH

In 2015 there was an increase in hate speech on Facebook. Politicians and society justifiably expected a robust response from Facebook. Since then, we have made massive investments and taken measures to ensure that we remove hate speech from the platform whenever we find it.

We have made great progress since that time. We have always emphasised that it is our responsibility to contribute our part towards solving this problem, but it is equally important that civil society should also respond to the challenge robustly, showing it decisively opposes radicalisation. It is essential to maintain the consensus of the whole of society and to counteract its increasing polarisation. In order to contribute to strengthening civil society, in 2016 Facebook launched the Online Civil Courage Initiative (OCCI), working with the Amadeu Antonio Foundation, the Institute for Strategic Dialogue (ISD) and the International Centre for the Study of Radicalisation and Political Violence (ICSR). We will continue to support the work of the OCCI.

Everyone is talking about terms such as hate speech, fake news and online echo chambers. But what is behind this? Research on the topics discussed in this report is still in its infancy. How do these phenomena affect society? What tactics are used by those who wish to attract followers for their radical ideas on the internet? What role does the internet play in this? And what role do social networks such as Facebook play? There is not yet any satisfactory answer to many of these questions. It is important to consider phenomena such as hate speech in an overall societal context.

The first step is therefore to analyse the problem in detail, as this OCCI research report does. Politics, civil society and companies have a common responsibility and must look for solutions together. At Facebook, we see our responsibility clearly and support this OCCI research report, which demonstrates the complexity of the challenges we face. Facebook will address the suggestions and criticisms this report raises. We are working hard to contribute our part to solve the problems which have been identified and especially over the past two years we have initiated and implemented a great deal. We are working continuously to improve ourselves. However, it is clear that we are in a learning process.

We hope that this report will help to further promote the dialogue between academia, civil society and politics. We look forward to the discussion it will create and wish to thank the authors for their work.

Foreword

By Prof. Dr. Peter Neumann

Neither hate speech nor radicalisation are entirely new phenomena. However, their manifestations and consequences have changed greatly over the past few decades, mainly because of the increasing role of the internet.

This is not surprising. Electronic communication has become an important part of our lives and at the same time has changed our lives so dramatically that it would be strange if the communication behaviour of extremists had not also changed. When 22-year-old foreign fighters in Syria take selfies and then publish them on social media, they are not only acting as extremists, but also – and above all – like 99% of their age group.

The fact that jihadist online magazines are found on the laptops of suspected terrorists and that they communicate with their comrades via messaging platforms and on social networks is not necessarily proof of online radicalisation, but rather shows that extremists, whatever their political views, are products of their age. Or do we seriously expect extremists still to write letters, book their flights via travel agents and take their photos to be developed?

The same applies to hate speech. As early as 15 years ago, the American social psychologist John Suler (2004) speculated on why there is so much hate on the internet, concluding it was because of a combination of (supposed) anonymity and the absence of moderating influences. He described the result as 'online disinhibition': not only do participants in virtual bubbles and/or countercultures reinforce their (often extremist) views, but they also spur each other on.

None of this is normal or tolerable in a pluralistic democracy. However, we must accept that the 'dark side' of the internet cannot simply be censored out.

Laws that oblige internet providers to remove content are therefore only – and at most – part of the solution. In view of the volume of information on the internet, they will never be 100% successful and require flanking measures. The internet is not only a 'problem', it can also be a solution. Those who demand the removal of content must also consider how internet platforms can be used to increase tolerance and mutual understanding. Many of the phenomena described in this report are an intrinsic part of the digital present. They can be combatted, but they will probably never completely disappear, unless we wish for an authoritarian society or to switch off the internet completely.

We will therefore be concerned with the topics of radicalisation and hate speech for many years to come. A part of this debate is the struggle to find the best political solution, allowing as much freedom of opinion as possible while complying with existing laws. This balance can only be achieved if we really understand what happens on the internet, where and when hate speech and radicalisation take place, and exactly how they function. This report provides an important contribution to this because we are still don't know the answers to many politically relevant research questions.

Probably the most significant question is the correlation between online hate speech and offline violence. Current studies appear to show that more violence occurs where a large amount of hatred is disseminated on social media. But what exactly is the relationship between hatred disseminated online and offline violence? Is online hate speech the cause of the violence, or merely the consequence of extremist communities which already exists? What are the mechanisms and circumstances, the tipping points, under which hate speakers perpetrate violence?

Foreword

It is important to understand the internet as an opportunity. How can we use online media to engage with people who are becoming radicalised? Under what conditions are there even possibilities for de-radicalisation? Which content can decrease the risk of calls for hate and violence and in which contexts do they work?

One of the greatest problems for researchers is that technology is developing more rapidly than research. Anyone who publishes their results five years after the start of a research project can expect that the technological environment in which hate speech and radicalisation take place has developed to such an extent that the results are no longer valid. For example, most of the messenger services through which people now communicate have existed for less than ten years.

The challenges are therefore enormous. However, the articles in this book show that, even in the digital age, research can make an essential contribution to achieving a balance between freedom of opinion and compliance with existing rules and laws, so that we can continue to live in a free and at the same time secure society.

Furthermore: I believe that research, politics and internet providers must co-operate more closely than previously, because empirically based research can play an important role in defusing political conflicts. This report makes an excellent contribution to this conviction.

Notes on definitions

Extremism

The ISD defines extremism as the advocacy of a system of belief that posits the superiority and dominance of one 'in-group' over all 'out-groups', propagating a dehumanising 'othering' mindset that is antithetical to the universal application of human rights. Extremist groups advocate, through explicit and subtler means, a systemic change in society that reflects their worlview.

Hate speech

The term 'hate speech' is the subject of public controversy. It is important to state that hate speech is not a legally specified category in Germany. The German Netzwerkdurchsetzungs-gesetz [Network Enforcement Act] primarily targets hate crime and legally punishable misinformation. Various definitions exist for hate speech.

The Council of Europe defines hate speech as:

> *"All forms of expression which disseminate, incite, promote or justify racism, xenophobia, antisemitism or other forms of intolerance based on hate, including intolerance which is expressed in the form of aggressive nationalism and ethnocentricity, discrimination and hostility to minorities, migrants and people with a migrant background."*

In its community standards Facebook defines hate speech as "a direct attack on a person due to protected characteristics: ethnic background, national origin, religious affiliation, sexual orientation, caste, gender, gender identity, handicap or illness". The authors of this article use various definitions of hate speech. In certain cases, the wide variety of definitions can result in different evaluations of particular circumstances and hence various statistics and studies cannot always be directly compared to each other. It is important to explain the definition of hate speech used in any study, as well as the method of obtaining the data.

Filter bubbles and echo chambers

Readers of the following articles will notice that to some extent genuine differences of opinion exist over certain concepts. For example, the question of whether we should use the terms 'filter bubbles' or 'echo chambers' and the current research situation with regard to these phenomena is answered differently by the authors.

These differences in interpretation demonstrate how controversial the concepts are, and the difficulty and importance of the debate about how they shape public discussion on these issues.

Reference

Suler, J. (2004) The Online Disinhibition Effect, CyberPsychology & Behavior 7(3).

Introduction

The value of the internet for present day society as an instrument to facilitate communication and access to information, and to enable social and political participation, is undisputable. But for many years governments, technology companies, civil organisations and researchers have struggled to find ways to counteract one of the darker aspects of the internet: the prevalence of hate speech and extremist content.

The complexity of this challenge has increased because of the increasing amount of disinformation on the internet, and because social media platforms and search engines which facilitate communication, entertainment and interaction on the internet have grown rapidly and developed further. Those who disseminate extremism, hate and disinformation usually react quickly to these changes and unfortunately are often more agile than those who wish either to understand their effects better or to limit them. As a result, approaches to countering hate speech and extremist content far too often lag behind these threats in speed and extent, regardless of whether they originate from government, industry or society in general.

These approaches can be divided into three major categories:

- Efforts to restrict availability of extremist and disinformation content and access to hate speech on the internet by reporting, filtering and removing content, and taking appropriate measures or invoking legal regulations;
- Efforts to compete with such content by providing a broader spectrum of perspectives through counter or alternative narratives, or more recently by fact checking; and
- Efforts to bolster the resilience of internet users through digital and civic education (typically of young people) and broad public awareness campaigns.

This delayed reaction is a result of the complexity of online challenges and the difficulties of balancing public security, protection of democracy and fundamental rights, including privacy and freedom of opinion, association, religion or beliefs, and the need to maintain global connectivity and the free and secure flow of information.

In June 2017, the German Bundestag passed the Netzwerkdurchsetzungsgesetz [NetzDG; Network Enforcement Act], in order to reduce these delays by legally obliging large social media platforms to remove obviously illegal content within 24 hours of receipt, or in cases of systematic breaches of this obligation, by imposing fines of up to €50 million. The NetzDG came into effect in October 2017 and after a transition period was fully introduced in January 2018.

Since then, those involved in preventing and combatting online hate speech, extremism and disinformation, and those working to protect democracy and fundamental rights and freedoms have closely observed the effects of this law in Germany. Against this background, this report aims to provide a series of up-to-date, interdisciplinary perspectives on the current debate about online hate speech and radicalisation in Germany.

Starting from the expert assessment by members of the steering committee of OCCI Germany and external authors, including a foreword by Peter Neumann, the report examines the tactics of Islamist and right-wing extremist groups online, the potential effects of disinformation, social media search algorithims on political polarisation and radicalisation, and the role of non-legislative civil responses to these challenges. This report describes current gaps in our understanding of these questions and makes a series of suggestions for political decisionmakers, the private sector and civil society.

Introduction

Chapters

In Chapter 1 **Simone Rafael** and **Alexander Ritzmann** briefly describe the debates in Germany about hate speech and online extremism, and the response to the introduction of NetzDG across the entire political spectrum. The authors discuss whether the legislation has resulted in more effective removal of illegal content without impairing legitimate freedom of expression or causing other unintended consequences, and ask how the legislation could be supplemented with broader co-operation and consultation with civil society and other stakeholders.

In Chapter 2 **Daniel Köhler** and **Julia Ebner** examine how Islamist and extreme right-wing groups use the possibilities of the internet to polarise society and radicalise and recruit vulnerable persons. The authors emphasise the similarities in the tactics and strategies of these groups in the online ecosystem and stress the need for fundamental understanding of their methods in order to develop effective and appropriate responses.

In the third chapter **Prof. Dr. Christian Montag** examines the effects of algorithmic filter bubbles on individual psychology and further political polarisation, and emphasises the need for additional interdisciplinary research in order to better understand this growing phenomenon. Although several studies question the effect of filter bubbles at an individual and societal level, the author recommends that the private sector, political decisionmakers and researchers take this phenomenon seriously.

In Chapter 4 **Karolin Schwarz** and **Josef Holnburger** consider the role of disinformation for disseminating hate speech and extremist perspectives on the internet and disrupting democratic elections. The authors assess the range and effects of prominent online disinformation campaigns in Germany and examine various attempts by social media platforms and search engines to counteract these effects.

In Chapter 5 **Dr. Matthias Quent** argues that the fight against hate speech and extremism on the internet cannot just be the responsibility of the government and the private sector, and that the prevalence of online hate speech and extremism on the internet can be understood as part of a broader, offline cultural backlash against the progressive achievements of modern democratic societies. In view of the prevalence of unpleasant or uncomfortable online content that does not contradict legal standards nor conditions of use and community guidelines of major platforms, Dr. Quent suggests that a robust response from civil society is required, including investment in education and promotion of the narratives of marginalised groups.

Chapter 6 is by **Sina Laubenstein** and **Alexander Urban**. They propose there should be a balanced response to online hate speech and extremism, which promotes counter responses and positive alternative narratives. The authors examine data to assess the success of initiatives for counter responses in two case studies on the German chapter of the No Hate Speech Movement of the Council of Europe and the Facebook group #ichbinhier. Finally, they point out the importance of having a clearly defined strategy, target group and core message, and the need for a sustainable, long-term, social media presence.

In the final chapter, **Jakob Guhl** and **Johannes Baldauf** summarise the recommendations made in this report and call for a continuous dialogue between politicians, technology companies and civil society, and the specification of a common framework for combatting hate speech and extremism on the internet.

Among other things, the authors recommend there should be greater support for those affected by hate speech on the internet, and education programmes that make young people aware of various online dangers and help them to become digital citizens. This final chapter is concerned with the central tensions and unanswered questions that need to be addressed by governments, technology companies and civil society when dealing with hate speech and extremism on the internet.

1. Background: the ABC of hate speech, extremism and the NetzDG

By Simone Rafael and Alexander Ritzmann

Abstract
In this introductory article Simone Rafael and Alexander Ritzman explain important background information about the current debates on hate speech, online extremism and the NetzDG.
Hate speech on social media had not been monitored adequately for a long time, until in 2015 the German Ministry of Justice established the Task Force for Dealing with Hate Speech in order to respond to the dissemination of misanthropic and extremist content on the internet. However, when NetzDG was adopted in 2017, politicians were criticised not only by extremists butalso by representatives of democratic civil society, who were similarly sceptical. The implementation and search for improvements or supplements to the NetzDG is in full progress.

1.1 The problem of hate on the internet

Right-wing extremism, racism, antisemitism, Islamophobia and other forms of group-related prejudice have existed in the digital environment as long as the internet has (Zick 2009). Initially no one wanted to see any abuse of the new, free medium for agitation against civil liberties – except those who were subjected to comments that advocated Nazism, were racist, denied the Holocaust, were Islamophobic or propagated conspiracy theories: on the one hand victims of misanthropic postings and on the other hand social media managers and moderators who had to come to terms with themselves, their colleagues and bosses. What can stay on our pages and what needs to be concealed or removed? Do our netiquette codes, discussion rules or social media terms of service fit in with this, or do they need to be changed? Are content removals from social media platforms made proactively or reactively?

Because a free internet without censure is one of the positive benefits of the internet, all removals were usually made without comment and the precise criteria for removal were kept secret. Nonetheless, even in the early years of social media it was already clear to social media moderators that anyone who permits misanthropic contributions to be made available online changes the course of democratic discourse. Social media moderators exclude people who are attacked verbally on the internet because they withdraw from discussions and are therefore no longer represented in the brave new internet world, which is allegedly accessible to all. Democratic values and standards should also apply in the digital environment.

However, in practice the implementation did not function immediately. Although networks had general terms of business and/or community standards, which prohibited discrimination, racism and antisemitism, there were great differences in how they were applied. There were no criminal prosecutionsin what the German Chancellor Angela Merkel called #Neuland [new or unknown land] in 2013 (Walezcek 2013).

The Amadeu Antonio Foundation observed that until the end of 2014, prosecution of most charges of hate speech on the internet was discontinued. These days, the handling of hate crime on the internet differs between the various German federal states, as does the knowledge and staffing of law enforcement authorities in the field of hate crime on the internet. Throughout Germany, the judiciary makes extremely heterogeneous verdicts, even in cases with similar wording or circumstances. It still remains difficult to find a definitive red line. As if this was not complicated enough, there is also the problem that the criminal laws against racism, antisemitism or Islamophobia are often not the right instruments on their own in the context of wider societal problems.

Freedom of expression is valued highly in Germany, so despite the chatter in right-wing internet spheres, a great deal of racist or misanthropic statements can be stated, without the author being subject to prosecution. Nonetheless, these postings, termed 'dangerous speech' in research,[1] require a response in order to prevent them from developing their toxic effect. Ideally these responses should be given not only by the state, but also by other participants: civil society – users, media, organisations and social media companies. But civil society as a whole is not especially well prepared for this.

When Facebook first started in Germany in 2008, there were dedicated individuals who made statements opposing hate, not yet called 'counterspeech' but fulfilling the same function. Over time, the methods changed. Although in 2009 there were online demonstrations in which users participated with their photo or profile image, these were later dropped, due to security concerns for the participants, who were personally attacked by aggressive neo-Nazis. Online registries for right-wing extremist profiles? On further consideration, these do not necessarily correspond to democratic discussion and behaviour. Monitoring of right-wing extremist and populist activities via Facebook pages? This proved to be a good method to raise awareness. There are many opportunities for development in this process.

By 2015 at the latest, the question of counter strategies became increasingly urgent in German-speaking countries. With the rise of Patriotic Europeans against the Islamisation of the Western World (Pegida) and the founding of Alternative for Germany (AfD), the self-confidence of the right-wing extremists grew, so they came out into the open.

While the corresponding players had previously acted within closed groups or on little-known pages without public knowledge, they now crowded onto non-right-wing Facebook pages and in groups, with their racist and anti-refugee content. They therefore became prominent and a serious problem: on the one hand for democratic discourse, which increasingly shifted to the far right and promoted the normalisation of misanthropic positions.

On the other hand, this became a problem for social media companies such as Facebook, which now had to consider whether shouting down democratic users also threatened their customers' loyalty.

1.2 And Facebook?

Social media companies must develop their handling of hate speech in a public atmosphere, in which customers are at times highly critical and expect immediate perfection. Initially, Facebook did not respond to inquiries or reproaches – which resulted in accusations that the company completely lacked transparency. In the German-speaking region, for many years this resulted in stubborn accusations, for example "Facebook does not delete hate content, because it does not matter to Facebook", "Facebook uses it, because it is traffic" or even "Facebook supports it, because they think Nazi content is OK". This also happened with reference to the idea of freedom of speech in the USA, which resulted in misanthropic and extreme right-wing language not being sanctioned. Here it was completely irrelevant whether Facebook actually did or did not remove the corresponding content. Until about 2015 these accusations tended to come from the democratic side. Extreme right-wing users felt comfortable and relaxed on Facebook and easily accepted occasional removals and simply created new profiles.

Because there was no transparency and Facebook did not even disclose information removal, barring criteria or the size and qualification of the processing teams, there was at most an empirical learning process, both for right-wing extremist groups as well as for groups who attempted to remove right-wing extremist content by referring to the Facebook general terms of business.

Since 2010, various civil society organisations have sought contact with Facebook and other social media in order to work out practical solutions in this field. This essentially includes encouragement to make a stand against right-wing extremism, racism and group-related misanthropic statements, which in the case of Facebook was initially made by small campaigns and later by the foundation of the Online Civil Courage Initiative (OCCI) in co-operation with ISD.

[1] *See https://dangerousspeech.org/*

OCCI supports good counterspeech on social media to counteract hate speech in the area which is not covered by criminal law. Their achievement is that new, competent organisations from the offline world were able to be trained and motivated for online action. People who are already committed to democracy are encouraged to keep participating – even if they are frustrated about articles which have not been removed. This is associated with awareness-raising that not all of the content which is considered to be inappropriate can or should be simply removed: our society cannot avoid dealing with topics of hate-filled content.

Nonetheless, companies can and should accept their social responsibility and set a signal. Facebook did this in parallel with the increase in anti-refugee agitation on the platform. Because of this, from 2015 onwards dissatisfaction grew in the extreme right-wing sphere, because, increasingly, low-threshold articles, pages and profiles were barred or removed according to community standards. This resulted in an increasingly emotionalised and polarising discussion within this scene: freedom of speech was being curtailed, was in danger, or no longer existed. The accusation from extremist groups was often: "Because I can no longer publish racist contents or call for violence on an internet platform, this is censorship."

Of course, companies can set and enforce the rules according to which their platform can be used. As this was not the case for a long time, the extreme right-wing sphere felt that they had been robbed of their playing field.

1.3 And then came hate speech …

From 2015, the German public became aware that there was hate speech on the internet. In parallel with corresponding discussions in the USA since 2009, 'hate speech' was used as a general term for hate-filled expressions intended to put down and denigrate individuals and groups of people, covering right-wing extremism, racism, antisemitism or other forms of group-related enmity on the internet.

This term had the advantage that it appeared to be more universal and succinct than those previously used, and the disadvantage that it diluted what the content actually referred to: an ideology of unequal values directed against universal human rights and fundamental democratic principles of equality of all people.[2]

In October 2015, the (then) German Federal Minister of Justice Heiko Maas initiated the Task Force for Dealing with Hate Speech. The major social media providers were involved in this – Facebook, Google, YouTube and Twitter – as well as employees of the Federal Ministry of Justice, non-governmental organisations (NGOs) and employees of the Federal Criminal Police Office (Rafael 2015). Not surprisingly, in view of the inviting institution, this task force was exclusively concerned with criminally relevant content and the question of how it was to be removed. The mere establishment of such a working group enraged the right-wing extremist internet sphere, and was accompanied by many accusations of censorship and personal attacks on participants of the task force.

1.4 … and extremist propaganda

Extremists try to be present wherever people obtain information and communicate. Because of this, over the past few years the new media ecosystems which resulted from social media played an increasingly important role in spreading extremist propaganda.

Extremists increasingly concentrated on optimally interlinking their online communication methods with their offline activities in order to maximise their circulation and effectiveness. Extremism in its violent form and terrorism in particular can be described as a type of bloody political theatre, in which violence is a means to a (political) end (Ritzmann 2016).

[2] *The Council of Europe defines hate speech as "All forms of expression which disseminate, incite, promote or justify racism, xenophobia, antisemitism or other forms of intolerance based on hate, including intolerance which is expressed in the form of aggressive nationalism and ethnocentricity, discrimination and hostility to minorities, migrants and people with a migrant background" (Council of Europe 1997).*

Escalation, polarisation and fuelling fear of neighbours are the preferred strategic methods in the toolbox of extremists who wish to overcome the free democratic order (Vidino, Marone and Entenmann 2017). Fear and hate are intended to divide society and provoke governments to over-react. Therefore, at least at a strategic level, extremism only works because it has an audience, which itself responds in an extreme manner. The response to extremist violence by the media, state and society therefore has an escalating or de-escalating effect. This either plays (unconsciously) into the hand of terrorists, or frustrates their strategy.

From an operative point of view, propaganda in which hate speech in the form of denigration of others (out-groups) often forms an essential part has two functions for extremists. First, their supporters (in-groups) are to be activated and motivated. Extremist propaganda can therefore be described as a group-related call to arms (Ritzmann 2018). In addition, propaganda is used for recruiting and expanding the in-group.

Essential components of hate speech propaganda are:

a) Victimhood narratives ("we are under attack!"),

b) Redemption scenarios ("only a caliphate, dictatorial state, workers' state can protect us!") und

c) Fulfilling individuals' desire to gain significance ("Be a hero, a mother of the nation, a builder!").

For this, extremism requires an ideology based on great narratives in order to create a binding ideology. This great narrative makes the difference between the so-called Islamic State and Mexican drug cartels. Both impose their aims and interests to some extent with extreme violence.

The same applies to the difference between right-wing extremist groups and (apolitical) criminal biker gangs. The legitimation of their acts, especially violence, is based on a particular ideology. This chosen 'truth' serves as an interpretative framework for categorising events and topics. Without ideology there is no extremism or terrorism, but 'only' criminality.

1.5 The effect and limits of propaganda

The internet, and in particular social media, are to some extent described as a decisive factor or at least as catalysts for the radicalisation process. However, the extent of the direct influence of propaganda on people can only be determined in individual cases (Horgan 2014). It is also disputed whether propaganda can initiate extremist views and behaviour, or whether it only reinforces existing sympathies or convictions (Policy Department for Citizens' Rights and Constitutional Affairs 2017). In particular the role of online filter bubbles – media content which is pre-sorted by algorithms and the resulting echo chambers as forums for like-minded people – gives cause for concern (Lauer 2017; Stöcker 2016).

In the current discussion of online filter bubbles it is often forgotten that these are not simply technological constructions, but rather reflect human needs and biological algorithms (Ritzmann 2017, 2018). Starting with thinking short-cuts (heuristics) and motivated reasoning, the human brain functions on the basis of a number of biological programmes, which up to now are only partially known (Epley and Gilovich 2016).

Confirmation biases and cognitive frames permanently pre-select information and prefer that which best matches our existing convictions (Wehling 2017). In contrast, information that questions our convictions and 'truths' (cognitive dissonance) is relativised and devalued. Our biological filters are therefore components of largely unknown processes. A great deal happens offline and entirely without the support of algorithms from Facebook or Google.

Psychology and neurosciences show that adults have pronounced defence mechanisms against external manipulation (Kaplan, Gimbel and Harris 2016). Without these biological filters, depending on the quality of the narratives, propaganda and personal relationships, people would continuously change their political or religious identity.

Adults therefore only have limited control over their initial core beliefs and identity. Making a drastic change to this requires a great deal of effort, a conscious manipulation of ourselves. The motivation for such self-manipulation can be caused by a personal crisis, possibly extreme emotional stress. Then as a result we seriously question what we have previously believed to be good and right. The associated urge for new orientation can create the necessary cognitive opening and make us receptive to propaganda in the form of 'promises of salvation'.

Children, adolescents and young adults whose personalities are developing as they look for truths and their role in society, whose filter bubble is not yet (fully) developed, are accordingly potentially more susceptible than older people to manipulation.

The probability of accepting extremist narratives also increases if a person already believes in global conspiracies (Lambarty 2017). These have additional emotional value when compared with beliefs in an ideology, in particular an extremist one. Conspiracy theorists consider themselves to be the few 'people who can see', elite in-groups who knows the truth. This is associated with an increase in self-value, which automatically results in devaluation of the 'blind and stupid' out-group.

1.6 ... and finally the NetzDG

The German Federal Minister of Justice wanted to make a stronger statement on the handling of hate speech and in spring 2017 submitted the draft of the so-called Network Enforcement Act (the NetzDG), which was passed by the Bundestag shortly before the summer recess. This law is primarily aimed at companies as actors in the fight against hate speech. Social media platforms with more than 2 million active users in Germany (such as Facebook, Twitter and YouTube) must remove "obviously illegal" content within 24 hours after it has been reported to them, and less obvious criminal content within seven days. They must present a detailed report of resources, removal teams and procedures every six months. Fines are imposed if there are systematic infringements of the legislation. Furthermore, social media platforms must appoint a so-called domestic authorised recipient. This has the positive effect for civil society that civil lawsuits against international companies are now possible in Germany and that transparency with regard to the removal practice will probably increase as a result of the reporting – the first reports were published in June 2018.

However, this also causes problems: the main point of criticism is the privatisation of jurisdiction in the sensitive area of freedom of opinion – that companies such as Facebook should decide on cases which are even disputed among lawyers. If the company does not want to take any risks, this could result in over-blocking – excessive removal of content which is not illegal (Amadeu Antonio Foundation 2017).

Accordingly, the introduction of the NetzDG has created the situation that the new law in its present form is not only maligned as a censorship law by right-wing activists, but has also been the subject of criticism by many democratic actors, for example in the Declaration for Freedom of Opinion.[3]

1.7 Current implementation of the NetzDG

The NetzDG came into effect on 1 January 2018. Hate content on social media according to the community standards and/or the NetzDG can now be reported, and users can report content directly to the Federal Ministry of Justice if they believe networks have not complied with their obligation for removal.[4] A report by the German news programme Tagesschau stated that after the first two months, 205 complaints relating to the NetzDG had been submitted to the Federal Office of Justice – but the office had expected 25,000 complaints (Tagesschau 2018).

[3] *See https://deklaration-fuer-meinungsfreiheit.de/*
[4] *See https://www.bundesjustizamt.de/DE/Themen/Buergerdienste/NetzDG/NetzDG_node.html*

Users have not observed any significant change from the introduction of the NetzDG: on 2 January 2018 the AfD members of parliament Beatrix von Storch and Alice Weidel complained that they had been maltreated by the NetzDG – however, they had posted racist comments, which could have been deleted according to the community standards anyway (Frankfurter Rundschau 2018; Rafael 2018; Zeit Online 2018).

At the beginning of 2018, the number of spurious complaints about users dedicated to democracy and human rights increased, because some right-wing activists attempted to use the agitation about the NetzDG for their own ends.[5]

1.8 Are there alternatives and useful supplements to the NetzDG?

With global forums such as the Global Internet Forum to Counter Terrorism and the initiative founded by the United Nations Counter-Terrorism Executive Directorate, Tech Against Terrorism, the internet industry is attempting to consolidate the exchange of knowledge and technology. Stakeholders from politics, administration and civil society are involved and research projects are being financed.[6] Tech against Terrorism has focuses on providing specialist knowledge and technologies to smaller companies.

At the European level, in contrast to the German Federal Government, a co-operative and consultative procedure has been preferred. At regular intervals, the EU Internet Forum established by the European Commission in December 2015 holds meetings between members of European authorities such as Europol, EU interior ministers, representatives of the internet industry (including Google, YouTube, Facebook, Microsoft, Twitter, JustPaste.it, Snap, WordPress and Yello) and the EU Commission, and other relevant stakeholders such as the Radical Awareness Network and the research network VOX-Pol. So far representatives from civil society have not been included.

The aim of this voluntary partnership is to "restrict access to terrorist content on the internet and to support partners from civil society in enlarging the scope of powerful alternative discussions on the internet" (European Commission 2017). The partnership established and financed the Civil Society Empowerment Programme, in which more than 250 European civil society organisations have participated in training sessions for the implementation of alternative or counterspeech campaigns.[7]

The EU Internet Form follows the approach of politically moderated self-regulation by industry sectors, as a substitute for legislation if the protagonists involved achieve the agreed objectives. A legislative solution is only attempted if this is not successful.

At the end of 2016 the internet industry committed to increase investment in technological solutions, especially for the identification and automatic removal of extremist content. For this purpose, a database of hashes was created in December 2017, on the basis of which a re-upload filter reports or prevents repeated uploading of content which has already been reported as illegal or infringing the conditions of use (ibid.) Interestingly, this re-upload filter technology has been used for years by Microsoft, Google and Facebook to prevent the repeated uploading of child pornography content. In 2015, in co-operation with the Counter Extremism Project, Professor Hani Farid, who developed the algorithm for automatic deletion of child pornography, presented an analogously functioning re-upload filter known as eGLYPH, which can report or remove content that has been previously classified as extremist (Ketterer 2016). There is still a lack of transparency over the selection and removal criteria of the database of hashes of the internet industry.

[5] *Facebook published its community standards in April 2018., where it defined hate speech as "direct attacks on persons due to protected characteristics: ethnic background, national origin, religious affiliation, sexual orientation, caste, gender, gender identity, handicap or illness. Immigrant status is to a certain extent a protected characteristic" (Facebook 2018).*

[6] *See https://www.techagainstterrorism.org/project-background/*

[7] *See https://ec.europa.eu/home-affairs/what-we-do/networks/radicalisation_awareness_network/civil-society-empowerment-programme/csep_db_en*

The German government states:

> *"In the context of voluntary measures by internet service providers, in the opinion of the Federal German Government, the removal of illegal internet content should also be performed in a transparent manner" (Deutscher Bundestag 2018).*

The necessary social and legal discussion about the balancing of security requirements, civil liberties and business interests cannot be replaced by exclusive reliance on technological solutions or legislation (Llansó 2018).

Rather, the complex nature of the problem must be addressed holistically, which requires closer co-operation between state, private and civil society stakeholders. A German internet forum with equal involvement of internet companies, government and parliamentary representatives, as well as representatives of civil society, could make an important contribution to this. The following chapters examine the various problem areas and challenges, and possible approaches to find solutions, in greater detail.

References

1. *Amadeu Antonio Foundation (2017) Network Enforcement Act Endangers the Culture of Debate on the Internet – Amadeu Antonio Foundation Rejects the Draft Version and Demands a Round Table Discussion, Belltower News, http://www.belltower.news/artikel/netzwerkdurchsetzungsgesetz-gef%C3%A4hrdet-debattenkultur-im-netz-amadeu-antonio-stiftung-lehnt [18.07.18].*

2. *Council of Europe (1997) Recommendation No. R (97) 20 of the Committee of Ministers to Member States on 'hate speech', https://rm.coe. int/1680505d5b [18.07.2018].*

3. *Deutscher Bundestag (2018) Response of the Federal Government to the Parliamentary Question by Members Andrej Hunko, Ulla Jelpke, Zaklin Nastic, Further Members and the Parliamentary Group die linke – Document 19/565 – Threat of Statutory Measures by the European Commission for the Removal of Internet Content, https://dip21.bundestag.de/dip21/btd/19/007/1900765.pdf [18.07.18].*

4. *Epley, N. and T. Gilovich (2016) The Mechanics of Motivated Reasoning, Journal of Economic Perspectives, 30(3), 133–40, http://faculty.chicagobooth.edu/nicholas.epley/EpleyGilovichJEP2016.pdf [18.07.18].*

5. *Europäische Kommission (2017): Bekämpfung des Terrorismus im Internet: Internetforum drängt auf automatische Entdeckung terroristischer Propaganda. http://europa.eu/rapid/press-release_IP-17-5105_de.htm [18.07.18].*

7. *Facebook (2018) Community Standards, https://www.facebook.com/communitystandards/ [18.07.18].*

8. *Frankfurter Rundschau (2018) Incitement to Hatred is not an Expression of Freedom of Opinion, http://www.fr.de/politik/heiko-maas-vs-afd-volksverhetzung-ist-kein-ausdruck-von-meinungsfreiheit-a-1419454 [18.07.18].*

9. *Horgan, J. (2014) The Psychology of Terrorism, Routledge.*

10. *Kaplan, J. T., S. I. Gimbel and S. Harris (2016) Neural Correlates of Maintaining One's Political Beliefs in the Face of Counterevidence, Scientific Report 6, 39589, https://www.nature.com/articles/srep39589 [18.07.18].*

11. *Ketterer, J. (2016) An Algorithm to Stop Extremism in Social Networks, Wired, https://www.wired.de/collection/life/ein-algorithmus-soll-extremismus-sozialen-netzwerken-stoppen [18.07.18].*

12. *Lambarty, P. (2017) Don't Trust Anyone: Conspiracy Ideas as Accelerators of Radicalisation?, EXIT Deutschland, 5, http://journals.sfu.ca/jed/index.php/jex/article/viewFile/72/198 [18.07.18].*

13. *Llansó, Emma (2018): Who needs courts? A deeper look at the European Commission's plans to speed up content takedowns. VOX-Pol. http://www.voxpol.eu/who-needs-courts-a-deeper-look-at-the-european-commissions-plans-to-speed-up-content-takedowns/ [18.07.18].*

14. *Lauer, S. (2017) The Myth of Filter Bubbles and Echo Chambers, Belltower News, http://www.belltower.news/artikel/der-mythos-von-filterblasen-und- echokammern-12829 [18.07.18].*

15. *Policy Department for Citizens' Rights and Constitutional Affairs (2017) Countering Terrorist Narratives, European Parliament, http://www.europarl.europa.eu/RegData/etudes/STUD/2017/596829/IPOL_STU(2017)596829_EN.pdf [18.07.18].*

16. *Rafael, S. (2015) Task Force Against Hate Content on the Internet: A Great Deal Still Needs to be Done, Belltower News, http://www.belltower.news/artikel/task-force-gegen-hassinhalte-im-internet-es-gibt-noch-viel-zu-tun-10780 [18.07.18].*

17. *Rafael, S. (2018) AfD Twitters the End of Freedom of Opinion, Belltower News, http://www.belltower.news/artikel/afd-twittert-das-ende-der-meinungsfreiheit-13113 [18.07.18].*

18. *Ritzmann, A. (2016) From Suicide to Mass Murder, Terroristische Propaganda und die Verantwortung der Medien [Terrorist Propaganda and the Response of the Media], TV Diskurs [TV discussion], https://tvdiskurs.de/beitrag/vom-selbst-zum-massenmord/ [18.07.18].*

19. *Ritzmann, A. (2017) Propaganda – Wirkung, Grenzen und Gegenmaßnahmen [Propaganda – Effect, Limits and Countermeasures], Interventionen [Interventions], Fachzeitschrift für Verantwortungspädagogik, http://www.violence-prevention-network.de/de/ publikationen/interventionen-zeitschrift-fuer-verantwortungspaedagogik [18.07.18].*

20. *Ritzmann, A. (2018) A Tribal Call to Arms: Propaganda and What PVE Can Learn from Anthropology, Psychology and Neuroscience, European Eye on Radicalization, https://eeradicalization.com/a-tribal-call-to-arms-propaganda-and-what-pve-can-learn-from-anthropology-psychology-and-neuroscience/ [18.07.18].*

21. *Stöcker, C. (2016) Einfluss auf die Gesellschaft: Radikal dank Facebook [Influence on Society: Radical thanks to Facebook], Der Spiegel, http://www.spiegel. de/netzwelt/netzpolitik/filterblase-radikalisierung-auf-facebook-a-1073450.html [18.07.18].*

22. *Tagesschau (2018) Fewer Complaints than Expected, https://www.tagesschau.de/inland/hassimnetz-101.html [18.07.18].*

23. *Vidino, L., F. Marone and E. Entenmann (2017) Fear Thy Neighbor: Radicalization and Jihadist Attacks in the West, George Washington University, IPSI, ICCT.*

24. *Walezcek, T. (2013) Merkel's 'Neuland' is Becoming an Internet Joke, Der Tagesspiegel, https://www.tagesspiegel.de/politik/die-kanzlerin- und-das-internet-merkels-neuland-wird-zur-lachnummer-im-netz/8375974.html [18.07.18].*

25. *Wehling, E. (2017) Politisches Framing, Wie eine Nation sich ihr Denken einredet – und daraus Politik macht [Political Framing, How a Nation Persuades its Thought – and Makes Politics from it], Halem-Verlag.*

26. *Zeit Online (2018) Investigations against Storch due to Incitement to Hatred, https://www.zeit.de/politik/ deutschland/2018-01/alternative-fuer-deutschland-afd-beatrix-von-storch-volksverhetzung-koelner-polizei- twitter [18.07.18].*

27. *Zick, A. (2009) Group-related Enmity – the Scientific Background, Belltower News, http://www.belltower.news/artikel/ gmf-gruppenbezogene-menschenfeindlichkeit-wissenschaftlicher-hintergrund [18.07.18].*

2. Strategies and tactics: communication strategies of jihadists and right-wing extremists[8]

By Daniel Köhler and Julia Ebner

Abstract
Both right-wing extremists and Islamist groups use social media to spread their political and ideological messages. To prevent extremist groups from abusing new media ecosystems and their mechanisms for their purposes, preventative measures must be based on sound knowledge of their methods. In this article Daniel Köhler and Julia Ebner explain which media strategies right-wing extremist and Islamist groups pursue in order to target internet users who are susceptible to radicalisation, recruit them for their purposes, intimidate political opponents and manipulate online discussion. Despite the different ideological convictions of the groups, their methods overlap.

2.1 The media battle

In 2005 the (then) deputy of Osama bin Laden and Number Two of Al Qaida (AQ), Aiman az-Zawahiri, wrote a letter to Musab al-Zarqawi, the leader of Al Qaida in Iraq (from which the so-called Islamic State later originated), in which he complained about the excessive brutality of the Iraqi Al-Qaida offshoot under Zarqawi. Fearful that this excessive violence would risk Al Qaida losing support of the population, Zawahiri demanded moderation:

> *"More than half of this battle takes place on the battlefield of the media. We are in a media battle for the hearts and minds of our umma."*

Zawahiri's fears were to be proved right, as the so-called Anbar Awakening of Sunni groups against the tyranny of Al Qaida in Iraq showed in 2006.

In spite of this, the originally critical comment became a proverb among jihadists, and even entered into youth pop culture on t-shirts and other products: "Half of jihad is the media." A correspondingly large number of comments and information about the use of the media as a part of militant jihad is now distributed among groups and networks within the community, and jihadist groups have a long tradition of using media strategically for propaganda and recruiting purposes. For example, in the 1980s the Afghan Mujahedin recruited international fighters by distributing printed newspapers, audio and video cassettes. At the time of the war in Bosnia, specially prepared videos of combat scenes from various battles were distributed. With the internet, it has been possible to produce products suitable for mass propaganda much more easily.

Research on jihadist propaganda in general or in particular (for example the use of music, video games, magazines) is now extensively available (Ciovacco 2009; Farwell 2014; Gråtrud 2016; Ingram 2016a; Lakomy 2017; Milton 2016; Skillicorn and Reid 2014; Torres-Soriano 2016; Whiteside 2016), but comparisons with right-wing extremist and jihadist media strategies have been rare until now. Communication strategies and trends in the use of digital media by Islamist and right-wing extremist movements show surprising similarities.

Right-wing extremists give the information-war the highest priority and attempt to misuse both traditional and new media ecosystems to achieve their aims. As early as 1998, the ex Klu Klux Klan leader David Duke wrote that the internet would be helpful for the "global revolution for the awareness of the white race" and would assist the movement in directly addressing its target groups independently from the mainstream media (Kessler 1999).

[8] *The so-called 'Overton window' is a theory by Joseph P. Overton: the idea that there is a framework for what are perceived as socially acceptable opinions and morally justifiable positions in public debate.*

Strategic polarisation

Strategy manuals of Islamists and right-wing extremists have in common that they state that they wish to achieve political change by means of strategic polarisation. Instrumentalisation of the media and exploitation of the weak points of new media are a central component of a broad-based strategy that aims to eliminate grey areas and gradually shift the 'Overton window'[8] – what can be said and is socially acceptable (Atwan 2015; Davey and Ebner 2017). Outwitting algorithms, hijacking 'trending hashtags' and optimal use of echo chamber effects are just some of the examples of the methods used to accelerate polarisation online and offline (Davey and Ebner 2017).

The main target group of both extremist fringe movements is young people. Because of this, Islamic State (IS) propaganda often imitates scenes from Hollywood films or computer games to spread its black-and-white presentations of good and evil to its target group, Generation Z, and to present it in a playful and attractive way for young people (Atwan 2015; Lesaca 2015). The call to 'Respawn in Jannah' ('Be reborn in paradise'), aping computer games such as Call of Duty, shows the attempt to combine jihadist vocabulary with the language of computer games.

Right-wing extremist mobilisation attempts, above all the online campaigns of the international alt-right, also frequently refer to internet and pop culture. For example, the right-wing terrorist Luca Traini, who fired on African migrants in February 2018, was presented by Italian neo-fascists as the God of Race War in reference to the computer game God of War (Ebner and Davey 2018).

A year before the letter from Zawahiri cited above, in 2004, one of the most important manuals of jihad was published online. Called *The Management of Savagery: The Most Critical Stage Through Which the Islamic Nation Will Pass*, this work is attributed to the (then) leading strategist of Al Qaida, Abu Bakr an-Naji (an-Naji 2004). It is one of the pioneering concepts for militant jihad and has had a defining influence on the later so-called IS.

The Management of Savagery states that the main aim of all activities is to end the predominance of Western ('unbelieving') nations over Islamic territories. For this it is essential to break the nimbus ('halo' in the original) of Western invulnerability by showing images of attacks and victims online in countless and continuous small operations, in order to generate a permanent feeling of insecurity, fear and chaos. This chaos (and the associated 'savagery') is not only aimed at the West, but also at the civilian populations of mainly Muslim countries. An-Naji expected a simple psychological process: if people suffer from insecurity and savagery, they will follow the party that offers security and order and can enforce it:

> *"When savagery happens in several regions – whether we administer them or they are neigh boring regions or further away – a spontaneous kind of polarization begins to happen among the people who live in the region of chaos. The people, seeking security, rally around the great personages of the country or a party organization or a jihadi organization or a military organization composed of the remainders of the army or the police of the regimes of apostasy" (an-Naji 2004, 110).*

Precisely the same logic can be found in the right-wing terrorist concept 'strategy of tension', which exploits the desire for security, law and order in society (Bale 1994; Jenkins 1990). Right-wing terrorist attacks should be committed under a 'false flag' in order to present left-wing terrorists or jihadists as the perpetrators. Right-wing extremist parties, which immediately condemn the weakness of the democratic state and demand a 'hard line' against crime, could profit from the feeling of insecurity. However, non-violent extremist movements such as the American alt-right and the identitarian movement talk of similar approaches on a metapolitical level. With the aid of targeted provocations and media stunts a 'strategic polarisation' of society is to be achieved in order to facilitate normalisation of their language and mainstreaming of their ideologies.

The aim is to force people who are undecided, moderate or even apolitical from the middle of society to decide on which side of the conflict they stand. The identitarians are especially good in the mainstream staging of their offline campaigns and linking online and offline activism, which differentiates them from previous extreme-right youth cultures in Germany.[9] Activists normally film their actions and immediately put them on YouTube to share them on social media. They are then further disseminated, 'liked' and commented on, which in turn can result in further offline support and a growing number of members (Hentges et al. 2014).

For an-Naji the media play a central role in maintaining the myth of the invulnerability of Western powers ('deceptive media halo') and are therefore a central weapon of both sides in a psychological war for the loyalty of the Muslim population. An-Naji naturally expects that the West will also use the media to carry out various campaigns of lies against the Mujahedin.

Accordingly, the Mujahedin must pursue a dual media strategy which targets two classes:

> *"(The first) class is the masses, in order to push a large number of them to join the jihad, offer positive support, and adopt a negative attitude toward those who do not join the ranks. The second class is the troops of the enemy who have lower salaries, in order to push them to join the ranks of the mujahids or at least to flee from the service of the enemy" (an-Naji 2004, 50–1).*

Right-wing extremist groups also increasingly pursue a dual media strategy: they attempt to gain the attention of traditional media with co-ordinated provocations in order to gain the attention of the masses; and in parallel they build up their media ecosystem and attempt to undermine the credibility of the established media. The idea of the *Lügenpresse* [lying press], often coupled with antisemitic conspiracy theories, is another concept shared by right-wing extremists and Islamists. (Hope not Hate 2017; Phillips 2018).

In *The Management of Savagery*, an-Naji makes it clear that the jihadist media strategy must be subjected to a plan, which breaks out of the 'prison of individualism' and is completely oriented to the masses. The masses are the future life insurance of the jihadist movement and the corresponding media plan must acknowledge faults in order to appear more credible and transparent than other media campaigns:

> *"Its specific target is to (motivate) crowds drawn from the masses to fly to the regions which we manage, particularly the youth after news of (our) transparency and truthfulness reaches them so that they may be fully aware of the loss of money, people, and worldly gains [lit. 'fruits']" (an-Naji 2004, 51).*

For an-Naji the media battle is equivalent to the local battlefield and has a corresponding significance in *The Management of Savagery*. It refers to the organisation of media committees with experts from various fields and details of Western media psychology.

In a similar manner to the Islamists, the media battle of right-wing extremists is primarily concerned with addressing the masses and expanding their audience. A leaked style guide from the world's largest neo-Nazi site, The Daily Stormer, explains that appearance is of primary importance for "spreading the idea of nationalism and antisemitism among the masses" (Marantz 2017). There are guidelines for the style of presenting quotations, and the form and even the tone of publications and public statements. For example, the importance of using simple psychological means to achieve positive messaging is emphasised for propaganda and media content to be effective:

> *"We are covering very negative content generally, but still as much effort as possible should be put into presenting a positive message. We should always claim we are winning, and should celebrate any wins with extreme exaggeration. This does not mean we downplay the enemy, just that we play up ourselves. We overestimate our influence" (Vox Popoli 2017).*

[9] *At the beginning of June 2018, many pages from the identitarian movement and their more prominent activists were barred on Facebook and Instagram.*

2.2 The art of recruiting

In 2009 a further central jihadist manual appeared on the internet. Written under the pseudonym Abu Amru Al Qa'idy, this work, called *A Course in the Art of Recruiting: a Graded, Practical Program for Recruiting via Individual Da'wa,* can be seen as an attempt to decentralise jihadist radicalisation by empowering already radicalised supporters in the target countries by means of easily used high level psychological methods.

This recruiting course has five central stages (described below). It includes among other things a quantification of partial successes as an aid for deciding when to initiate next steps and stages ('survey of successes'). In this detailed handout, Al Qaida (and later IS, which adop-ted the manual) attempts to bridge the critical distance between skilled and experienced recruiters from jihadist organisations and recruits (called 'candidates' in this manual). The course offers a step-by-step do-it-yourself guide for all interested supporters of the movement. Its target group for jihadist radicalisation is clearly stated: young, non-religious Muslims and converts with good education (secondary school or university) who live away from their home towns. In this way it hopes to address the especially 'pure souls' with high ideals, who are highly unlikely to be informants or members of intelligence services. The five stages of recruiting are described as a gradual introduction of the candidate to the self-recognition of militant jihad as an individual duty.

In the first stage ('Introduction') the focus is on resistance against injustice. Jihad or other religious topics should not be discussed. Driving a wedge between candidates, their family, environment or lifestyle should also be avoided. Under no circumstances should pressure be exerted. The recruiter should use general information material and create an initial basis of trust.

The second stage ('Coming closer') contains a specific plan for the candidate, with particular media content which is to be regularly consumed. These are not yet of a jihadist nature, but are oriented to current events and tragedies which make the need for resistance clear. The Israel–Palestine conflict is suitable for this, as Muslims are very unlikely to oppose criticism of Israel. Ideally, candidates should be contacted every day to participate in their lives, fulfil their needs, listen to them, be good to them and encourage them to persuade other people. Meals together and presents are emphasised as aids for relationship-building efforts. Now, Islamic topics should gradually be discussed. The aim of this stage is to learn all about the candidate (their interests, hobbies, environment, values).

In the third stage ('Awakening the faith') the recruiter should deliberately reinforce, praise and esteem the positive traits of the candidate. These positive traits are now linked to Islam and the new, true faith. Only now, in the third stage, does the recruiter start to talk about Islamic rites and duties (for example prayer), with great patience, in order to avoid incurring any mistrust from the candidate. It is recommended to talk more about paradise than hell when discussing religious topics.

Access to the candidate's world should by gained through participating in current events (for example discussing dreams, family experiences). It is also recommended that the recruiter and candidate together go on picnics, visit cemeteries (to consider death and the afterlife) and do good deeds.

The fourth stage ('Planting the concept') sets out the topic of jihad as an individual duty for the first time. Here, it is internalised that Islam stands above democracy and the family and that those in the family must be 'hated' for God (reference to the core Salafist principle 'al wala wal bara') if they are against the 'true faith'. Only now is the full breadth of jihadist propaganda reached and the candidate is presented as a foreigner who is under continuous attack in their environment, but can depend on a new, loyal group.

Finally, in the fifth stage ('Establishing the brigade') the recruiter convinces the candidate that she or he must carry out militant jihad of their own as a consequence of their faith (or new identity). The recruiter helps with practical arrangements and the central topic of conversations should be 'martyrs'. Right-wing extremist groups regularly share instructions for optimum recruiting and mobilisation processes, especially door-opening topics and hints on appearance.

For example, in order to address as many people as possible, the organisers of the Charlottesville protests specified an internal dress code and defined other visual details which would make them appear 'normal' and 'cool' in the media. Hence, explicit Nazi symbols and white clothing, which is traditionally associated with nationalism, was forbidden (Davey and Ebner 2017).

In the identitarian media guerrilla manual (Informationskrieg-Manual 2017), good introductory and door-opening topics are identified in order to convince so-called 'normies' – normal citizens – of their ideologies. This calculated, gradual radicalisation of average users on the internet is referred to by right-wing extremist activists as 'red-pilling' in reference to the cult film *Matrix*:[10]

> *"For red-pilling it is therefore advisable to start with topics which already have a certain acceptance, i.e. nothing controversial. Perfect topics and a gentle entry are political correctness, feminism, gender and their negative excesses. Many people have already noticed the destructive effects of these globalised social engineering methods. Anyone who is open to these topics is often open to more. This is so to speak a gateway drug. In an atmosphere of approval, people are more ready to open up to other topics."*

2.3 Case study 1: the IS media strategy

It is undisputed that the so-called IS as a terror organisation or "terrorist semi-state" (Honig and Yahel 2017) has raised the use of online and offline media for propaganda and recruiting purposes to a previously unknown level. Accordingly, much research has been carried out on special aspects of IS propaganda and its effect, for example on the IS offline media strategy (Winter 2016), the IS use of social media (Berger and Morgan 2015; Farwell 2014; Huey, Inch and Peladeau 2017; Milton 2016; Talbot 2015; Whiteside 2016) or IS print media (Colas 2016; Ingram 2016b; Musial 2016; Vergani and Bliuc 2015, 2017). It is unanimously agreed that IS knows how to produce and use its own and external media in a highly professional manner, both technically and strategically.

The suspicion that IS had a detailed media strategy was confirmed to the general public most recently in May 2015 in an IS video clip.The video with the title 'Media Operative, You Are a Mujahid, Too' was produced in the IS province Salahuddin (Salah al Din) in North Iraq and impressively presents the work of 'media warriors' as equivalent to physical combat. The printed version of a special media manual with the same title, published by the IS library al-Himma, was briefly available, and experts could only analyse the second edition after April 2016 (Winter 2017).

There are three central pillars for the propaganda strategy of IS in this media manual:

1) A positive and alternative narrative about IS;

2) Extensive refutation (counterspeech) of enemy propaganda; and

3) Targeted media attacks ('media projectiles').

As a part of this strategy, media work is presented as at least equivalent to, if not even more important than, the physical fight against the unbelievers. Western media are deliberately and consciously used as a weapon against the West. A core element of the IS media strategy is the emotional and theological underpinning of propaganda work, especially for supporters who for whatever reason have decided not to travel to Syria and Iraq or to participate in fighting and violence.

IS presents propaganda work as a continuous tradition, which can be traced back to the prophet Mohammed and together with grandiose hymns to the media warriors ('media mujahid') this activity is presented as having great importance and value for supporters.

The first pillar of the IS media strategy (positive and alternative narrative) aims to 'open the eyes' of the viewer and in particular to create pleasure and satisfaction.

[10] *In the science fiction film Matrix, the main character Neo is offered two pills, a red pill and a blue pill. If Neo swallows the red pill he becomes aware that he has previously lived in a computer simulation controlled by machines and not in true reality. The process of radicalisation of normal citizens is referred to as 'red-pilling' by right-wing extremists, because now the alleged true political reality is disclosed to them.*

The aim is to create an attractive IS brand which has value for the general public and not just for a special audience, and is therefore one of the most important innovations by IS:

> *"The Islamic State's foundational appeal is not rejection of the status quo or defiance in the face of tyranny. Rather, it is an offer of a positive alternative – a brand that presents a compre hensive solution without dwelling too much on the problem. This represents something of a shift in salafi-jihadist outreach and is the Islamic State's single most important innovation in the realm of strategic communications" (Winter 2017, 16).*

The second column (counterspeech) calls on IS media warriors to set up an arsenal of arguments and counter evidence to defend themselves against the "intellectual invasion" (Winter 2017, 17) of the unbelievers. The central reasoning for this in the IS media strategy is surprisingly close to right-wing extremist thought patterns:

> *"Ignorance will take root among the people and it would be but a few decades before this generation of fighters in the name of Allah the Almighty would be lost and you would not be able to find anyone to continue the journey. Even if you found some left, they would not be of the level required to manage the global conflict with the evil states of unbelief" (Winter 2017, 17).*

A kind of spiritual or mental genocide by 'colonisation of the hearts' of Muslims with the propaganda of the unbelievers is described, which ultimately undermines and dissolves the resistance and defensive capability of the Muslim ummah. Right-wing extremists use a similar argument when they refer to the degradation of the defensive capability or biological quality of the 'Arian race' through migration and multicultural society as a deliberate strategy by democrats to bring about the 'death of the race'. In both cases an existential crisis is described, which must be prevented at all costs.

While IS sees the spiritual power necessary for effective resistance against unbelievers as endangered, right-wing extremists fear the deliberate destruction of the physical defensive capability of the 'Arian race'.

The third column (media projectiles) consists of the use of IS and external media as a psychological weapon to support or even replace military or terrorist operations (Winter 2017, 18). Everything should be done to annoy, provoke and emotionally drain the enemy and induce knee-jerk reactions. This explains the efforts of the IS media organisation to produce particularly shocking images for certain target groups (for example especially horrific executions and the use of children in propaganda). With these components, the IS media strategy is at present the most detailed instruction for use of the media for recruiting and propaganda work in the jihadist scene.

2.4 Case study 2: the media strategy of the international alt-right

The international alt-right, which has now reached Germany in the form of right-wing extremist troll armies and an increasingly digitally active new right, has to some extent an even more sophisticated media strategy than IS. In an analysis, the Data & Society Research Institute in New York showed how between 2016 and 2018 the alt-right succeeded in deliberately instrumentalising the new media ecosystems for their purposes. Those working in traditional media faced the dilemma of believing in their duty to inform, without at the same time increasing the public visibility and legitimacy of right-wing extremists (Phillips 2018).

Younger and older journalists had differing levels of knowledge about trolling and manipulation of online media and there was inconsistent handling of it (Phillips 2018). Therefore, many online activists were free to develop and test new strategies for manipulating and influencing journalists.

These are some of the especially popular methods used in the past few years:

a) 'Triggering' – attempting to cause an over-reaction by 'mainstream media' through provocative words or actions;

b) 'Doxxing' – disclosure of personal information to intimidate journalists; und

c) 'Source hacking' – deliberate sharing of false information with credible sources such as research institutes or local media, which are then quoted nationally by journalists (Donovan 2018).

The aim of these activities is to distort public perception and influence political discussion. Hence, over the past few years political fringe groups have increasingly combined to achieve their aims with the aid of co-ordinated troll, hate and disinformation campaigns on social media. Since the American alt-right celebrated Donald Trump's election victory as a result of their large-scale online troll campaigns, right-wing extremist activists in Europe have copied their tactics in order to influence political debate in their countries in favour of right-wing populist parties. (Ebner and Davey 2018).

These methods have now arrived in Germany. In encoded chat groups on Discord, thousands of German right-wing extremists exchange information about how to win their 'information war' against the political mainstream and the 'lying press'. Germany's (then) largest troll factory, Reconquista Germanica was founded shortly before the parliamentary elections in 2017 and gained 7,000 members within a few weeks, though several later left. Reconquista Germanica still has several hundred active members, and sees itself as an electronic army, which follows strict hierarchical ranking structures and communicates with military vocabulary.

The members of Reconquista Germanica range from patriots, AfD members and identitarians to Reichsbürger and neo-Nazis. Every day these self-appointed 'generals' and 'officers' look for different targets – from refugees to television presenters – and then publish the hashtags and times in order to outwit the algorithms of social media platforms.

To ensure that their provocative posts and comments land in the top trends and are visible to as many people as possible, they also use fake user accounts and infiltration techniques. "Act as if you are a normal account, which posts about football, BBQs, parties, Karl Marx or similar topics", the media guerrilla manual of the identitarians states (Ebner and Davey 2018).

Detailed instructions for techniques for engaging in the 'information war' are given. For example, a 'sniper mission' is a targeted verbal attack on a 'major enemy account' with the aim of discrediting and derogating the person behind it. In contrast, the instruction for undertaking a 'massive air strike' states: "Directly target the accounts of opponents: politicians, celebrities, state radio etc. and fill up the comments. As has been said: change the account every 2–3 tweets." Trolling actions become a kind of computer game in which there is a fun factor of hounding minorities or political opponents. Among the descriptions and orders for the 'information war', in the right-wing extremist troll factories such as Reconquista Germanica, Nazi symbols, Holocaust denial and hints of a race war can be found. "Because of this, I especially urge you: get a knife," wrote a user in the Crisis Prevention Centre chat.

Instructions for making stun guns from hair trimmers and recommendations for firearms circulate within the group. Although Discord has barred the Reconquista Germanica channel several times, the group continues to re-appear under new names. It now has strict recruiting processes, in which background checks on the identity and ideology of applicants are carried out.

The hate campaigns take place in the virtual world, but are not without consequences in the real world. They can inspire violent actions, intimidate opponents and influence elections. In the two weeks before the German parliamentary elections in 2017, activists of Reconquista Germanica managed to place seven of their hashtags (including #TraudichDeutschland [trust in Germany], #nichtmeinekanzlerin [not my chancellor], #merkelmussweg [Merkel must go] and #reconquista in the top 20 hashtags in Germany).

Because of their co-ordination techniques, right-wing extremist fringe groups have a virtual monopoly of hate comments in the comment columns: 5% of all active accounts are responsible for 50% of the 'likes' for hate comments. This distortion of perception generates increasing political pressure and damages the online discussion culture (Kreißel et al. 2018).

With these methods, organised trolls, whether right-wing extremists or Islamists, can achieve the media 'tipping point' at which it is hardly possible for traditional media to ignore their campaigns and there can be negative effects on the credibility of the mainstream media (Phillips 2018). The NetzDG is therefore not a suitable means of counteracting such troll strategies. As co-ordinated hate and disinformation campaigns can go viral and reach their target audience within a very short time, promotion of civil society resilience and informing the public about these tactics are considerably more important countermeasures than removing content.

References

1. *an-Naji, A. B. (2004) The Management of Savagery: The Most Critical Stage Through Which the Islamic Nation Will Pass, translated by William McCants, John M. Olin Institute for Strategic Studies at Harvard University.*

2. *Atwan, A. B. (2015) Islamic State: The Digital Caliphate, Saqi.*

3. *Bale, J. M. (1994) The' Black' Terrorist International: Neo-fascist Paramilitary Networks and the 'Strategy of Tension' in Italy, 1968–1974, PhD, University of California, Berkeley.*

4. *Berger, J. M. and J. Morgan (2015) The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter, Brookings Institution, http://www.brookings.edu/~/media/research/fies/papers/2015/03/isis-twitter-census-berger-morgan/isis_twitter_census_berger_morgan.pdf [16.07.18].*

5. *Ciovacco, C. J. (2009) The Contours of Al Qaeda's Media Strategy, Studies in Conflict & Terrorism, 32(10), 853–75.*

6. *Colas, B. (2016) What Does Dabiq Do? ISIS Hermeneutics and Organizational Fractures within Dabiq Magazine, Studies in Conflict & Terrorism, 1–39.*

7. *Davey, J. and J. Ebner (2017) The Fringe Insurgency: Connectivity, Convergence and Mainstreaming of the Extreme Right, Institute for Strategic Dialogue, http://www.isdglobal.org/wp-content/uploads/2017/10/The-Fringe-Insurgency-221017.pdf [16.07.18].*

8. *Donovan, J. (2018) How White Nationalists Fooled the Media about the Florida Shooter, Data & Society Research Foundation, https://datasociety.net/output/how-white-nationalists-fooled-the-media-about-the-florida-shooter/ [16.07.18].*

9. *Ebner, J. and J. Davey (2018) Mainstreaming Mussolini: How the Extreme Right Attempted to 'Make Italy Great Again' in the 2018 Italian Election, Institute for Strategic Dialogue, http://www.isdglobal.org/wp-content/uploads/2018/03/Mainstreaming-Mussolini-Report-28.03.18.pdf [16.07.18].*

10. *Farwell, J. P. (2014) The Media Strategy of ISIS, Survival, 56(6), 49–55.*

11. *Gråtrud, H. (2016) Islamic State Nasheeds as Messaging Tools, Studies in Conflict & Terrorism, 39(12), 1050–70.*

12. *Hentges, G. et al. (2014) Die Identitäre Bewegung Deutschland (IBD): Bewegung oder virtuelles Phänomen? [The German Identitarian Movement: Movement or Virtual Phenomenon?], Forschungsjournal soziale Bewegunen – PLUS, http://forschungsjournal.de/sites/default/files/fjsbplus/fjsb-plus_2014-3_hentges_koekgiran_nottbohm_x.pdf [16.07.18].*

13. *Honig, O. and I. Yahel (2017) A Fifth Wave of Terrorism? The Emergence of Terrorist Semi-States, Terrorism and Political Violence, 1–19.*

14. *Hope not Hate (2017) The International Alt-Right, https://alternativeright.hopenothate.com.*

15. *Hu Huey, L., R. Inch and H. Peladeau (2017) '@ Me if You Need Shoutout': Exploring Women's Roles in Islamic State Twitter Networks, Studies in Conflict & Terrorism, 1–19.*

16. *Informationskrieg-Manual V 4.0: Handbuch für Mediengueriillas [Information War Manual V 4.0: Manual for Media Guerrillas] (2017) Publication of the Identitarian Movement.*

17. Ingram, H. J. (2016a) An Analysis of Inspire and Dabiq: Lessons from AQAP and Islamic State's Propaganda War, Studies in Conflict & Terrorism, 40(5), 357–75.

18. Ingram, H. J. (2016b) An analysis of Islamic State's Dabiq magazine, Australian Journal of Political Science, 51(3), 458–77.

19. Jenkins, P. (1990) Strategy of Tension: The Belgian Terrorist Crisis 1982–1986, Studies in Conflict & Terrorism, 13(4–5), 299–309.

20. Kessler, J. (1999) Poisoning the Web: Hatred Online: an ADL Report on Internet Bigotry, Extremism and Violence, Featuring 10 Frequently asked Questions about the Law and Hate on the Internet Anti-Defamation League.

21. Kreißel, P., J. Ebner, A. Urban and J. Guhl (2018) Hate at the Touch of a Button: Right-wing Extremist Troll Factories and the Ecosystem of Coordinated Hate Campaigns on the Internet, Institute for Strategic Dialogue and #ichbinhier, https://www.isdglobal.org/wp-content/uploads/2018/07/ISD_Ich_Bin_Hier_2.pdf [16.07.18].

22. Lakomy, M. (2017) Let's Play a Video Game: Jihadi Propaganda in the World of Electronic Entertainment, Studies in Conflict & Terrorism, 1–24.

23. Lesaca, J. (2015) On Social Media ISIS Uses Modern Cultural Images to Spread Anti-modern Values, Brookings, https://www.brookings.edu/blog/techtank/2015/09/24/on-social-media-isis-uses-modern-cultural-images-to-spread-anti-modern- values/ [16.07.18].

24. Marantz, A. (2017) Inside the Daily Stormer's Style Guide, New Yorker, https://www.newyorker.com/magazine/2018/01/15/inside-the-daily-stormers-style-guide [16.07.18].

25. Milton, D. (2016) Communication Breakdown: Unraveling the Islamic State's Media Efforts, West Point, https://www.ctc.usma.edu/posts/communication-breakdown-unraveling-the-islamic-states-media-efforts [16.07.18].

26. Musial, J. (2016) 'My Muslim sister, indeed you are a mujahidah' – Narratives in the Propaganda of the Islamic State to Address and Radicalize Western Women: An Exemplary Analysis of the Online Magazine Dabiq, Journal for Deradicalization, Winter 2016/17, 9, 39–100.

27. Phillips, W. (2018) The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators Online, Data & Society Research Institute, https://datasociety.net/output/oxygen-of-amplification/ [16.07.18].

28. Skillicorn, D. B. and E. F. Reid (2014) Language Use in the Jihadist Magazines Inspire and Azan, Security Informatics, 3(9), 1–16.

29. Talbot, D. (2015) Fighting ISIS Online. MIT Technology Review. https://www.technologyreview.com/s/541801/fighting-isis-online/ [16.07.18].

30. Vox Popoli (2017) The 'Andrew Anglin' Style Guide, https://voxday.blogspot.com/2017/09/the-andrew-anglin-style-guide.html [16.07.18].

31. Torres-Soriano, M. R. (2016) The Caliphate is not a Tweet Away: The Social Media Experience of al Qaeda in the Islamic Maghreb, Studies in Conflict & Terrorism, 39(11), 968–81.

32. Vergani, M. and A. M. Bliuc (2015) The Evolution of the ISIS' Language: A Quantitative Analysis of the Language of the First Year of Dabiq magazine, Sicurezza, Terrorismo e Società, 2, 7–20.

33. Vergani, M. and A. M. Bliuc (2017) The Language of New Terrorism: Differences in Psychological Dimensions of Communication in Dabiq and Inspire, Journal of Language and Social Psychology.

34. Whiteside, C. (2016) Lighting the Path: The Evolution of the Islamic State Media Enterprise (2003–2016), International Centre for Counter-Terrorism, https://icct.nl/publication/lighting- the-path-the-evolution-of-the-islamic-state-media- enterprise-2003-2016/ [16.07.18].

35. Winter, C. (2016) ISIS' Offline Propaganda Strategy, Brookings, 31 March, https://www.brookings.edu/blog/markaz/2016/03/31/isis-offline-propaganda-strategy/ [16.07.18].

36. Winter, C. (2017) Media Jihad: The Islamic State's Doctrine for Information Warfare, International Centre for the Study of Radicalisation and Political Violence, http://icsr. info/2017/02/icsr-report-media-jihad-islamic-states- doctrine-information-warfare/ [16.07.18].

3. Filter bubbles: how do filter bubbles affect (political) opinion, taking personality into account?

By Prof. Dr. Christian Montag

Abstract
In the discussions about the role of social media in online radicalisation and political polarisation, the concept of 'filter bubbles' has been repeatedly used as an explanation. In his article, Christian Montag points out that there are still gaps in our current knowledge over the causes and effects of filter bubbles. Nonetheless, he advocates that the topic of online filter bubbles should be taken seriously. Above all, Montag considers a differential psychological approach to be promising for further research into the causes and political consequences of filter bubbles.

3.1 An anecdotal observation from everyday life: the normalisation of the weirdo

I was recently travelling by train from Ulm to Vienna. A middle-aged man wearing a t-shirt with the slogan 'Good Morning Vietnam' was seated opposite me. Under the slogan there was a large, bright red Vietnamese flag. In the middle of the flag was a large yellow star. I looked at the t-shirt with interest.

It was not long before we started a conversation. My pleasant travelling companion soon explained that he had a special hobby. More precisely, he collected t-shirts with flags. No matter where his friends travelled, he asked them to bring a t-shirt with a flag from the country which they travelled to. In this way, my travelling companion had obtained a large number of t-shirts with flags.

I would like to use this story of everyday life to illustrate the influence of the internet and in particular of social media on our (political) opinions and attitudes. I hope that it will become clear below why this short story about my train companion is important for this complex topic. For the sake of simplicity, imagine that my own hobby was collecting t-shirts with a flags as already described.

In a thought experiment we will now travel back in time to before the existence of the internet. In that age it would have been difficult to find like-minded people with a hobby such as collecting t-shirts with flags, as the basic rate of people with such a hobby would be rather small. The term 'basic rate' describes how frequently a certain characteristic occurs in a population or examined sample. Unfortunately, I have not found any information about the actual number of people with this hobby in my internet research, but I would like to suggest that 0.001% of the population collect these t-shirts, and use it for the following calculation: Germany has some 80 million people, and if 0.001% of the population had this hobby there would be exactly 800 people collecting t-shirts with flags.

At present, Germany has 2,060 towns (estimates vary). For the sake of simplicity, in the calculation we will assume that all towns in our example are of the same size. Under these conditions, in the pre-internet age I would almost certainly not have met any other people with the hobby of collecting t-shirts with flags in my town. With the suggested figures, we have just 0.39 people with this hobby for 38,834.95 inhabitants in each town. If I was interested in finding out the basic rate of collectors of t-shirts with flags I would have found out that I have an unusual hobby.

Following the rapid development of the world wide web and spread of social media, perception of unusual hobbies and therefore of the perceived basic rate has changed. Since the beginning of the world wide web, fan groups have formed for various things, and perhaps somewhere there is also a fan group on the the internet in which people have lively discussions about collecting t-shirts with flags.

As has been said, in our example, in Germany, which has approximately 80 million people, up to 800 people could theoretically be found online who pursue this unusual hobby and perhaps communicate about it daily on the internet.

On a psychological level, by interacting with so many like-minded people, the perceived basic rate of people who collect t-shirts with flags would increase in comparison with those who collected these t-shirts before the age before the internet. I am not sure how much higher the subjectively perceived basic rate actually is due to the internet. This is still to be researched. With the illustrated example I would simply like to make it clear that in the age before the internet, I would have been relatively alone with my unusual hobby. Through daily communication with people on the internet who share my hobby, my hobby nowadays no longer seems to be special or even strange.

It is worth mentioning that in the example cited here people actively use the internet to find like-minded persons. This idea is important when we talk about filter bubbles and their effects later on. The difference is that most of the filter bubble debate concerns effects over which users only have a limited active influence.

At the end of the first part of this article we would like to consider that some people with a potential interest in the said hobby would not even have thought about its existence before the age of the internet. But in the internet age, through social media and fan pages, it is highly probable that these people will find others who collect t-shirts with flags through social media and fan pages, or may become fans of t-shirts with flags through their initial contact with this hobby. In this way, the community of t-shirts with flags fans will gradually increase, possibly exponentially at some point. Perhaps after one year the number of fans will no longer be 800 but rather 850 people. In a few years it could possibly rise to 1,500. From being an unusual and possibly even strange or weird hobby for weirdos, through the possibilities of the internet it could become something completely 'normal'.

The example discussed here describes a harmless hobby and in any case it is a good thing if people can get to know each other online and talk about their interests. Also, topics falsely associated with a stigma, such as psychological disorders, could be rehabilitated because many affected people talk about them online, but what happens if people talk to like-minded persons with dangerous ideas and become radicalised?

What happens if a person increasingly and at some point only talks to people online who support the use of weapons? Do they eventually conclude that there are more like-minded people than the number who actually exist in society? Could this lead them to become more confident and radical in their positions?

Before I start to talk about the actual topic – filter bubbles and their influence on (political) attitudes – I have deliberately chosen a harmless example such as the hobby of collecting t-shirts with flags in the brief introduction. From this simple example it is already clear that a one-sided and frequent occupation with a harmless topic can change one's perception of the actual basic rate of a characteristic in the population. What now happens to my (political) attitude if, in addition, so-called filter bubbles come into play?

3.2 What are filter bubbles?

The term 'filter bubble' was coined by Pariser (2011) who introduced it into literature. In order to understand the term, we must recognise that on the internet people can be exposed to a flow of biased news reports. Before this is examined in greater detail, it must be pointed out that many internet users decide for themselves only to read or subscribe to particular news messages. For example, a person can decide only to read news from *Die Tageszeitung* (rather leftist) or F*rankfurter Allgemeine Zeitung* (rather right-wing) and to ignore everything else.

In English-speaking regions this would be the online version of the *New York Times* (rather left–liberal) versus Fox News or even Breitbart (right-hand edge of the spectrum). In addition to this self-chosen form of preselection of news by the user, since the recent Cambridge Analytica scandal there is discussion about the influence of news feeds on people's attitudes. Many academics fear that users could be manipulated by news feeds. A search engine or a social media platform such as Facebook could filter the news it shows someone in a way that corresponds with that person's interests, as demonstrated by their search history or the 'likes' they have set.

To find out what a person's preferences are operators of online platforms hope to keep users on their platform for as long as possible to increase the digital data they leave online on it, which in turn can be monetarised.

The presentation of a news feed which is adapted to a user's preferences is especially problematic if this generates a more extreme or restricted online discourse. In short: a xenophobic person could become even more xenophobic through a news feed that is designed by algorithms, as this person would increasingly come to the opinion that deviating opinions hardly exist in the wider population and that their ideology is shared by everyone. This makes it easier for a person to take on an extreme attitude, as they perceive their views to be mainstream. The actual basic rate is correspondingly overesti-mated – as demonstrated in the example above of the person who collects t-shirts with flags.

Some researchers believe that an essential component of any democracy is that pluralistic standpoints are debated in a political discourse. Pre-filtering of news could result in bubbles, in which people do not learn anything new (they are only shown or confirmed what they have 'liked' in the past) and only their own opinion is echoed.

In the worst scenarios, filter bubbles would possibly even facilitate the radicalisation of certain groups of the population, and knowledge of the actual diverse opinion and mood existing in a society could greatly reduce. In the context of filter bubbles, in English the term 'echo chamber' is often used. An echo chamber is not the same as a filter bubble, but much broader in scope – an environment in which someone encounters only beliefs or opinions that coincide with those they hold, so their existing views are reinforced and they do not consider alternative ideas. This process of having one's views reinforced by only encountering beliefs that coincide with one's own is triggered by an algorithm for filter bubbles.

When internet users meet like-minded people in online forums, in order to share their often preconceived and above all similar opinions, they exist in echo chambers. Many people find it considerably less strenuous to talk to people with the same opinion as their own than to those with different positions. While filter bubbles are an internet phenomenon, echo chambers existed before the internet age.

3.3 How dangerous are filter bubbles today?

A review by Zuiderveen Borgesius et al. (2016) considered the question of the strength of current empirical evidence for the negative effects of filter bubbles on political opinions. Although the authors cite several cases that suggest filter bubbles can actually result in radicalisation (for example Knobloch-Westerwick and Meng 2011; Stroud 2010), overall the findings appear to suggest that although there are measurable and statistically significant indications that filter bubbles have negative effects on people's political opinions, they are slight or moderate. The authors summarise that "at present there is no empirical evidence that warrants any strong worries about filter bubbles" (Zuiderveen Borgesisus et al. 2016, 10).

These findings should be questioned further, however. First, the authors themselves point out that the majority of previous studies are USA centred and it is questionable whether their results can be transferred to other political systems such as the multiple party system existing in Germany.

Second, there has not been much research on this subject and more recent studies such as one by Flaxman et al. (2016) find that filter bubbles influence search engines and social media and affect the "mean ideological difference" between the groups investigated (2016, 298). This study examines whether the ideological distance between certain groups of people changes. However, the observed magnitudes of the effects of filter bubbles are once again weak.

Overall, because there are few studies on the effects of filter bubbles, at present no conclusion can be reached on these questions, and there are several factors which in my opinion have not been sufficiently considered up to now in this research. These factors will be briefly stated and explained below in the context of current studies from adjacent fields of research.

3.4 The importance of a differential psychological approach for research on the effect of filter bubbles

Do the possible effects of filter bubbles affect the political opinions of all people in the same way? The studies cited up to now generally investigated whether filter bubbles have a negative effect on the political attitude of a person and could possibly have an effect on the radicalisation of groups of people. As has been mentioned, statistically relevant findings could be derived, but are in the weak to moderate range.

However, it can certainly not be concluded that there is no danger to society from filter bubbles, as the radicalisation of even a small group of people can cause great problems and damage to a society, but there is also no reason for panic.

Unfortunately, there are few studies that consider the differential psychology of filter bubbles. Differential psychology is a discipline that attempts to understand the inter-individual differences between people (Montag 2016). Among other things, it attempts to answer questions such as: Why am I as I am? Why do people differ from each other?

Important concepts of differential psychology are personality and a person's cognitive abilities. A person's personality – their stable characteristics – can be observed over a long period and to some extent in different everyday situations to uncover emotional or motivational (behavioural) tendencies, and cognitive thought patterns (Montag and Panksepp 2017; see also Mischel and Shoda 1995). For example, one can find out how sociable or shy a person is in their dealings with other people.

In the context of the present set of questions, a differential psychological investigation could be undertaken to find out whether a person tends to consume only a few sources of news or perhaps is especially susceptible to biased reporting.

One of the central personality models suggests there are five personality traits, which form the acronym OCEAN, explained below (McCrae and John 1992). With the aid of a lexical approach using speech analysis, personality psychologists have derived these five personality traits which can be used to categorise and globally describe any person.

This is the OCEAN personality model:

- **O** penness to experience – people who like to try new things and are intellectual and curious.
- **C** onscientiousness – dependable people who are punctual and careful.
- **E** xtraversion – people who are sociable, lively and assertive.
- **A** greeableness – people who are warm-hearted and caring.
- **N** euroticism – people characterised among other things by anxiety, obsessiveness and a tendency to depressive moods.

Each of these five personality traits is a dimensional construct – a person tends somewhat more to extraversion or its opposite, introversion.

Why is the consideration of personality of general scientific interest? Why must personality be considered in the context of the possible negative effects of filter bubbles on the political attitudes of a person? On one hand the importance of personality variables relates to the fact that personality traits are associated with many important life variables (Montag 2016).

For example, conscientiousness is associated with a healthier lifestyle (Bogg and Roberts 2004), or an extreme degree of extraversion is associated with greater success as a salesperson (Grant 2013). On the other hand there are studies which show that certain personality traits tend to be associated with particular political attitudes. For example, a study by Lee et al. (2010) has shown that greater social conformity is associated with lower values of openness.

Older studies such as that by Pratto et al. (1994) found that central personality variables such as a person's gender influence social dominance orientation: men achieve higher values than women. They also found that highly pronounced social dominance orientation is characterised by affirmation of statements such as "Inferior groups should stay in their places" (Choma and Hanoch 2017; 289).

Echoing the result of the 2016 US election, a study by Choma and Hanoch (2017) showed that high values for social dominance orientation and right-wing authoritarianism (for example that the country must be protected against moral degradation) could predict whether people are Trump supporters. This study showed that lower cognitive abilities are associated with higher social dominance orientation and right-wing authoritarianism. Cognitive capabilities must therefore be analysed in the context of the effect of filter bubbles on various user groups.

In line with this finding, a new study by Zmigrod et al. (2018) showed that Brexit supporters had less cognitive flexibility than Brexit opponents in an experimental setting.

3.5 Psychological profiling and filter bubbles

Over the past few years several studies have shown that psychological profiling or digital phenotyping is possible by studying the digital trails that people leave when using digital devices such as smartphones or browsers on a desktop computer (Montag et al. 2016). This type of research belongs to the field of psychoinformatics; psychological diagnosis can also be carried out using IT or computer science methods (Markowetz et al. 2014).

Studies have found that 'likes' on Facebook provide a great deal of information on personal variables such as gender, political orientation or sexual orientation (Kosinski et al. 2013).

Prediction of personality traits can also be carried out by studying 'likes' on Facebook (but not not yet sufficiently accurately at an individual level), and smartphone use, for example how long someone stays on WhatsApp or Facebook (Montag et al. 2015).

One study was even able to demonstrate that there is a relationship between the brain volume of nucleus accumbens and the duration or frequency of Facebook use on smartphones (Montag et al. 2017). Simply put, the nucleus accumbens is the 'reward system' of the brain. Lower volumes of grey matter in this area of the brain are associated with longer or more frequent use of Facebook on smartphones, so one could predict the level of Facebook use from brain scans. However, this can only be done at group level.

Finally, in addition to this interesting combination of biological and psychological information, it is noted that personality diagnosis can be carried out through text mining (Iliev et al. 2015) or analysis of vocabulary a person uses on social media channels. Studies such as those by Schwartz et al. (2013) and Kern et al. (2014) suggest that personality traits can be deduced from the vocabulary used. For example, people who frequently use the word 'kill' or expletives in communicating on Facebook proved to be less agreeable than people who don't. This type of data is interesting when investigating the effect of filter bubbles. Matz et al. (2017) suggested that a precise match between personality and advertising messages can considerably increase click and purchase rates. This applied in particular if there is a match between the personality traits extraversion or openness and a corresponding advertising message. Accordingly, a news feed tailored to a person's personality could certainly reinforce the filter bubble effects described above.

In summary, it can therefore be demonstrated that not only can internet users be thoroughly psychologically examined on the basis of their digital trails, but also that these users can be incentivised to read certain messages via specially tailored news, make purchases or even make a particular cross in a polling booth.

Further research is needed to find out how strong these effects actually are in everyday life.

3.6 Further reasons to take the potential effects of filter bubbles seriously

Following the data scandal with Facebook and Cambridge Analytica, there is increasing debate as to the actual power of the data obtained and evaluated by psycho-informatic methods. Some research suggests that fake news is transported considerably faster and further than true news – especially via social media (Vosoughi et al. 2018). Possibly false information is especially powerful for the creation and maintenance of filter bubbles. Vosoughi et al. suggest that fake news has an especially high novelty character, particularly in a political setting. People respond quickly to new information because in the course of evolutionary human history quick responses helped to ensure survival. This could explain the rapid spread of this kind of information.

3.7 What is to be done?

Even though there are already initial insights in research into the possible complex effects of filter bubbles on our society, scientists are still only beginning to understand them. This important research is made more difficult by the often rapidly changing infrastructure of the platforms, which are frequently closed to scientists.

Therefore, I make the following recommendations:

- Platforms such as Facebook should be immediately opened up to scientists, for example in order to investigate whether fake news can be successfully restricted by the operator (as has already happened in the context of the Social Science One programme to examine the effects of social media on elections). Protection of private data must be ensured and verified.
- Studies on platforms such as Facebook should be categorised as safe by an ethics commission before they start. This should reduce the probability of data scandals, and users of Facebook and similar platforms would have greater protection against manipulation.
- Psychological mechanisms that suggest the creation of a filter bubble should be better determined, and suitable mechanisms must be created to reduce the possible effects of filter bubbles on radicalisation, even of small groups.
- Differential psychological approaches could be of special importance for understanding whether and why particular groups of people are more susceptible to the effects of filter bubbles than others.
- Other business models for using social media should be considered. At present, users normally pay for a service with their data. If platforms such as Facebook could be financed by other means than advertising, for example through a monthly user fee, the news feed could be provided with more news from friends etc.
- The news feed should present more balanced news items, checked in advance for accuracy, from the entire democratic spectrum.

Bibliography and references

1. *Bakshy, E., S. Messing and L. A. Adamic (2015) Exposure to Ideologically Diverse News and Opinion on Facebook, Science, 348(6239), 1130–2.*
2. *Bogg, T. and B. W. Roberts (2004) Conscientiousness and Health-Related Behaviors: A Meta-analysis of the Leading Behavioral Contributors to Mortality, Psychological Bulletin, 130(6), 887–919.*
3. *Choma, B. L. and Y. Hanoch (2017) Cognitive Ability and Authoritarianism: Understanding Support for Trump and Clinton, Personality and Individual Differences, 106, 287–91.*

4. *Grant, A. M. (2013) Rethinking the Extraverted Sales Ideal: The Ambivert Advantage, Psychological Science, 24(6), 1024–30.*

5. *Flaxman, S., S. Goel and J. M. Rao (2016) Filter Bubbles, Echo Chambers, and Online News Consumption, Public Opinion Quarterly, 80(S1), 298–320.*

6. *Iliev, R., M. Dehghani and E. Sagi (2015) Automated Text Analysis in Psychology: Methods, Applications, and Future Developments, Language and Cognition, 7(2), 265–90.*

7. *Kern, M. L., J. C. Eichstaedt, H. A. Schwartz, L. Dziurzynski, L. H. Ungar, D. J. Stillwell and M. E. Seligman (2014) The Online Social Self: An Open Vocabulary Approach to Personality, Assessment, 21(2), 158–69.*

8. *Knobloch-Westerwick, S. and J. Meng (2011) Reinforcement of the Political Self Through Selective Exposure to Political Messages, Journal of Communication, 61(2), 349–68.*

9. *Kosinski, M., D. Stillwell and T. Graepel (2013) Private Traits and Attributes are Predictable from Digital Records of Human Behavior, Proceedings of the National Academy of Sciences, 110(15), 5802–5.*

10. *Kosinski, M., S. C. Matz, S. D. Gosling, V. Popov and D. Stillwell (2015) Facebook as a Research Tool for the Social Sciences: Opportunities, Challenges, Ethical Considerations, and Practical Guidelines, American Psychologist, 70(6), 543–56.*

11. *Lee, K., M. C. Ashton, B. Ogunfowora, J. S. Bourdage and K. H. Shin (2010) The Personality Bases of Socio-Political Attitudes: The Role of Honesty: Humility and Openness to Experience, Journal of Research in Personality, 44(1), 115–19.*

12. *Markowetz, A., K. Błaszkiewicz, C. Montag, C. Switala and Schlaepfer (2014) Psycho-informatics: Big Data Shaping Modern Psychometrics, Medical Hypotheses, 82(4), 405–11.*

13. *Matz, S. C. and O. Netzer (2017) Using Big Data as a Window Into Consumers' Psychology, Current Opinion in Behavioral Sciences, 18, 7–12.*

14. *Matz, S. C., M. Kosinski, G. Nave and D. J. Stillwell (2017) Psychological Targeting as an Effective Approach to Digital Mass Persuasion, Proceedings of the National Academy of Sciences, 114(48), 12714–19.*

15. *McCrae, R. R. and O. P. John (1992) An Introduction to the Five-factor Model and its Applications, Journal of Personality, 60(2), 175–215.*

16. *Mischel, W. and Y. Shoda (1995) A Cognitive-affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure, Psychological Review, 102(2), 246–68.*

17. *Montag, C., K. Błaszkiewicz, R. Sariyska, B. Lachmann, I. Andone, B. Trendafilov and A. Markowetz (2015) Smartphone Usage in the 21st century: Who Is Active on WhatsApp?, BMC Research Notes, 8(1), 331.*

18. *Montag, C. (2016) Persönlichkeit: Auf der Suche nach unserer Individualität [Personality – In search of our individuality], Springer-Verlag.*

19. *Montag, C., É. Duke and A. Markowetz (2016) Toward Psychoinformatics: Computer Science Meets Psychology, Computational and Mathematical Methods in Medicine, 2016.*

20. *Montag, C. and J. Panksepp (2017) Primary Emotional Systems and Personality: An Evolutionary Perspective, Frontiers in Psychology, 8, 464.*

21. *Montag, C., A. Markowetz, K. Blaszkiewicz, I. Andone, B. Lachmann, R. Sariyska, B. Trendafilov, M. Eibes, J. Kolb, M. Reuter and B. Weber, (2017) Facebook Usage on Smartphones and Gray Matter Volume of the Nucleus Accumbens, Behavioural Brain Research, 329, 221–8.*

22. *Montag, C. and S. Diefenbach (2018) Towards Homo Digitalis: Important Research Issues for Psychology and the Neurosciences at the Dawn of the Internet of Things and the Digital Society, Sustainability, 10(2), 415.*

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

24. *Pratto, F., J. Sidanius, L. M. Stallworth and B. F. Malle (1994) Social Dominance Orientation: A Personality Variable Predicting Social and Political attitudes, Journal of Personality and Social Psychology, 67(4), 741–63.*

25. *Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal and L. H. Ungar (2013) Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach, PloS One, 8(9), e73791.*

26. *Stroud, N. J. (2010) Polarization and Partisan Selective Exposure, Journal of Communication, 60(3), 556–76.*

27. *Vosoughi, S., D. Roy and S. Aral (2018) The Spread of True and False News Online, Science, 359(6380), 1146, 1151.*

28. *Zmigrod, L., Rentfrow, P. J. and T. W. Robbins (2018) Cognitive Underpinnings of Nationalistic Ideology in the Context of Brexit, Proceedings of the National Academy of Sciences, 115(19) E4532–40.*

29. *Zuiderveen Borgesius, F., D. Trilling, J. Moeller, B. Bodó, C. H. de Vreese and N. Helberger (2016) Should We Worry About Filter Bubbles?, Internet Policy Review, 5(1).*

4. Disinformation: what role does disinformation play for hate speech and extremism on the internet and what measures have social media companies taken to combat it?

By Karolin Schwarz and Josef Holnburger

Abstract
Targeted disinformation plays an important role in the dissemination of hate speech and extremist ideologies on social media. In the political arena, so-called 'fake news' has become a controversial and frequently politically instrumentalised topic owing to its possible role in influencing democratic elections. Karolin Schwarz and Josef Holnburger look objectively at the facts – how widespread are disinformation campaigns, how successfully do they spread and what measures have social media platforms introduced until now in order to prevent disinformation?

The term 'fake news' reached prominence recently during the 2016 US presidential election campaign. The large amount of false information disseminated on social media, which mainly presented Donald Trump in a positive light (Allcott and Gentzkow 2017, 212), dominated the election campaign in the USA. An analysis by the Buzzfeed author Craig Silverman showed that disinformation disseminated during the election campaign was more frequently liked, shared and commented on than all the articles on the 19 news pages on Facebook with the largest reach (Silverman 2016). Hannah Parkinson (2016), author for the Guardian, and Max Read (2016), editor-in-chief of the *New York Magazine*, even feared that disinformation played the decisive role in the outcome of the election.

In contrast, the comprehensive assessment of the dissemination of disinformation during the US election campaign by Hunt Allcott and Matthew Gentzkow (2017) shows a less dramatic picture. The authors investigated the dissemination of disinformation articles by means of a representative survey: approximately 15% of those questioned stated that they had seen the disinformation headlines selected by Allcott and Gentzkow during the election campaign, but just 8% considered the communicated content to be credible (Allcott and Gentzkow 2017, 227). These figures may appear less alarming, however the effect and mobilisation of and by disinformation must not be underestimated. In this discussion on hate speech and extremism, we will provide an overview of the background and effects of disinformation, with particular consideration for the dissemination of disinformation in Germany. We look at what measures have already been taken and their effect.

4.1 What is 'fake news'?

A special problem in reporting and researching fake news is the lack of precision and controversial nature of the term itself. It is now also used as a political slogan against media and journalists, for example by Donald Trump, who issues fake news awards to established media representatives (Spiegel Online 2018).

The lack of precision regarding the term is shown by the large number of different classifications, definitions and delimitations of it. Hence, in the context of disinformation, in its typology, the frequently cited non-profit organisation First Draft describes an entire ecosystem, which consists of seven sub-groups of false information and disinformation (First Draft 2017). The recently published study by the Neue Verantwortung [New Responsibility] Foundation introduces the category 'poor journalism' in addition to the term 'fake news' (Sängerlaub et al. 2018, 11 et seq.).

In contrast, this overview is based on a broader definition of fake news as intentionally false information – so-called disinformation. The creation and dissemination of such disinformation is usually based on political or monetary motivation and primarily spread on social media.

4.2 Circulation and relevance of disinformation in Germany

In the course of the 2017 German parliamentary elections, a wave of fake news was expected, analogous to the one that had taken place in the USA in 2016, which could have a serious effect on the election. A hacker attack on the infrastructure of the German parliament fuelled this fear (Sängerlaub et al. 2018, 75). However, many fact-checking institutions concerned with the clarification of disinformation[11] concluded that there was no such wave (ibid.).

So far there are differing findings into the general reach of disinformation in Germany. For example, Buzzfeed News evaluated the most successful articles about Angela Merkel on Facebook by measuring the number of likes, comments and shared links they had attracted. Seven of the top ten articles published between July 2012 and 2017 could be classified as disinformation (Schmehl 2017a).

In a further evaluation, Buzzfeed News found that the most successful disinformation article in German on Facebook in 2017 (Schmehl 2017b) was a report about a study which alleged to demonstrate the ineffectiveness of vaccinations (published by the site anonymousnews.ru). It had approximately 78,500 Facebook interactions in Germany in 2017. Since its publication on Buzzfeed News in December 2017, the circulation of the disinformation article increased to 171,500 interactions.

In comparison: within the same period, the article with the greatest reach in the *Frankfurter Allgemeiner Zeitung* only had approximately 27,500 interactions, one in the *Süddeutsche Zeiting* achieved 60,700 and one in *Die Welt* had 144,100 interactions.[12]

These evaluations show that the circulation of disinformation should not be underestimated. To some extent it can reach a larger audience than established news media on Facebook, though such a distribution is rather an exception and the number of interactions only an estimate of the possible circulation. It is difficult to determine how many people click on a link and internalise and adopt the communicated content.[13] In addition, traditional journalism can use offline circulation outside social media and therefore ensure continual visits to their pages. In contrast, disseminators of disinformation are often one-hit wonders.[14]

A less dramatic picture of the spread of disinformation is drawn by a recently published study by the Neue Verantwortung Foundation (Sängerlaub et al. 2018). The ten case studies investigated in the context of the German parliamentary election showed a rather small overall circulation. An exception was a report of alleged rioting youths with migrant backgrounds at a celebration in the town of Schöndorf (ibid. 2018, 35 et seq.). This report was loosely based on a press report by the Deutsche Presseagentur, which was taken up and exaggerated by protagonists.

Various types of disinformation about the situation in Schöndorf were liked, shared and commented on approximately 500,000 times. In contrast, a correction to the original report did not reach a large audience (ibid. 2018, 39). This demonstrates a great problem: while disinformation can spread quickly and extensively on social media, any correction to it does not reach as many people as the original did, above all not those who originally shared the disinformation content (Kreil 2017). The clarification only had a greater reach than the disinformation in one out of ten cases the Neue Verantwortung Foundation study investigated (Sängerlaub et al. 2018, 79).

[11] *In particular the fact finders of the ARD (http://faktenfinder.tagesschau.de/), the non-profit research centre Correctiv (https://correctiv.org/correctiv/) and the Austrian non-profit association Verein Mimikama (https://www.mimikama.at/) were or are concerned with clarification of fake news in Germany.*

[12] *Evaluation of the circulation was made possible with Buzzsumo (https://app.buzzsumo.com/). The current figures relate to the figures on 30 April 2018.*

[13] *A broad-based study by Columbia University showed that about 60% of the links shared on Twitter were never clicked on by other users. Many users share content merely on the basis of the headlines (Gabielkov et al. 2016).*

[14] *This does not mean that there are not pages which continually disseminate disinformation.*

Social media dynamics and the way the human psyche works may explain why fewer people read clarification articles than those who read the original. A large number of studies have observed that emotional and emotionalising articles are shared particularly often on social media (Berger and Milkman 2012; Ryan 2012; Stieglitz and Dang-Xuan 2013). They found that the emotion of anger has an especially influential role – a particularly activating effect.[15] Users who are stimulated by a particularly appalling headline or message are more likely to click on a link and more willing to share an article (Ryan 2012).

Since March 2016, it has been possible to react to articles on Facebook by using so-called reaction buttons (haha, wow, sad, angry, love). A click on a reaction button has an even higher weighting than a click on 'like' – then users are shown similar articles more frequently in the future (Bell 2017).[16] If someone clicks on an angry reaction button after reading an article about an appalling piece of disinformation this can result in similar articles being displayed to them in the future, because interactions with reports of this type are more frequent.

4.3 Disinformation and its role for hate speech and extremism

It is not surprising that eight of the ten most widely disseminated disinformation articles during the German parliamentary election campaign of 2017 were on refugees and crime (Sängerlaub et al. 2018, 3). Even outside the election campaign, false reports of allegedly criminal refugees were disseminated, especially on Facebook (Schmehl 2017a). The blogs halle-leaks (https://blog.halle-leaks.de/) and anonymousnews (http://www.anonymousnews.ru/) are especially prominent – on these pages there is embedded advertising in the articles for pepper sprays, batons and other weapons which are illegal in Germany (Pittelkow and Riedel 2018).

The comments under the disseminated disinformation match the articles, which are designed to provoke outrage, and are correspondingly negative.

Sängerlaub et al. found there were 70% negative comments under a false report about an alleged quotation by the former chair of the Council of the Protestant Church in Germany, Margot Käßmann (Sängerlaub et al. 2018, 83). The narratives put in place by disinformation are propagated in further discussions on social media.[17]

On social media many quotations from politicians are taken out of context or are often even completely fictitious and in some cases have been in circulation for several years. Until now, legal action against the authors of fake quotations has usually been unsuccessful. A verdict in a precedent case in Berlin is expected and is in the course of litigation (Schwarz 2017). Previous studies on the dissemination of disinformation primarily concentrate on the dissemination of links on social media.

Unfortunately, no adequate studies of the circulation of images and videos in the context of disinformation and its dissemination on social media have been carried out to date; these surveys are necessary to better determine the reach of disinformation (Schwarz 2018, 133).[18]

[15] *With regard to the special role and relevance of the emotion of anger, Brodnig (2016) and Ebner (2018) are especially recommended.*

[16] *However, there is also the limitation that a click on a reaction is only one of thousands of parameters considered when assessing the relevance of articles for users. The criteria Facebook uses are not publicly known – and are subject to continual changes.*

[17] *Monitoring of the narratives which are deliberately put in place by disinformation can be found in Amadeu Antonio Foundation (2017).*

[18] *Studies on the spread of memes and images at University College London are a step in the right direction; see https://www.technologyreview.com/s/611332/this-is-where-internet-memes-come-from/*

4.4 Measures by platform operators against disinformation

Initiatives and measures by platform operators differ greatly depending on the platform, as does their reception by the public. While measures by Facebook against disinformation are frequently discussed in the media, politics and society, the focus of German debate is rarely on Google and YouTube, and even less on Twitter. Among other things, this could be due to the rather small number of users of these platforms in German-speaking countries. However, it is wrong to assume from this that disinformation on these platforms is not important to the public.

In fact, in certain situations, massive amounts of false news are circulated on Twitter. For example, there are attempts to make people affected by terror attacks or natural disasters such as floods or comparable events feel insecure. False news can be aimed at journalists, who adopt the false information as a part of their reporting. After an attack on a concert in Manchester on 22 May 2017 when there were more than 20 fatalities, photos of alleged concert visitors who were reported as missing were posted on Twitter. Some were used by the media including the German newspaper *Bild* and the *British Daily Mail*.

Facebook

After the election of Donald Trump, Facebook was criticised for having been able to influence the result of the election through targeted disinformation. There are several reasons for this. With the exception of YouTube, the number of Facebook users in the USA is considerably higher than the number of users of other social media (Smith and Anderson 2018). In addition, the presidential candidates Clinton and Trump used Facebook as an election platform. Alongside this, non-political protagonists competed for clicks on Facebook with false news, often with financial motivations.

Since March 2017 false information and misleading content are labelled on Facebook in co-operation with various media outleys. What initially started in the USA with partners such as fact checkers from Snopes, PolitiFact and the news agency Associated Press spread to France, Germany and the Netherlands shortly after. Fact checks by Facebook co-operation partners are now displayed in 15 countries, including India, Mexico and the Philippines.

Fact checkers are mainly paid by Facebook itself, which gives rise to concerns among external critics and some fact checkers themselves about possible conflicts of interest (Levin 2017a). According to statements by Facebook, the search for co-operation partners in Germany is difficult, so up to now only Corrective.org checks and labels false news and disputed content in Germany.

Until spring 2018, only linked external content could be labelled. In the meantime, Facebook has announced that in the future photos and videos which users upload directly to the platform will also be checked (Ingram 2018). This is of enormous importance in Germany and many other countries, because much disinformation content is spread through memes or photos and videos, which are taken out of their context and placed in a completely new one. Until now, in Germany Facebook has not extensively informed its users that content can be reported as false.

In addition, the reporting function is listed among other functions, which are not processed by fact checkers but by Facebook moderation teams who ensure that some content is removed or barred. In contrast, false reports which are checked by co-operation partners are labelled and – according to Facebook – displayed less often in users' news feeds. However, at first sight, these differences in the reporting procedure are not clear.

Several co-operation partners have complained about the lack of transparency by Facebook (Levin 2017a). For example, the fact checkers involved did not know how much content was labelled, the consequences of labelling, and which websites were checked most frequently. In fact, so far only one evaluation is known about, which was initially communicated to the fact checkers and then made public: According to Facebook, once it has been checked, content is circulated an average of 80% less frequently than previously (Silverman 2017). However, some reports suggested that labelling had less effect on the distribution of such articles. There is a danger that additional traffic may even be generated by labelling (Levin 2017b). In any case it takes an average of three days before labelling is carried out. Normally, most content on Facebook, whether it contains false information or not, has probably exceeded its peak in circulation after three days.

Facebook has announced a series of further measures to prevent disinformation in the news feed. It has agreed to use technical solutions Facebook to prevent the creation of fake accounts to disseminate political content and will bar existing fake accounts in this way.

Facebook claims to have removed 583 million fake accounts worldwide in the first quarter of 2018[19] and gave its users information on how to identify false reports. Furthermore, in surveys users were requested to vote on the credibility of various media outlets. The results are now used to categorise the content of media outlets in the news feed (Smith and Honan 2018), where local media are preferably displayed.[20] In the USA, links to articles on media pages are supplemented with information from Wikipedia about the publishing medium (Hughes 2018).

In the US election campaign in 2016, many advertisements were placed that were only displayed to particular groups of users. It was suspected and to some extent proven that false news or extreme exaggerations were spread through these 'dark ads', and Facebook has since announced a package of measures to make ads more transparent.[21] Among other things it will provide a database of political advertising (Goldman and Himel 2018).

YouTube and Google

In the past, Google has often been criticised because misleading content and conspiracy theories have been included in the autocomplete function of the search engine. For example, if 'are Jews' was typed in the search bar, one of the supplementing suggestions was 'evil'.

Users were therefore stimulated to ask whether Jews are evil. Entering the terms 'are women' or 'are Muslims' produced similar results. After international criticism, Google removed these search suggestions (Gibbs 2016).

In April 2017, Google introduced the 'fact check tag' (Kosslyn and Yu 2017), giving media the possibility of fact checking possibly misleading content with various metadata, which ensured that these articles were placed very prominently in the search results.

In Germany, only one medium has used this option: Correctiv.org (Niggemeier 2017). This is presumably partly because in order to implement this technology, the content management of the particular news website has to be extended with a plugin and there is little willingness to take this step in many news offices.

In the debate about the influence on elections of false reports, the Google subsidiary YouTube was mentioned considerably less often than Facebook. The platform was recently mainly criticised after a series of conspiracy theories were spread on YouTube after attacks on a concert in Las Vegas and in a high school in Florida, and viewed by a vast audience. Some of these videos were prominently placed in YouTube's automatically generated YouTube Trends. According to critics, YouTube had generated even more publicity for crude conspiracy theories. Furthermore, tests show that users often end up in a kind of vicious circle of conspiracy theories: if a video on a particular topic is clicked, the next video is shown after it has finished. Regardless of whether the last video viewed was from a news medium with a good reputation or a conspiracy theory on the moon landing, content follows which is not based on facts (Chaslot 2018).

After the videos of conspiracy theories about the rampage in Las Vegas had been placed very prominently in YouTube's search results, the company changed its search algorithm to give preference to videos from credible media (Guynn 2017).

At the specialist conference South by Southwest in March 2018, YouTube CEO Susan Wojcicki announced that in future videos that contained conspiracy theories would link to content which provides contextualisation. Among other things, Wikipedia will probably be included (Graham 2018). This announcement was criticised by academics because although Wikipedia articles do not normally contain false information, there are occasional exceptions and academics thought them unsuitable as a source for fact checking (Graham 2018).

[19] *Facebook Community Standards Enforcement Preliminary Report: https://transparency.facebook.com/community-standards-enforcement#fake-accounts*

[20] *Announcement of News Feed update: https://transparency.facebook.com/community-standards-enforcement #fake-accounts*

[21] *Further information on the package of measures: https://newsroom.fb.com/news/2018/04/transparent-ads-and-pages/*

Staff from the Wikipedia Foundation and Wikipedia editors were also sceptical, fearing that many conspiracy theorists would attempt to re-write the relevant Wikipedia articles, which would result in a considerable amount of work for Wikipedia's network of volunteers (Matsakis 2018).

Twitter

Until 2018 Twitter managers had not done enough to limit the dissemination of disinformation, though in February 2018 Twitter considerably restricted the possibility of publishing identical content simultaneously through different user accounts (Roth 2018). The disclosure of personalised advertisements, which were sent to voters in the US election in 2016, caused Twitter staff to announce that they would create a database in which users could view advertisements, including their target groups and financiers. So far this has remained an announcement.

In 2018 Twitter reacted publicly to a series of disinformation items about a shooting incident on the premises of YouTube (Harvey 2018). At the beginning of April 2018, a large number of users had spread photos of alleged offenders. In a statement, Twitter later announced that it had barred several hundred accounts and asked users to remove tweets with misleading content. Previously barred users were prevented from creating new accounts. Twitter uses the function 'moments' in which tweets can be grouped into collections. According to Twitter staff, during the incident a collection with credible facts was used in several languages and countries.

The media have now reported that Twitter is trialling allowing users to report misleading content (Dwoskin 2017), but such a function has not been published.

For a while, verification of individual Twitter profiles was restricted after critics complained that Twitter had verified the accounts of radical right-wing users. If Twitter were to link this function to the dissemination of credible content, this could possibly help limit the spread of disinformation. Research shows that unverified user accounts in the USA contribute to the spread of false news (Amador Díaz López et al. 2017).

However, verified accounts also spread disinformation, which can be circulated very widely, as the Neue Verantwortung Foundation showed in its study of disinformation during the German parliamentary elections in 2017 (Sängerlaub et al. 2018). Among others, the user accounts of the AfD politician Jörg Meuthen and the former Christian Democratic Union member of parliament Erika Steinbach were noticed. Both have a blue tick after their name in Twitter, which shows they are verified profiles.

4.5 Conclusion

Even though the dissemination of disinformation on social media was probably not decisive for the results of the 2017 parliamentary elections in Germany, the previous development and circulation of primarily politically motivated disinformation is at least cause for concern.

Disinformation may contribute to the radicalisation of users by fuelling and strengthening resentment, rewarding outrage in a manner which can hardly be corrected. Minority groups such as refugees are especially frequently the targets of anger and outrage. The problem here is primarily that the algorithms of Facebook and Twitter reinforce the competition for outrage – the interaction with disinformation results in further outrageous reports being displayed. People who are politicised by disinformation are very difficult to re-integrate into the general political debate.

The effects of disinformation on society have not been adequately investigated. It is essential that platform operators allow researchers to access relevant data.

Until now YouTube, Google and Facebook have not taken adequate measures to counter the spread of disinformation, above all in the dissemination of photos and videos. While the circulation of links to content of news sites, blogs and the so-called alternative media can be examined using analysis tools, this is hardly possible for media which are saved directly on the platforms as they are often downloaded and then uploaded again by other users.

Also, more needs to be done to improve media literacy across all age groups. Until now, neither the media nor politicians have found a way to restrict disinformation. One of the greatest challenges is to play any subsequent refutation of disinformation to consumers in a form that does not further increase their belief in the disinformation narrative. In addition, fact checks must be designed so that they appeal to all readers.

Bibliography and references

1. *Allcott, H. and M. Gentzkow (2017) Social Media and Fake News in the 2016 Election, Journal of Economic Perspectives, 31(2), May, 211–36.*

2. *Amadeu Antonio Foundation (2017) Toxic Narratives Amadeu Antonio Foundation, https://www.amadeu-antonio-stiftung.de/w/files/publikationen/monitoring-2017.pdf [03.04.2018].*

3. *A. Díaz López, J., A. Oehmichen and M.I Molina-Solana (2017) Characterizing Political Fake News in Twitter by its Meta-Data, https://arxiv.org/pdf/1712.05999.pdf [01.05.2018].*

4. *Bell, K. (2017) You Might Want to Rethink What You're 'Liking' on Facebook Now, Mashable, https://mashable.com/2017/02/27/facebook-reactions-news-feed/ [17.02.2018].*

5. *Berger, J. and K. L. Milkman (2012) What Makes Online Content Viral?, Journal of Marketing Research, 49(2), 192–205.*

6. *Brodnig, I. (2016) Hass im Netz: was wir gegen Hetze, Mobbing und Lügen tun können [Hate on the Internet: What we can do against Agitation, Bullying and Lies], Brandstätter.*

7. *Chaslot, G. (2018) How Algorithms Can Learn to Discredit the Media, https://medium.com/@guillaumechaslot/how-algorithms-can-learn-to-discredit-the-media-d1360157c4fa [10.05.2018].*

8. *Dwoskin, E. (2017) Twitter Is Looking For Ways to Let Users Flag Fake News, offensive content, https://www.washingtonpost.com/news/the-switch/wp/2017/06/29/twitter-is-looking-for-ways-to-let-users-flag-fake-news/ [10.05.2018].*

9. *Ebner, J. (2018) Wut: was Islamisten und Rechts- extreme mit Uns Machen [Anger: What Islamists and Extremists are Doing to Us], translated by Thomas Bertram, Theiss.*

10. *First Draft (2017) Understanding and Addressing the Disinformation Ecosystem, https://firstdraftnews.org/wp-content/uploads/2018/03/The-Disinformation-Ecosystem-20180207-v2.pdf.*

11. *Gabielkov, M., A. Ramachandran, A. Chaintreau and A. Legout (2016) Social Clicks: What and Who Gets Read on Twitter?, ACM Sigmetrics and IFIP Performance 2016, June, https://hal.inria.fr/hal-01281190/document [03.05.2018].*

12. *Gibbs, S. (2016) Google Alters Search Autocomplete to Remove 'Are Jews Evil' Suggestion, Guardian, 5 December, https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-suggestion [01.05.2018].*

13. *Graham, J. (2018) YouTube Tries to Solve its Conspiracy Problem with Wikipedia: Some Critics Cry Foul, https://eu.usatoday.com/story/tech/talkingtech/2018/03/14/youtube-tries-solve-its-conspiracy-problem-wikipedia-some-critics-cry-foul/425340002/ [01.05.2018].*

14. *Goldman, R. and A. Himel (2018) Making Ads and Pages More Transparent, https://newsroom.fb.com/news/2018/04/transparent-ads-and-pages/ [10.05.2018].*

15. *Guynn, J. (2017) YouTube Alters Algorithm After Searches for Las Vegas Shooting Turn Up Conspiracy Theories, https://www.usatoday.com/story/tech/2017/10/05/youtube-alters-algorithm-after-searches-las-vegas-shooting-turn-up-conspiracy-theories/736548001/ [01.05.2018].*

16. *Harvey, D. (2018) Serving the Public Conversation During Breaking Events, https://blog.twitter.com/official/en_us/topics/company/2018/Serving-the-Public-Conversation-During-Breaking-Events.html [01.05.2018].*

17. Hughes, T. (2018) Helping People Better Assess the Stories They See in News Feed, https://newsroom.fb.com/news/2018/04/news-feed-fyi-more-context/ [01.05.2018].

18. Ingram, D. (2018) Facebook Begins 'Fact-checking' Photos and Videos, https://www.reuters.com/article/us-facebook-fakenews/facebook-begins-fact-checking-photos-and-videos-idUSKBN1H52YX [01.05.2018].

19. Kreil, M. (2017) Social Bots, Fake News und Filterblasen [Social Bots, Fake News and Filter bubbles], 34c3 – Chaos Communication Congress, https://media.ccc.de/v/34c3-9268-social_bots_fake_news_und_filterblasen [03.04.2018].

20. Kosslyn, J. and C. Yu (2017) Fact Check Now Available in Google Search and News Around the World, https://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around- world/ [01.05.2018].

21. Levin, S. (2017a) Facebook Promised to Tackle Fake News, But the Evidence Shows it's not Working, Guardian, 16 May, https://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-working [01.05.2018].

22. Levin, S. (2017b) 'Way Too Little, Way Too Late': Facebook's Factcheckers Say Effort Is Failing, Guardian, 13 November, https://www.theguardian.com/technology/2017/nov/13/way-too-little-way-too-late-facebooks-fact-checkers-say-effort-is-failing [01.05.2018].

23. Matsakis, L. (2018) YouTube will Link Directly to Wikipedia to Fight Conspiracy Theories, https://www.wired.com/story/youtube-will-link-directly-to-wikipedia-to-fight-conspiracies/ [01.05.2018].

24. Niggemeier, S. (2017) Jeder kann für Google Fakten checken – aber kaum einer tut es [Everyone can check facts for Google – but hardly anyone does], https://uebermedien.de/14854/jeder-kann-fuer-google-fakten-checken-kaum-einer-tut-es/ [01.05.2018].

25. Parkinson, H. J. (2016) Click and Elect: How Fake News Helped Donald Trump Win a Real Election, Guardian, http://www.theguardian.com/commentis-free/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies [22.04.2018].

26. Pittelkow, S. and K. Riedel (2018) Rechtsextremist Festgenommen: Profit durch Hass [Right-wing Extremist Arrested: Profit from Hate], http://faktenfinder.tagesschau.de/inland/migrantenschreck-anonymousnews-101.html [08.04.2018].

27. Read, M. (2016) Donald Trump Won Because of Facebook, http://nymag.com/selectall/2016/11/donald-trump-won-because-of-facebook.html [03.04.2018].

28. Roth, Y. (2018) Automation and the Use of Multiple Accounts, https://blog.twitter.com/developer/en_us/topics/tips/2018/automation-and-the-use-of-multiple-accounts.html [01.05.2018].

29. Ryan, T. J. (2012) What Makes Us Click? Demonstrating Incentives for Angry Discourse with Digital-Age Field Experiments, Journal of Politics, 74(4), 1138–52.

30. Sängerlaub, A., M. Meier and W.-D. Rühl (2018) Fakten statt Fakes – Verursacher, Verbreitungswege und Wirkungen von Fake News im Bundestagswahlkampf 2017 [Facts Instead of Fakes – Causers, Dissemination Routes and Effects of Fake News in the 2017 German parliamentary election], Neue Verantwortung Foundation, https://www.stiftung-nv.de/sites/default/files/snv_fakten_statt_fakes.pdf [03.04.2018].

31. Schmehl, K. (2017a) 7 der 10 erfolgreichsten Artikel über Angela Merkel auf Facebook sind Fake News [7 of the 10 most successful articles about Angela Merkel on Facebook are fake news], https://www.buzzfeed.com/karstenschmehl/die-top-fake-news-ueber-angela-merkel [08.04.2018].

32. Schmehl, K. (2017b) Das Hier sind 8 der Erfolgreichsten Falschmeldungen aus 2017 [These are the 8 Most Successful False Reports from 2017], https://www.buzzfeed.com/de/karstenschmehl/8-der-erfolgreichsten-falschnachrichten-2017 [20.04.2018].

33. Schwarz, K. (2017) Renate Künast und Martin Schulz verklagen rechten Blog wegen Fake-Zitaten [Renate Künast and Martin Schult sue blog because of fake quotes], https://motherboard.vice.com/de/article/vb3ajd/renate-kunast-und-martin-schulz-verklagen-echten-blog-wegen-fake-zitaten [10.05.2018].

34. Schwarz, K. (2018) Alles Fake? Zwischen Alarmismus und Verharmlosung [Is Everything Fake? Between Alarmism and Trivialisation], in G. Hooffacker, W. Kenntemich and U. Kulisch (eds) Die neue Öffentlichkeit: Wie Bots, Bürger und Big Data den Journalismus Verändern [The New Public Sphere: How Bots, Citizens and Big Data are Changing Journalism], Springer VS, 125–33.

35. Silverman, C. (2016) This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook, BuzzFeed, https://www.buzzfeed.com/craigsil-verman/viral-fake-election-news-outperformed-real-news-on-facebook [03.05.2018].

36. Silverman, C. (2017) Facebook Says its Fact Checking Program Helps Reduce the Spread of a Fake Story by 80%, BuzzFeed, https://www.buzzfeed.com/craigsilverman/facebook-just-shared-the-first-data-about-how-effective-its [21.06.2018].

37. Smith, A. and M. Anderson (2018) Social Media Use in 2018, http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/ [01.05.2018].

38. Smith, B. and M. Honan (2018) Facebook Has Begun to Rank News Organizations by Trust, Zuckerberg Says, BuzzFeed, https://www.buzzfeed.com/bensmith/facebook-has-begun-to-rank-news-organizations-by-trust [01.05.2018].

39. Spiegel Online (2018) Kritik an US-Medien: Trump vergibt 'Fake News Awards' [Criticism of US Media: Trump Issues 'Fake News Awards'], http://www.spiegel.de/politik/ausland/donald-trump-vergibt-fake-news-awards-a-1188476.html [03.05.2018].

40. Stieglitz, S. and L. Dang-Xuan (2013) Emotions and Information Diffusion in Social Media – Sentiment of Microblogs and Sharing Behavior, Journal of Management Information Systems, 29(4), 1 April 217–48.

41. Wardle, C. (2017) Fake News: It's Complicated, https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79 [10.02.2018].

42. Weedon, J., W. Nuland and A. Stamos (2017) Information Operations and Facebook, https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf [01.05.2018].

5. Civil society: defending the global village: strategies against the cultural backlash on social media

By Dr. Matthias Quent

Abstract
Combatting hate speech and extremism on the internet cannot just be left to the government and social media platforms. Above all, with the prevalence of content which is problematic but neither illegal nor breaches the community standards of large platforms, decisive opposition from civil society is required. Dr. Matthias Quent views hate on the internet as part of an extensive cultural backlash against progressive achievements of modern democratic societies, which is also taking place offline. In order to restrict the effect of hate speech on social media, Quent calls for education, solidarity and a strengthening of the narratives of marginalised groups to be given priority over repressive measures.

Insults, humiliations, prejudice and organised intimidation campaigns, above all by extreme right-wing protagonists, are a threat to the self-declared claim by Facebook to bring "the world closer together" (Facebook 2018). In public discussion, justified controversies about the power of algorithms and their risks of discrimination, disinformation, data protection, tax questions and the question of what is hate on the internet overshadow the historical and sociological classification of the new social media platforms. My thesis is that as part of an extensive cultural backlash movement, messages of hate on the internet contradict the claim that people throughout the world are coming closer together. This article combines observations of articulation, formation and radicalisation of prejudice against marginalised groups offline and online. The focus is on how academia, civil society, social media platforms and state protagonists can collaborate to transfer suitable practical prevention concepts to social media for counteracting hate and strengthening democratic culture.

A differentiation is made between object-oriented approaches which relate to victims and subject-oriented approaches relating to protagonists. The first focus on the consequences of hate messages for its victims and democratic culture. With the protagonist or subject-oriented approach, the focus is on dealing with producers and organisers of hate messages.

5.1 On the road to a digital global society

In the 1990s many people emphasised the potential of the internet to establish a 'colour-blind society', one without any form of racism being evident in communication. This could result in an "egalitarian, electronic village... in which there would be no race, gender or weakness" (Tynes et al. 2016, 2). The fact that such a process would not take place without opposition from racist and sexist users was foreseeable. The importance of social media for democratic resistance movements, not only in democratic states but global emancipation movements such as #metoo show that the progressive potential of the internet is more than just an advertising promise by platform operators. The age of purely analogue politics is over. Like it or not, democracy researchers, politicians and civil society must concern themselves with conditions which have been changed by social media – and forge new alliances.

In contrast to pessimism about progress, from a historical –sociological perspective it must be emphasised that especially in and with social media, for the first time in human history the ideal of a cosmopolitan global society is becoming a virtual reality. Large parts of humanity have the opportunity of participating, networking and communicating with relatively few restrictions and free of charge, with voluntary disclosure of their personal data. By necessity, divergent worlds collide: in comment columns and groups, feminists meet masculinists and liberal city dwellers meet Nazis from rural areas.

Prejudiced people may be confronted with modern cosmopolitan society and the near equality of users from all possible religious, ethnic, cultural, social and ideological backgrounds under the conditions of globalised capitalism. Technical possibilities are developing faster than human concepts of society: even though the barriers between people are reduced by social media, latent attitudes of group-related enmity do not disappear with two clicks. It is not surprising that these derogations are also articulated under conditions of hate speech on the internet (anonymity, invisibility, community, instantaneousness and particular vulnerability; Brown 2017).[22]

Changes toward a virtual cosmopolitanism whose context in a digital age is primarily defined by international, commercially organised social media force national politicians to expand transnational co-operation. At the same time, parts of society and politics wish for renationalisation, as is shown among other things by the Brexit vote and the election successes of right-wing parties.

People who were socialised with nationalist, gender-related or racially justified mindsets of privilege find themselves in contradictory situations on global social media. They stand between a national–patriarchal past, which still dominates many offline areas, and a cosmopolitan digital future.

Positioning online and offline remains divided. For example, in comment columns on the internet, men are not more assertive than women because of their loud voices, aggressive body language or to some extent concealed techniques of 'male dominance' (Bourdieu 2012). Especially on the internet, contempt and humiliation of (emancipated) women is all the more blatant. However, national identities or membership of minorities can largely be made invisible on the internet, so cultural hierarchies are deconstructed in the discourses on social media.

Social media therefore offer the opportunity of coming closer to the promise of the Enlightenment: not the origin, gender or ancestry of the author should be decisive for the award of 'likes', but rather quality, humour or persuasiveness.

Prejudiced hate messages strongly emphasise the desire for delimitation, exclusion and hierarchy, up to the expulsion of users from the virtual community by means of enormous intimidation. Derogation on the internet, especially when directed against women and social minorities, results from the fear of true equality, which is perceived as a threat.

5.2 The digital anti-modernity

Hate on the internet is not a new phenomenon, but rather the digitalisation of cultural backlash politics against processes of cultural change and progress. The political scientist Inglehart describes the growing importance of post-material values such as cosmopolitanism and multiculturalism in Western societies as a "silent revolution".

Inglehart and Norris (2016) emphasise that links between the populations of various nations have greatly increased since the Second World War and the previously dominant belief in a homogeneous national state has been replaced by a cosmopolitan mindset. Political scientists explain that the growth of right-wing populist forces is due to the fact that previously socially privileged groups see themselves threatened with a loss of cultural status. The backlash is the reaction of groups that were formerly culturally dominant in Western Europe, whose members reject progressive values and react angrily to the perceived undermining of their privileges and status as well as to the change in cultural values (Inglehart and Norris 2016, 3).

Global developments, for example increased immigration into Europe, exacerbate the disappointment of expectations of the national state and mobilisation of right-wing defences increases. However, why is there so much hate online?

[22] *In contrast, Rost et al. (2016) argue that the great majority of hate speech is posted under real names.*

On the internet, what is otherwise primarily experienced by discriminated groups of society becomes publicly visible. The US sociologist Kimmel (2015) argues that especially because social progress in equality has increasingly come to dominate everyday life in Western societies over the past few decades, the "angry white men" have fled into virtual enclaves in order to express derogatory ideas, which have been justifiably tabooed offline in the course of equalisation over the past few decades. It is natural that hated political correctness, which reduces their historical privileges, is opposed with a defiant incorrectness by protagonists of the cultural backlash. The boundaries between organised campaigns of right-wing extremist activists, who use social media as instruments to disseminate their narratives and shift the virtual hegemony, become blurred.

Although they are reactionary and anti-modern, the opponents of the ongoing cosmopolitan silent revolution, who are beginning to form a backlash movement, are extremely flexible in their propaganda methods and adapting to changing technical conditions and possibilities. The extreme right uses the technical progress of (post) modern society in organised online campaigns. Racists, right-wing culture pessimists and extremists in Europe, who otherwise see indications of the end of Western society in social transformation processes, transform their politics with great efficiency.

Modern appearance and professional use of new technologies, largely without ideological taboos, has tradition among the extreme right: even the National Socialists distributed large numbers of the then technically innovative Volksempfänger radios in order to reach every household, and the predilection of the Nazis for technical 'wonder weapons' in the fight for the 'final victory' is well known.

In his excellent analysis of the 20th century the historian Hobsbawn wrote:

> *"As fascism in principle rejected the legacy of the Enlightenment of the 18th century and the French Revolution it also rejected the ideology of modernity and progress. However, it had no difficulty in linking a lunatic assortment of ideologies with regard to practical questions with technological modernity unless ideological grounds opposed its scientific research... And he provides evidence that without the slightest difficulty, people can combine completely crazy ideologies about everything in the world with excellent mastery of the high technology of their age. The late 20th century with its fundamentalist sects, which fight with the weapons of television and computer-controlled benefit events have made us even more familiar with this phenomenon" (Hobsbawm 2009, 155 et seq.).*

Social media platforms create a vast space for social and political conflicts. It is therefore not surprising that today's right-wing radicals and anti-modern groups, for example the 'identitarians', use modern social media in a highly professional manner.[23] What conclusions must be drawn from this when looking at measures against hate?

5.3 Solidarity against hate

In the context of the struggles for civil and fundamental rights, the term 'hate speech' does not designate random harsh statements, radical criticism or insults per se, but rather statements which reproduce prejudice and discriminate against marginalised groups. Unlike (cyber) bullying, hate speech is always group-related: in addition to the injured persons, the consequences of hate speech affect entire social groups (e.g. Jews, migrants, people with handicaps and similar).

[23] *At the beginning of June, many pages from the identitarian movement and their more prominent activists were barred on Facebook and Instagram.*

It is not the emotion hate, but rather the prejudiced, verbal derogation of particular groups which is the defining characteristic of hate speech (Geschke 2017).

In a study by the German internet association eco (2016) more than one-third of those questioned reported that they had experienced racist hate messages on the internet. Verbal derogation of weaker groups not only opposes social cohesion and democratic internet culture, but rather is directly harmful to those affected. Hate speech can cause psychological harm to those who are confronted with it and increase social divisions or even give rise to violence in particular cases (Costello and Hawdon 2018, 55). Online harassment can cause changes in the online behaviour of those affected and impair their well-being. The Australian study *Cyber Racism and Community Resilience* concluded that the victims of cyber-racism react to hate messages with anger and frustration, and one in ten of the questioned victims had physical reactions such as headaches, stomach problems, cramps, palpitations or sweating (Jakubowicz et al. 2017, 76). Tynes et al. (2016) shows a there is a correlation between online discrimination and poor mental health as well as externalising behaviour. These findings confirm the robust findings of research: experiences of discrimination, online and offline, have a negative effect on the life of the victims (Dieckmann et al. 2017).

Action against antisemitism, racism, discrimination and ethnocentricity must be carried out online and offline. The mechanisms of scapegoating and social marginalisation of vulnerable groups are similar. The response to this collective victimisation by harmful speech should be to strengthen marginalised groups and their perspectives, including by widening awareness through monitoring and informing people about hate speech.

On the basis of extensive empirical analysis, Jakubowicz and colleagues (2017) state that one of the most important measures that can be taken against racism on the internet is the creation and support of 'online communities of resistance and solidarity' by group-related and general protagonists from civil society, state institutions, grass roots activists and scientists.

These communities should respond reactively to hate messages by:

1. Naming racism online and in everyday life,
2. Working together online to have racist content removed, and
3. Contradicting racist narratives with counterspeech campaigns.

These communities should also develop campaigns that:

– Emphasise positive values such as diversity,

– Communicate knowledge about the culture and traditions of vulnerable groups,

– Provide narratives emphasising the damaging nature of racism for individuals and society,

– Help people to become aware of what other groups, for whatever reason, consider to be racist,

– Counteract historical and other narratives which reflect the prejudice of dominant cultural groups to the disadvantage of minorities,

– Provide narratives which normalise positive inter-group relationships (ibid. 224 et seq.).

Furthermore, information and education campaigns, on social media and in schools, universities and businesses are needed to reduce the consequences of online and offline discrimination.

It is essential to communicate to (potential) perpetrators of hate messages that it is not the victims who are responsible or 'at fault', but rather the aggressors and perpetrators. Ideally there should be independent advice centres, which provide the victims of discrimination with psychological, social and – in case of doubt – legal support. Especially on the internet, they could provide this low threshold and if necessary anonymous help and advice; by lobbying they could help increase awareness and visibility of discrimination.

Law enforcement authorities must not shirk their responsibility of prosecuting people who post illegal content, insults or threats. They can learn from civil society in the necessary digitalisation processes. In their own interest, social media should effectively protect their users against hate and strengthen democratic internet communities

5.4 Restriction of the effect of hate groups

Organised campaigns by hate groups are aided by discriminating content, which is continually posted by individual users. The sociodemographic backgrounds of the authors of hate messages have only been rudimentarily researched. Initial analyses from Germany suggest that a comparatively small number of accounts are responsible for a relatively large number of hate comments and that many postings are part of organised right-wing campaigns (Tagesschau Faktenfinder 2018). This confirms international findings, which observe a dominance of right-wing extremist internet content.

Typically far-right hate groups on the internet disseminate convictions based on white male superiority and racial homogeneity. Groups considered to be a threat to this ideology – such as ethnic minorities, immigrants, Muslims, Arabs, Jews, feminists, homosexuals, the government and political liberals – are the most frequent targets of right-wing hate (Costello and Hawdon 2018, 56). The same study of online hate found that men disseminate online hate material 1.76 times more frequently than women (ibid., 57). This supports the hypothesis that an antifeminist cyber-backlash is taking place on the internet. In addition, the research found the probability of producing hate is eight times greater for internet users who report they have been the targets of 'online hate' than for those without perceived derogatory experiences. Accordingly, hate reproduces itself in hate-filled online environments (ibid., 59).

Further investigation is necessary in order to analyse the sociological backgrounds of hate speech and to develop countermeasures for specific target groups. While there are many offline projects and methods to prevent the harmful influence of right-wing extremists, primarily on young people, for example in schools, meeting places and by media, toxic presentations on social media (especially significant in the form of propaganda videos and music on YouTube) are often only a few clicks away. The removal of hateful content, which is often seen as the strongest reaction to them, is neither a panacea nor without alternatives, however. Ultimately there is a risk that the achievements of the Enlightenment and freedom of opinion will be sacrificed to a repressive logic, which can be used against political protagonists and opinions of all kinds. How can the methods of democratic discourse and political education be combined with the new challenges of social media in order to weaken the effects of anti-democratic internet phenomenon?

Civil society protagonists are a source of important information about hate groups and their backgrounds. Online activists operate 'watch accounts' and research associations. Platform operators have been able to use and support this know-how, for example by emphasising the content of research associations next to the content of hate groups. Contextual information relating to the self-presentation of right-wing extremist groups could be displayed on social media, which could link to independent information articles about the agenda of a particular group. There the frequently coded self-portrayals could be categorised in co-operation with experts, and information could be provided about the mechanisms and aims of the group in an understandable manner appropriate for the medium.

In a similar manner to those who spread disinformation, in co-operation with experts, social media operators could draw attention to right-wing extremist groups and display links to critical reports on their pages. In this way, the filter bubbles of right-wing extremist groups could be broken open. Users could then still decide whether or not they 'like' any particular page, but would be encouraged to consider arguments related to the disinformation in their decisions, in the same way as they decide whether or not to heed health warning notices on cigarettes.

Such methods should be carefully prepared and tested in small case studies in order to ensure they do not have adverse consequences, such as a reinforcing the cohesion of the right-wing extremist scene or public perception of disinformation on the platforms.

5.5 Information rather than repression

Interventions against the authors of (organised) hate campaigns are important and provide solidarity with those who are marginalised by hate speech and the revelation of discrimination; they can promote alternative narratives and counterspeech.

Information about the (political) backgrounds, mechanisms and strategies of online hate is especially important in creating social resilience and restricting the effect of hate group propaganda. However, the aim should not be to refute every (interchangeable) allegation introduced by trolls and haters as this would only bolster their position in the discourse.

Engaging with propaganda and derogatory speech can be misunderstood as acceptance of their discourse. Repression, for example by removal of content, only acts against the symptoms, according to the principle 'out of sight, out of mind'. It is therefore especially important to provide information about the general patterns, protagonists and aims of hate. This cannot prevent hate in the short term, but it can restrict its damaging effects and circulation.

References

1. *Bourdieu, P. (2012) Die männliche Herrschaft [Male Domination], Suhrkamp.*

2. *Brown, A. (2017) What Is So Special About Online (As Compared to Offline) Hate Speech?, Ethnicities, 13.*

3. *Costello, M. and J. Hawdon (2018) Who Are the Online Extremists Among Us? Sociodemographic Characteristics, Social Networking, and Online Experiences of Those Who Produce Online Hate Materials, Violence and Gender, 5(1), 55–60.*

4. *Dieckmann, J., D. Geschke and I. Braune (2017) Diskriminierung und ihre Auswirkungen auf die Gesellschaft [Discrimination and its Effects on Society], https://www.idz-jena.de/wsddet/diskriminierung-und-ihre-auswirkungen-fuer-betroffene-und-die-gesellschaft/ [30.04.2018].*

5. *eco (2016) Survey: Jeder Dritte ist schon rassistischen Hassbotschaften im Internet begegnet [One in Three People Have Encountered Racist Hate Messages on the Internet], https://www.eco.de/news/eco-umfrage-jeder-dritte-ist-schon-rassistischen-hassbotschaften-im-internet-begegnet-2/ [19.04.2018].*

6. *Facebook (2018) Terms of Use Online, https://www.facebook.com/legal/terms/update?ref=old_policy [03.05.2018].*

7. *Geschke, D. (2017) Alle reden von Hass. Was steckt dahinter? Eine Einführung [Everyone is Talking About Hate. What Is Behind This? An Introduction], http://www.idz-jena.de/wsddet/alle-reden-von-hass-was-steckt-dahinter-eine-einfuehrung/ [19.04.2018].*

8. *Hobsbawm, E. J. (2009) Das Zeitalter der Extreme. Weltgeschichte des 20. Jahrhunderts [The Age of Extremes: A World History of the 20th century], Dt. Taschenbuch-Verlag.*

9. *Inglehart, R. F. and P. Norris (2016) Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash, HKS Faculty Research Working Paper, 26.*

10. *Jakubowicz, A., K. Dunn, G. Mason, Y. Paradies, A.-M., Bliuc and N. Bahfen (2017) Cyber Racism and Community Resilience. Strategies for Combating Online Race Hate, Palgrave Hate Studies, Springer-Verlag.*

11. *Kimmel, Michael (2015) Angry White Man: The USA and its Angry Men, Orell Füssli.*

12. *Rost, K., L. Stahel and B. S. Frey (2016) Digital Social Norm Enforcement: Online Firestorms in Social Media, PLoS One, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155923.*

13. *Tagesschau Faktenfinder (2018) Lautstarke Minderheit [A Noisy Minority], http://faktenfinder.tagesschau.de/inland/hasskommentare-analyse-101.html [03.05.2018].*

14. *Tynes, B. M., C. A. Rose, S. Hiss, A. J. Umaña-Taylor, K. Mitchell and D. Williams (2016) Virtual Environments, Online Racial Discrimination, and Adjustment among a Diverse, School-Based Sample of Adolescents, International Journal of Gaming and Computer-mediated Simulations, 6(3), 1–16.*

6. Case studies: which types of campaign against hate and extremism on the internet work, which do not, and why?

By Sina Laubenstein and Alexander Urban

Abstract
In order to counteract hate speech and extremist messages on social media, over the past few year's civil society organisations have increasingly used counterspeech campaigns. But do such online initiatives have the desired effect? How can the success of a good counterspeech campaign be measured? Sina Laubenstein and Alexander Urban believe that it is especially important that campaigns do not act blindly, but rather have a clear strategy, specify a definite target group for their message, and build up a strong presence on social media. As examples of such successful campaigns, the authors present the German local group of the No Hate Speech Movement and the Facebook group #ichbinhier.

The internet and social media influence forms of communication and interpersonal relationships: although the internet provides new forms and forums for exchange and participation, it also presents society with new challenges. Hate and incitement against groups often spread unchecked and without comment on social media. Civil society is only slowly becoming engaged in the online environment, even though the number of counterspeech campaigns in Germany has increased significantly over the past two years. However, the increasing number of counterspeech initiatives and campaigns has not resulted in a reduction in extremist propaganda on the internet. Often, they have had the opposite effect: from research and the evaluation of a few case studies it is apparent that counterspeech campaigns can often have negative consequences (Hemmingsen and Castro, 3). Ultimately, discriminatory narratives are at least reproduced by referring to them, and there is a danger of legitimising them (Sponholz 2016, 519).

Counterspeech campaigns are nonetheless recommended as a response to hate speech and extremist propaganda on the internet – and for good reason. Without counterspeech campaigns, the field would be abandoned without contest to those advocating hate speech and allows haters to dominate and poison discourse in the online environment (Braddock and Horgan 2016, 398).

However, counterspeech campaigns are not all the same: Although it is important and right that there are new initiatives and that hate is opposed, not all campaigns against hate and enmity on the internet work. But when can a campaign be considered to be effective and when has it failed?

6.1 What are the factors for a successful counterspeech campaign?

One problem when evaluating online counterspeech campaigns is their differing aims, relating to target group, approach, messenger, medium and impact (Tuck and Silverman 2016, 8 et seq.). Although many initiatives are concerned with hate and enmity in the online environment, only a few organisations appear to have created an online strategy or addressed it significantly on social media.

Usually, counterspeech campaigns are evaluated according to three criteria: awareness, engagement and impact (Silverman et al. 2016, 23 et seq.); they include variables such as the overall circulation which the campaign achieves on the internet and the number of impressions[24] and video views.

Only circulation in the online environment is considered here, not reporting in the media or the offline presence of a counterspeech campaign, although both are essential indications of a successful counterspeech initiative which attracts attention beyond the boundaries of the virtual world.

[24] *Impressions are the number of times a tweet or a post appears in the timeline or search results of an account.*

When evaluating engagement the communication of counterspeech campaigns with users – including comments and private messages, and 'likes' and shares of campaign content with the users' networks – is measured. Again, only the online environment is considered owing to methodological difficulties, but it can be assumed that users discuss the content and activities of counterspeech campaigns outside their virtual networks. The extent to which and how users are influenced by the counterspeech campaign should be measured when evaluating engagement (Initiative für Zivilcourage Online 2016). However, it is difficult to measure sustainable effects such as a positive change in attitude or behaviour of followers or haters, and this is only considered to a limited extent in the following analysis of functioning counterspeech campaigns.

6.2 Practical examples: functioning counterspeech campaigns from Germany

The youth movement No Hate Speech and the grassroots Facebook initiative #ichbinhier are two well-known counterspeech campaigns of the past few years. They have different approaches, but both focus primarily on the internet. The No Hate Speech Movement and the recently founded Facebook group #ichbinhier are dedicated to improving the civility of discourse on Facebook. #ichbinhier has grown rapidly within a very short time, demonstrating that people are not prepared to accept brutal comments on social media. These two counterspeech campaigns are discussed below using the aforementioned evaluation criteria in order to show which types of campaign function in Germany, and why. This can be helpful when creating counterspeech campaigns, although naturally there is no guarantee that counter-narratives will always be successful if they have a similar structure.

6.3 The No Hate Speech Movement

In 2012 the Council of Europe initiated the international youth movement No Hate Speech after various youth organisations had pointed out the increasing danger of hate on the internet and requested an initiative at European level. The aim of the movement was in particular to mobilise young people to stand up for fundamental human and democratic rights in the online environment.

The key features of the No Hate Speech Movement in Germany

From the start, the No Hate Speech Movement was in a privileged position. As one of the first initiatives in Germany exclusively dedicated to the topic of hate on the internet, the No Hate Speech Movement rapidly became one of the central partners for civil society in this field.

This can be attributed to the empowering approach of the counterspeech campaign: the focus of the movement is not the response to haters, but rather the perspectives of people affected by hate online. In addition, in Germany the No Hate Speech Movement networked various protagonists for pragmatic reasons: a broad alliance comprising representatives from politics and commerce, civil society organisations and activists achieves far more attention than a single initiative. Furthermore, the campaign benefits from its humorous approach, which to date had not existed in Germany or other national no hate speech movements.

An exemplary action characteristic of the success of the No Hate Speech Movement in Germany was the campaign for an action day for the victims of hate crime on 22 July 2017. To support the victims of hate speech and create public awareness of the topic and how to deal with it, the No Hate Speech Movement in Germany published three videos on this action day, which showed protagonists who had been victims of misanthropic comments on the internet, and how they had dealt with it.

Simply with these videos, the action day directly addressed more than 255,000 people via Facebook. The campaigners also published a No Hate Speech profile picture frame, which was also very popular internationally and strengthened the role of the German counterspeech initiative beyond its national borders. The articles on 22 July 2017 were viewed by some 300,000 people, following co-operation with protagonists and campaigns with the largest reach, including the grassroots movement #ichbinhier, the Council of Europe and the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth.

The success of this single campaign as the No Hate Speech Movement resulted from not just its well-thought-out social media strategy but also the preliminary work undertaken. The project team sent a press release to the relevant media and political and civil society protagonists in advance, benefitting from the broad network of the sponsoring association, which is networked with many media professionals.

Evaluation of the counterspeech No Hate Speech Movement

From the outset, the No Hate Speech Movement in Germany was able to generate a great deal of attention, partly by using the tool Thunderclap. Various protagonists used this platform to spread details of the launch of the initiative within their respective audiences. On the day of the campaign launch, 22 July 2016, the movement had reached approximately 645,000 people directly via Facebook, Twitter and YouTube. The number reached indirectly was probably much larger. The hashtag of the campaign dominated the trends on Twitter in Germany.

In retrospect, the movement benefitted greatly from its highly successful start in Germany because public influencers shared the content of the campaign widely, including with members of the Federal Government and the (then) Foreign Minister Frank-Walter Steinmeier.

Especially in the field of social media, the counter-speech initiative has learned a great deal from its own mistakes and mistakes of other protagonists and campaigns. It now focuses more strongly on mobilising civil society commitment and increasingly implements the international human rights approach: not merely to oppose something but instead show what it is committed to and stands for. This intention is made more difficult by the name chosen for the campaign by the Council of Europe and the international youth movement, which still leads to misunderstandings including aggressive attacks on social media. To address this, campaigners created a social media strategy and adapted language and content to particular networks and different audiences.

The initiative has generated various campaigns for different target groups. While initially (and still) a relatively wide audience was (and is) reached and addressed, individual campaigns targeting specific online and offline audiences were also created. This appears to differentiate the No Hate Speech Movement from other counterspeech campaigns, nationally and internationally: the No Hate Speech Movement has an online and offline strategy. The implementation of the campaign is not static, but adapts to new situations and developments.

The final evaluation of the Council of Europe for all No Hate Speech campaigns throughout Europe confirms this assumption: although there are online movements in other countries, the other national No Hate Speech initiatives have been very static and gained little momentum among young people. Furthermore, only the German initiative has produced and shared its own content, which was rewarded with 'likes' and comments by users.

The study *Videos Against Extremism* discusses how social media adapt the message of the No Hate Speech Movement and address particular target groups (Frischlich et al. 2017). The authors conclude among other things that clear and understandable messages have a more sustainable effect than those that are not clear. This can be confirmed by considering the social media strategy of the No Hate Speech Movement.

However, at present the counterspeech initiative is reaching far fewer people than in the past. This could be because the algorithm on Facebook has changed, which can only be addressed with massive advertising campaigns. It presents a challenge to civil society initiatives, because promises of sponsoring often depend on the organisation in question being able to prove it has a certain reach, even if this reach does not indicate much about the effectiveness of these campaigns.

Lessons: mobilisation of civil society

A large part of the success of the counterspeech initiative No Hate Speech Movement arose from offline actions, although activities on social media were also very influential. The movement's reach would certainly have been smaller if accompanying measures had not been used, including protaganists communicating with external professional groups and participating in many networking and training events.

The No Hate Speech Movement has built up a broad network of actors who draw attention to the work of the campaign in their fields of work, including the German Football Association, the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth, the Federal Ministry for Foreign Affairs, the Amadeu Antonio Foundation and activists such as Anne Wizorek. The intent and purpose of this network is to enable interaction and communication between many protagonists, and to transmit messages to the public through this broad alliance.

From the outset the initiative has collaborated with prominent supporters and influencers: the comedy video series *Bundestrollamt für gegen digitalen Hass* [Federal Troll Office Against Digital Hate], with Idil Baydar as Jilet Asyse, Nemi El-Hassan from the 'Datteltäter' and Raul Krauthausen, acted as messengers and used their popularity to draw attention to the content of the movement. The first series of videos was published before the official launch of the counterspeech initiative and generated a large reach in advance (addressing more than 400,000 people). Although Frischlich et al. (2017) concluded that humorous and satirical videos tend to be counterproductive, the counterspeech initiative was able to avert this effect by allowing hate speech victims to speak themselves, and using them as credible messengers. Frischlich et al. (2017) noted this was an important way to target particular groups successfully. During the course of the campaign, short videos were repeatedly shown of victims of hate speech and how they coped; they were empowered as a result.

Since the beginning of 2016, the No Hate Speech Movement has collaborated on several topics with various people from different fields, including civil society organisations such as Jugend Rettet and Hooligans gegen Satzbau, influencers such as Oguz Yilmaz (formerly Y-Titty) and Tarik Tesfu, the actress Emilia Schüle, journalists and authors such as Ingrid Brodnig, Kübra Gümüsay and Carline Mohr.

The social media work of the No Hate Speech Movement was strategically underpinned by its offline activities, including large events such as the gaming exhibition Gamescom and the Federal Government open day. Civil society was thus able to influence and be involved in the work of the movement.

But that is not all: the creative results of workshops and the photos submitted of events, independent of the campaign team, are part of the social media work of the initiative and have clearly shown that civil society is willing to stand up for social interaction – but its mouthpiece did not exist.

Before its launch in Germany, the No Hate Speech Movement actively sought supporters to address and mobilise civil society. With its activities in comment columns, the Facebook group #ichbinhier has also actively involved civil society and therefore justifiably gained the attention and support of public figures such as Dunja Hayali and more recently the band DONOTS. Even though their approaches are different, both campaigns use traditional media and so-called influencers to underline their commitment on social media. The Facebook group #ichbinhier is discussed below.

6.4 The Facebook group #ichbinhier

Origins of the group

The Facebook group #ichbinhier was founded by Hannes Ley in December 2016 with the aim of improving the civility of discourse on Facebook. Like many other Facebook users, in 2016 Ley observed an increasingly toxic climate in (online) society, especially in the comment sections on social media. Comments on Facebook could show disparagement, exclusion, hostility and even support for violence against entire population groups. Anyone who attempted to counter such comments dispassionately was quickly insulted, intimidated or even threatened.

Hannes Ley no longer wanted to watch this happening without doing something, and he founded the German action group #ichbinhier on Facebook, modelled after the Swedish group #jagärhär. Membership of #ichbinhier grew to several thousand members within a few weeks. The primary aim of the members was to raise the level of discussion in the comment sections online to a minimum standard and therefore create the conditions for a decent debating culture.

Such an initiative came at precisely the right time, not only for people who had individually opposed inflammatory comments on Facebook, but also for those who had withdrawn from the comment columns in shock, but were increasingly worried about social peace and democracy. In September 2018 more than 43,700 people had joined the #ichbinhier group.

Fundamentals and principles

As the need to improve the quality of discussions online was considered to be a general, over-riding matter, concerning the whole of society, it quickly became clear that actions to further this aim must be neutral and have nothing to do with judging anyone's politics. The group should serve as a platform for objective discussion and through comments by its members contribute to communicating balanced opinions to silent readers, instead of holding heated and highly aggressive debates. The founding of the group required a framework to be established within which the group would act.

In order to obtain a unified, comprehensive (and literally) average impression of the media presence of inflammatory comments (hate speech), the moderating team decided to focus the group's activity on:

- Facebook pages of media of any political leaning with more than 100,000 followers,
- not on private pages,
- not on Facebook groups and
- not on political party pages.

In exceptional cases the fields of activity were extended to the Facebook pages of people from public life, NGOs, foundations and other initiatives, if organised and externally controlled campaigns (shitstorms) were targeting them, for example Dunja Hayali, the Dresden Philharmonic Orchestra, Margot Käßmann and the GoVolunteer initiative.

Aims and functions

The large media pages with more than 100,000 followers are scrutinised to find out whether their articles contain unobjective, generalised, inflammatory and/or derogatory comments. If such a comment column is found, an 'action' is started within the Facebook group. Any member can then oppose the hate and agitation in any way, entirely free of restrictions. The only rule is that objectivity must be maintained.

Use of the hashtag #ichbinhier is discretionary for members, but beneficial as way to create transparency and identify group members. In contrast with other organisations such as Reconquista Germanica or corresponding Facebook groups, #ichbinhier does not operate in secret. This has resulted in attempts to hijack the hashtag or use it negatively, though so far without serious consequences. This is probably because of the transparent, neutral and authentic operation of #ichbinhier and its good reputation; members very quickly unmask defamatory or cynical comments.

The aim of taking an 'action' against an inflammatory comment has always been to communicate balanced and objective opinions through other users' comments and to support objective comments by group members (with 'likes'), so these comments move upwards in the comment columns and become more prominent than hate-filled or unobjective comments.

The definition of incitement to hatred, hate speech and freedom of opinion

In order to differentiate hate speech from justified anger, criticism or black humour, the moderation team initially followed the German dictionary Duden's definition of 'incitement to hatred': the "totality of unobjective, spiteful, libellous, disparaging statements or actions... which generate feelings of hate, hostile moods and emotions against someone or something". Moderators also consulted a publication by the Committee of Ministers of the Council of Europe (Council of Europe 2007), which defined hate speech as comments which have the aim of degrading and humiliating individual people, groups, communities or ethnic groups or of branding and excluding persons as 'different' on the basis of their ancestry, origin, colour, gender, sexual orientation, religion and physical handicaps

These are some other characteristics of incitement to hatred:

– Deliberately spreading uninformed or false statements.

– Using stereotypes and asserting prejudices.

– Insulting, disparaging and dehumanising victims for their membership of a group (for example being asylum seekers').

– Calling for physical violence against those abused.

Example of one month's statistics

From the very start, two criteria were used to assess the success of the group to some extent: the number of 'likes' and the number of silent followers who took part. Part of the strategy of #ichbinhier is for members to mutually 'like' the comments made by others in the group. It is therefore not surprising that approximately 86% of the 'likes' in November 2017 of a total of 90,985 'likes' received (of which 2,028 were top level comments – those directly visible to users) came from members of the group.

The remaining 14% of 'likes' come from silent followers who for the sake of simplicity are defined as non-members. Initially, that does not sound a lot, but it must be considered that each member distributes more 'likes' per head, which relativises this ratio. It is interesting that many non-members interact with top level comments from #ichbinhier members: 71.4% of all accounts which the 2,028 top level comments from #ichbinhier liked were not held by members of the group. This ratio fluctuates, depending on which medium is examined. It is noticeable that when media pages are actively moderated, more silent followers like our comments (more on this later).

Lessons – is this really hate?

During the past few years it has become clear that there is a deep dissatisfaction among parts of the population, which cumulates in the following observation: "I can finally say this and people listen to me." Or in an irrational envy of others, primarily refugees: "They are given things, but what about us?!" Over both of these statements hovers the feeling of membership of a group, a (partially) understandable rage or anger against the government, which has contributed to a subjective feeling of reduced security or being left behind socially. This has led to a pessimistic and cynical atmosphere in the comment sections of social media. However, this analysis is only half of the truth.

What else is involved?

What is much more dramatic and serious is the continuous propagation of insecurity and discord by the supplementary and deliberate use of cynicism, malice, half-truths and manipulated statistics in addition to what has already been described. Daily reading of the comment columns leads to a further realisation: over a long period, the same disparaging and sometimes inflammatory comments and posts can be seen over and over again, for example comments claiming that journalists, the media and the counterspeech activists are deliberately lying. Verbal influencers and co-ordinated campaigns naturally follow Orwell's remark that 'lies become truth' in the hope of creating a parallel online world, in which everything is questioned, and dubious and dishonest alternative internet or news pages are believed. This is one of the greatest dangers to democracy.

Social media as a stage

Bearing the information just been presented in mind, only one conclusion is possible: those for whom our democracy and the pluralistic exchange of opinion is important must continue to demonstrate character and commit to decency in comment sections and fight hate on the internet. Of course, all social media users can state their opinions, but they should do this in a level-headed way and always be aware that Facebook is nothing other than a large stage with a potentially very large audience. Against this background, the fact that verbal opponents like to provoke should be accepted with as much calm and composure as possible. Depending on the context of a particular conversation, asking questions, displaying humour, requesting sources, calling attention to the disparaging tone of others, or returning to the core topic when there are distractions ('whataboutism' or 'derailing') can be suitable methods to show followers who are interested in a decent debate.[25] One's own reflections, which include a readiness to grant the other person the right to be correct, can de-escalate a potentially inflamed discussion.

The role of the media

Many media pages provoke particular reactions and emotions simply with the design and framing of their articles (headlines, introductions to an article, selected photos). The more clicks generated, the greater the advertising revenue. Without proper moderation by social media editors, it can be observed that less differentiated, generalising or even inflammatory comments generate the most responses.

Unmoderated comment columns easily become the targets of organised groups or troll or fake accounts, which use the opportunity to place cleverly and increase the quantity of their half-truths, disinformation, political messages and patterns of argument (in the form of 'likes' and encouragement) and further shift the boundaries of what can be said. On the other hand, fact-based, easily understandable, well-researched and objective reporting in combination with good management of comment columns avoids the aforementioned effects and therefore contributes to maintaining decency in online discussions.

6.5 Conclusion: counterspeech as a response to hate and extremism on the internet

As is shown by the two examples from German counterspeech campaigns can function in the online environment, but only to a certain extent. Campaigns of this type can certainly generate attention and interaction with other users if they are well thought out and have a large number of supporters. However, this is often difficult for civil society campaigns, which often lack sufficient resources to measure whether and how effective counter-narratives are, even though models exist that enable such evaluation. Among others, the ISD has carried out research on the monitoring and evaluation of counterspeech campaigns, but it is questionable whether civil society organisations have the financial or staff capacity to evaluate them comprehensively.

[25] *Especially humour is not always a suitable method, as this can be perceived as patronising or arrogant.*

Furthermore, counterspeech is usually, if not always, defined as a response to hate and enmity and therefore reactive. Because of this, the narratives dictated by haters and extremist groups are legitimised, and alternatives rarely offered.

Counterspeech campaigns run the risk of acting in a primarily defensive and reactive manner instead of initiating new and innovative measures – and actually starting to have a positive effect on prevalent attitudes in society. Not all counterspeech campaigns are successful; some have negative effects, for example the British campaigns More than a Refugee[26] and Hug A Jihadi,[27] which both caused a right-wing backlash. In summary, although counterspeech campaigns can function, and the examples in this chapter provide important insights into the design and implementation of successful campaigns, they are not enough by any means. Both initiatives are a start, but we need more than just a response to hate and enmity on the internet. New narratives must be developed to fight the brutalisation of the internet.

References

1. *Braddock, K. and J. Horgan (2016) Towards a Guide for Constructing and Disseminating Counternarratives to Reduce Support for Terrorism, Studies in Conflict & Terrorism, 39(5), 381–404.*

2. *Council of Europe (2007) Guidelines of the Committee of Ministers of the Council of Europe on Protecting Freedom of Expression and Information in Times of Crisis, https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=09000016805ae60e [16.07.18].*

3. *Hemmingsen, A.-S. and K. I. Castro (2017) The Trouble with Counter-Narratives, Danish Institute for International Studies, http://pure.diis.dk/ws/files/784884/DIIS_RP_2017_1.pdf [16.07.18].*

4. *Frischlich, L., D. Rieger, A. Morten and G. Bente (2017) Videos Against Extremism: Putting Counter-narratives to the Test, Bundeskriminalamt, Kriminalistisches Institut, Vol. 51.*

5. *Initiative für Zivilcourage Online [Initiative for Civil Courage Online] (2016) Information Pack on the Topic of Counter Speech, https://www.isdglobal.org/wp-content/uploads/2017/10/OCCI-Counterspeech-Information-Pack-German.pdf [16.07.18].*

6. *Silverman, T., C. J. Stewart, Z. Amanullah and J. Birdwell (2016) The Impact of Counter-Narratives, Institute for Strategic Dialogue, https://www.isdglobal.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE_1.pdf [16.07.18].*

7. *Sponholz, L. (2016) Islamophobic Hate Speech: What is the Point of Counter Speech? The Case of Oriana Fallaci and The Rage and the Pride, Journal of Muslim Minority Affairs, 36(4), 502–22.*

8. *Tuck, H. and T. Silverman (2016) The Counter-Narrative Handbook, Institute for Strategic Dialogue, http://www.isdglobal.org/wp-content/uploads/2016/06/Counter-narrative-Handbook_1.pdf [16.07.18].*

[26] *See https://www.youtube.com/watch?v=Lxbdvo2vFwc*
[27] *See https://www.sbs.com.au/news/dateline/story/hug-jihadi*

7. Suggested solutions: hate speech and extremism in the context of the NetzDG – recommendations to policymakers, social networks and civil society

By Jakob Guhl and Johannes Baldauf

7.1 Recommendations from the previous articles

In the previous articles, the authors have made several suggestions as to how the specific problems they are facing in their fields of expertise and activism can be counteracted. These recommendations are briefly summarised again here.

In Chapter 1, Simone Rafael and Alexander Ritzmann suggest that according to the model of the EU Internet Forum established by the European Commission, representatives of tech companies, research and civil society (who unfortunately do not yet play a role in the EU Internet Forum) should be brought together to try to restrict online access to extremist materials and to promote alternative narratives from civil society by means of self-regulation of the internet industry.

In Chapter 3, on filter bubbles, the psychologist Christian Montag above all calls for deeper research into the existence, function and effect of filter bubbles in political discourse. Here, it would be especially important for scientists to have easier access to platforms such as Facebook in order to investigate filter bubbles on social media more closely. According to Montag, differential psychological approaches can be particularly helpful to discover what personality traits coincide with susceptibility to filter bubbles. It cannot be ruled out that certain groups of people are more susceptible to the effects of filter bubbles than others. Such research is essential in order to develop ideas of how the creation of filter bubbles and their possible effects on radicalisation of people can be prevented.

Josef Holnburger and Karolin Schwarz in Chapter 4 call for further research by social media platforms into the function and dissemination methods of disinformation, and recommend that social media users are given media and information competence training. It is important that professional fact-checking journalism can reach the consumers of disinformation better than has been the case up to now. Holnburger and Schwarz consider that this is the responsibility of social media companies.

In Chapter 5 Matthias Quent calls for the strengthening of the narratives of marginalised groups affected by hate speech on the internet, by reporting racist content and counterspeech campaigns against such discourses. He recommends that easily accessible advice centres should be created for victims of online hate speech, education campaigns be run to reduce online and offline discrimination, and that people be informed about the aims and tactics of the central protagonists of internet hate.

Finally, in Chapter 6 on the evaluation of counterspeech campaigns, Alexander Urban from #ichbinhier and Sina Laubenstein from the No Hate Speech Movement emphasise that not all campaigns against hate can be judged to be effective. Therefore, it is important that there is a strategy and clearly defined target group for each counterspeech campaign. A clear approach helped the No Hate Speech Movement to achieve a broad reach and a large number of interactions, while #ichbinhier was successful in convincing many 'silent followers' with its members' calm, objective and balanced comments.

For this concluding article we have compiled a series of further recommendations for policymakers, social network and civil society, which we hope can make a positive contribution to a cross-sector response to the dissemination of hate speech and extremism.

7.2 Recommendations: politics

There should be co-operation to bring about a digital uprising by decent people. In Chapter 6, an evaluation of counterspeech campaigns, Sina Laubenstein and Alexander Urban find that "those for whom our democracy and the pluralistic exchange of opinion is important must continue to demonstrate character and commit to decency in comment sections". We can only concur with this. We need a digital uprising by decent people to stand up to hate speech and extremism on the internet. It is important that the various groups that are especially important when combatting online hate act in parallel and understand their efforts in a common project.

With the passing of the Internet Enforcement Act (NetzDG) in Germany, politicians have made it clear that they will not look on inactively if hate speech and anti-constitutional statements are disseminated on social media. Despite all the to some extent justifiable criticism of the act from legal,[28] technological,[29] human rights,[30] journalistic and civil society[31] perspectives, this should be considered a very positive sign. As Simone Rafael and Alexander Ritzmann show in Chapter 1, an appropriate response by politicians to hate speech on the internet was long awaited. In 2015 the (then) Federal Minister of Justice, Heiko Maas, founded the Task Force for Dealing with Hate Speech and invited representatives from civil society and major social media representatives to the discussion table. However, through the NetzDG, the topic of hate speech was mostly reduced to consideration of legal questions and social networks were assigned responsibility for enforcing the law by removing illegal content.

It is now important to combine expertise and resources to enable a digital uprising by decent people, which combines a broad alliance of voices, including from civil society. The majority of social media users, the democratic users, must be able to speak out against incitement of hatred and unconstitutional content, as well as hate-filled and discriminatory content which is not criminal, and therefore does not fall under the NetzDG.

A new task force against hate speech?

Although there have been great differences in opinion between politicians, social media company representatives and civil society in the past, which will no doubt continue into the future, the creation of groups such as the task force should be reconsidered. Politicians, especially those in the Federal Ministry of Justice and the Federal Ministry of Family Affairs, Senior Citizens, Women and Youth, should clearly indicate that they remain interested in holding a constructive dialogue between the various sectors. Through debate with representatives of social media companies and civil society organisations it could be determined to what extent there is a need to adjust for legal issues and governmental support of civil society efforts against hate speech and extremism on the internet.

As Simone Rafael and Alexander Ritzmann describe in Chapter 1, with the EU Internet Forum, the European Commission now uses a similar procedure to the one used by the task force described above to bring together representatives from politics, tech companies and research to discuss how to restrict online access to extremist material and promote alternative narratives from civil society. Co-operation in such forums is very important for negotiations between the various actors, who frequently have different and sometimes conflicting positions in these debates. For example, it is critical that topics such as the tension between freedom of opinion and freedom from harassment are identified and openly examined. Such debates should take the international context into account.

A common framework for combatting hate speech and extremism on the internet

Those involved in discussing the problems of hate speech and extremism on the internet from different perspectives should ideally question ostensibly simple solutions, for example the idea that hate speech and extremism on the internet can only be solved by removing problematic content from the larger platforms. Has anything really been gained if hate-filled users simply move to platforms that are less actively moderated, such as Gab[32] or the Russian Facebook alternative VK? On these platforms they are among even more ideologically like-minded people, in the filter bubbles analysed by Christian Montag, and are even more difficult to access with counterspeech and alternative narratives.

[28] *For example Bitkom (2017).*

[29] *The blogger Sascha Lobo (2018) described the NetzDG as "technically uninformed".*

[30] *The human rights organisation Human Rights Watch (2018) warned that "forcing companies to act as censors for government is problematic in a democratic state and nefarious in countries with weak rule of law".*

[31] *Above all in the form of the Declaration for Freedom of Opinion: https://deklaration-fuer-meinungsfreiheit.de/.*

[32] *Gab.ai is a social media platform modelled on Twitter, which is very popular with right-wing extremists because the content is hardly moderated.*

A further common fallacy suggests that hate speech can be effectively reduced by forbidding anonymous profiles. But in 2016 researchers at Zürich University found that non-anonymous users write more aggressive comments than anonymous users (Rost et al. 2016).

To avoid such false conclusions, we believe that it is essential that politicians and representatives of social networks and civil society develop a common framework that defines central terms, specifies targets, develops indicators to measure progress in reaching targets, and selects methods of evaluation. Agreement of a common framework with the aid of these four steps can prevent politicians and representatives from social networks and civil society from drifting apart in their analyses, responses and assessments of the problems surrounding hate speech and extremism on the internet, as has been the case up to now. The framework should:

1. Define central terms and risk analysis
2. Define goals against hate speech and extremism on the internet
3. Identify possible responses and courses of action
4. Create indicators to measure progress
5. Evaluate methods

Such a long-term framework, including frequent communication with representatives of social media companies and civil society, would demonstrate that politicians continue to take the subject seriously and not presume that it has been dealt with sufficiently with the passing of the NetzDG. One of the dangers we observed when discussing the NetzDG is that problems of hate speech and racism online are reduced to questions of criminality and moderation initiatives, education projects and counterspeech campaigns are unused. A common framework supported by experts from civil society and researchers should advocate a multi-layer response to hate speech and extremism, which includes the removal of particular content as well as counterspeech campaigns and alternative narratives.

In order for the common framework to help, there should be no exclusive focus on the larger social media platforms but instead the entire online ecosystem, including smaller platforms in which misanthropic content is disseminated, should be examined.

Continuous support of civil society commitment
Continuity in the fight against hate speech is also required from politicians in financing education and prevention projects, especially by supporting civil society initiatives. For example, the federal programme Living Democracy!, run by the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth, can play an important role in the fight against hate speech in general and especially on the internet.[33] It is extremely praiseworthy that there is now a programme to encourage people to take action against hate on the internet, in addition to other efforts to promote democracy.[34]

Continuity is an important key word. By promoting civil society on the internet over the long term, politicians can go further than merely endorsing well-meaning attempts and pilot projects, but rather create permanent structures that oppose hate and extremism on the internet.

This is all the more important because not all politically problematic content can or should be removed: some hate speech is not criminal, and determined opposition to it by the silent majority of democratic users from civil society is of primary importance.

Joint projects with private actors
Joint projects with support from politicians and people in the private sphere, for example representatives of social media companies or private foundations, would be advisable in order to combine resources and expertise as well as improve the sometimes apparently antagonist relationship between social media companies and politics.

[33] *See https://www.demokratie-leben.de/*
[34] *See https://www.demokratie-leben.de/mp_staerkung-des-engagements-im-netz-gegen-hass-im-netz.html*

Projects such as Das NETTZ, which is supported by the Robert Bosch Foundation and the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth, serves as a link for various civil society protagonists against hate on the internet, and could be a model for others.[35] Promoting research into the effectiveness of counter-speech campaigns can only improve their quality and the willingness of civil society to contribute to such initiatives. Studies such as *Videos Against Extremism*, which investigated the effectiveness of counter-narrative videos against right-wing extremism and Islamism, are a good start (Rieger et al. 2018). The effect of counterspeech campaigns should always be included in research on counter-narratives so future campaigns can use a strong evidence base.

Equipping digital citizens against online dangers
This report has shown that problems such as hate speech, disinformation and extremism on the internet have changed rapidly within a short time, partly because new technologies have made it possible to disseminate content in a targeted manner and on a large scale. These challenges will continue to change, as technologies will develop further in new and exciting ways. In view of this increasing technical complexity, young people in particular must be equipped for the dangers which they are exposed to online.

Politicians and tech companies should therefore support the development of programmes that promote understanding of the increasingly wide spectrum of online harms. Although digital education can play an important role in protecting young people against such risks, it is essential that education programmes also involve young people in a more positive dialogue about their online behaviour and communities so they become positive and proactive digital citizens. It is very important to help the next generation to develop methods by which they can design the internet in a respectful and open manner if it is to be maintained as an instrument for freedom and networking in the future.

7.3 Recommendations: civil society

Digital literacy: "it's no longer OK to not know how the internet works"
In December 2011, the journalist Janus Kopfstein (2011) formulated the sentence "Dear Congress, it's no longer ok to not know how the internet works". This statement related to the Stop Online Piracy Act[36] hearing in the US Senate, in which the deficits in technical knowledge and internet culture of American representatives became apparent.

Even though Janus Kopfstein's quote had a different context, it is still transferable to the digital commitment against right-wing extremism and group-related enmity: "Dear civil society, it is no longer OK to not know how the internet works."

What do we mean with this provocative statement?
Over many years, there have been no projects by well-known organisations in the field explicitly dedicated to the digital environment. Without empirical experience, there can be no expertise. Ultimately, it is thanks to the untiring work of the Amadeu Antonio Foundation and individual actors that there is any digital expertise in the German-speaking region.

Strong presence
The digital environment is not a niche area, but central for political communication and social discourse. Therefore there is a need for organisations to propose innovative projects and have a stronger online presence – not just to have a perceivable voice on social networks, but also to be able to deal with attacks. Civil society organisations are increasingly frequently the targets of the same sort of shitstorms and hate campaigns as marginalised groups are.

[35] *See https://www.das-nettz.de/*
[36] *The Stop Online Piracy Act is a draft law which is intended to enable the owners of rights to prevent the unauthorised dissemination of content which is subject to copyright.*

Organisations must prepare themselves better for this and develop approaches based on data and experience for moderation and community building. It is advisable for them to have permanent staff for social media support, which ensures good and regular moderation of their pages.

Support for victims
The precise number of people who have been attacked online is not documented, but shitstorms, threats, doxxing and silencing are common components of right-wing extremist strategies against individuals. There are no established structures where those affected can obtain help and support. Up to now, victim advice centres have developed only selective expertise for online cases. Unless such support structures are created to respond quickly and competently to abuse online, digital engagement will decrease. Therefore, it is advisable to raise awareness deliberately and train victim advice centres to deal with these problems, or even to set up advice centres with an explicit online focus.

Research and observe
Online hate speech, right-wing extremist and hate campaigns are short-lived. To respond in real time and appropriately, continuous research, observation and targeted monitoring is required. Although there are mobile advice teams and initiatives throughout Germany whose members monitor and document local activities, no comparable structures exist for social networks.

It is central to the work of civil society activists that they are digitally literate so they can monitor hateful and extremist actors in the digital environment. We recommend that a federal, inter-organisational association should be set up, which performs cross-platform monitoring.

7.4 Recommendations: social networks

Companies are almost exclusively the focus of combatting hate speech because of reporting in the media and the strategic decision of the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth to focus on the way platform operators remove content instead of the social problem of racism. It is also the result of a lack of (or belated) awareness of the socio-political consequences and dark sides of global networking (Chapter 5).

The fact that efforts are being made by platform operators cannot be overlooked: co-operating cross-company in the fight against terrorism and child pornography, adopting community standards, publishing internal regulations for dealing with toxic content, co-operating to combat disinformation, and taking a series of measures that protect against influencing elections show that the problematic situation has been recognised and solutions are being sought to improve the platforms. Furthermore, Google has created #nichtegal and Facebook the OCCI in order to promote and strengthen the engagement of civil society in the digital sphere.

Greater involvement of civil society expertise
Finding globally valid rules for coexistence is a difficult task, with which even the international community continually struggles. In regional and national debates about the handling of toxic content on the internet, this global perspective is seldom considered and companies are expected to have a satisfactory solution at hand for every state or region.

Ultimately, this concerns the question of whether it is possible to find values that are globally valid. For example, in dealing with right-wing extremism and hate speech, the central question is whether freedom of opinion is a basic right, which has a greater weighting than the protection of human dignity, the fundamental right of participation in society and the free development of personality.

These are traditionally tough debates between those in civil society and politicians, who negotiate these values. Companies would be well advised to develop appropriate formats to allow room for the debates about the validity and weighting of values. It is undisputed that very clever people are working on this question within companies, but their community relies on participation, and politicians and members of civil society want to and must be involved in these processes – from discussions at the meta-level to asking detailed questions, for example on the precise boundary between hate speech and freedom of opinion. In addition to global and regional forums for these debates, experts from civil society and academia must be more (or more visibly) involved, because the desire for participation is behind many demands for greater transparency.

In order for academics and civil society representatives to make a profitable contribution to these debates, they must understand the phenomena and developments better. This requires having qualified access to the data, so legal and ethical questions on data access and protecting the privacy of users must be considered. Facebook's collaboration with the Social Science One Initiative provides a good model for how researchers can access verified and comprehensive data.

Platform functions and measures: considering mechanisms of online influence

At the moment, the measures and discussions on extremism and hate speech on the internet are mainly concentrated on problematic content, while the methods for disseminating such content do not receive sufficient attention. The influence of various platform functions and algorithms should be considered in the development of specific measures. Extremist and malicious actors already use the latest technologies and social media functions in order to develop new methods for recruiting and scaremongering on the internet; for example with the aid of inorganic multipliers or micro-targeting measures. While most companies have made their removal procedures and criteria considerably more transparent, they should also introduce higher standards of transparency about the design and algorithms behind their products and functions. This would counteract the abuse of existing and new technologies for extremist and anti-democratic purposes.

Socio-political responsibility

Every company has a socio-political responsibility, especially if its business is global communication and networking. As has been described above, such a responsibility is highly complex and companies are increasingly facing up to this responsibility. Companies such as Facebook, which focus on community and its protection, need to be socio-politically active and take clear positions.

Protecting the community ultimately leads to protecting society. Companies should make considerably greater and more active efforts to protect society and its democratic culture, for example by giving financial support to civil society protagonists, and paying more attention to the previously almost ignored area of victims of hate speech.

In the hate speech debate, as well as in this publication, there is frequently a call for a new uprising of decent people. Even though the original call for such an uprising in October 2000 was made by the (then) German Chancellor Schröder – a politician – the call was strongly echoed by economists, so that an effective alliance of actors in politics, civil society and companies could be formed. In the present climate of global backlash movements and the central role played by social media it would demonstrate a high level of social responsibility if tech companies were the initiators of a new uprising of decent people.

References

1. *Bitkom (2017) Bitkom statement on the government draft of an Internet Enforcement Act. https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Stellungnahmen/2017/Downloads/04202017_Stellungnahme_Bitkom_2_RefE_NetzDG.pdf?__blob=publicationFile&v=3 [18.07.18]*

2. *Human Rights Watch (2018) Germany: Flawed Social Media Law. NetzDG is Wrong Response to Online Abuse. https://www.hrw.org/news/2018/02/14/germany- flawed-social-media-law [18.07.18]*

3. *Kopfstein, J. (2011) Dear Congress, It's No Longer OK to not Know How the Internet Works, Motherboard Blog, https://motherboard.vice.com/en_us/article/pggamb/dear-congress-it-s-no-longer-ok-to-not-know-how-the- internet-works-5886b6cbc860fd45c9f2dfe3.*

4. *Lobo, S. (2018) The Fiasco of the Hate Speech Law: The Blunt Instrument NetzDG, Der Spiegel, http://www.spiegel.de/netzwelt/web/netzdg-berechtigtes-getoese-um-ein-daemliches-gesetz-a-1185973.html [18.07.18]*

5. *Rieger, D., L. Frischlich, A. Morten and G. Bente (eds) (2018) Videos Against Extremism? Putting Counter-narratives to the Test, Research Unit Terrorism, Extremism of the Federal Criminal Police Office.*

6. *Rost, K., L. Stahel and B. S. Frey (2016) Digital Social Norm Enforcement: Online Firestorms in Social Media, PLoS One, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155923.*

ISD London | Washington DC | Beirut | Toronto
Registered charity number: 1141069

www.isdglobal.org

ISD | Powering solutions to extremism and polarisation

PO Box 7814, London
United Kingdom, W1C 1YZ
www.isdglobal.org