

---

JULY 2021

---

---

# Courage Against Hate



FACEBOOK



---

# Introduction

---

The Courage Against Hate initiative has been brought together by Facebook for the purpose of sparking cross-sector, pan-European dialogue and action to combat hate speech and extremism. This collection of articles unites European academic analysis with practitioners who are actively working on countering extremism within civil society.

Hate and extremism have no place on Facebook and we have been making major investments over a number of years to improve detection of this content on our platforms, so we can remove it quicker - ideally before people see it and report it to us. We've tripled - to more than 35,000 - the people working on safety and security at Facebook, and grown the dedicated team we have leading our efforts against terrorism and extremism to over 350 people. This group includes former academics who are experts on counterterrorism, former prosecutors and law enforcement agents, investigators and analysts, and engineers. We've also developed and iterated various technologies to make us faster and better at identifying this type of material automatically. This includes photo and video matching tools and text-based machine-learning classifiers. Last year, as a result of these investments, we removed more than 19 million pieces of content related to hate organisations last year, over 97% of which we proactively identified and removed before anyone reported it to us.

While we are making good progress, we know that working to keep hateful and extremist content off Facebook is not enough, because this content proliferates across the web and wider society, often in different ways.

That is why we also need partnerships to do this work well. In order to be truly effective in stopping the spread of terrorist and extremist content across the internet, we need to join forces with others. Four years ago, we started meeting with other technology companies to discuss the best ways to counter terrorists' attempts to use our services through the [Global Internet Forum to Counter Terrorism \(GIFCT\)](#). We've subsequently also partnered with experts in government, civil society and academia to discuss and study the latest trends in terrorist and violent extremist activity online. Partnerships like this allow technology companies

like us to see beyond trends on our own platform, and better understand the interplay between online and offline. They also help us build programs with credible civil society organizations to support counter-speech at scale, and how best to deploy multi-sector efforts to challenge hate speech and extremism while respecting fundamental principles of human rights.

Courage Against Hate is a strategic policy initiative brought together with this goal in mind. We hope it will help us collectively develop a more holistic understanding about the diverse and evolving nature of the problem and prompt a multi-disciplinary conversation around what policies, further analysis and programs are needed for the fight against hate, extremism and terrorism to be truly effective. This publication highlights that we are by no means starting from zero. Substantial work has been done in the last five plus years. New methodologies have been developed, tools have been tested, programs have launched innovative approaches to reaching at-risk and vulnerable audiences. We hope that this publication shows how much technology companies, governments, researchers and civil society all have very important roles to play, and how much partnerships are needed to tackle the societal issue that hate speech represents.

---

## RESEARCH BRIEFINGS

Right-wing extremism is constantly changing and adapting to global developments. [Hope not Hate](#)'s Joe Mulhall and Safya Khan-Ruf text examine the trajectory of the rhetoric, tactics and direction of European far-right in the midst of a global pandemic in 2020 and the rise of the International Black Lives Matter movement. But right-wing extremism is not only reacting to such developments: Matthias Quent, Christoph Richter and Axel Salheiser from the [Jena Institute for Democracy and Civil Society](#) showcase in their text how the far-right is also adapting to subcultures and has modernized itself to fully embrace the digital space.

Bethan Johnson ([Centre for the Analysis of the Radical Right](#)) similarly explores the increasing trends of white supremacy across Europe and North America in her analysis piece on Siege Culture - one of this subcultures' original publications which makes up the backbone of the modern movement. She explores the increasing trends of white supremacy across Europe and North America, pointing to one of these subcultures' original publications and texts that make up the backbone of the modern movement. Researchers from Uppsala University and

the [Swedish Defence Research Agency](#) further assesses the challenges and opportunities of automatic hate speech detection including domain transferability, reliability, languages and expressions, and how to avoid detection.

These expert briefings highlight once again that threats from the far-right need to be constantly monitored in order to learn from emerging trends and develop nuanced successful countermeasures, in what is a fundamentally adversarial field.

---

## PRACTITIONER CASE STUDIES

In order to showcase exemplary programmatic solutions and best practices in counterspeech, we further commissioned six case studies. The aim of these case studies was to map how a selection of European practitioners are putting forth efforts to strategically combat hate speech and extremism, as well as how tactics for preventing and countering violent extremism, such as counterspeech/counter-narratives, are best deployed, and help a broader audience understand the interplay between online and offline interactions in this field.

The resulting case studies presented in this report are diverse and cover a range of focus areas and programmatic approaches. Sweden-based [#iamhere](#) speaks to their approach to counterspeech, specifically their method to reduce hate speech and misinformation online. [Galop](#) in the UK focuses on their approach to combating hate speech and extremism based on mutual support and cooperation. Meanwhile, [Moonshot](#), presents their work on the Redirect Method, using targeted advertising to connect people searching the internet for violent extremist content with safer alternatives. The [Institute for Strategic Dialogue](#) (ISD) sheds light on the Online Civil Courage Initiative (OCCI) aimed at uniting a diverse set of actors, providing them with the latest research into hate and extremism online and fostering a collaborative environment via conferences and hackathons. Austria-based [Zivilcourage und Anti Rassismus-Arbeit](#) (ZARA) puts forward their practices aimed at combating racism and online hatred and to promote, strengthen, and increase the value of civil courage within the Austrian society. Finally, [Textgain](#) and [Media Diversity Institute](#) look at how their respective projects Detect Then Act and Get The Trolls Out use AI and manual media monitoring to mitigate discrimination and intolerance.

---

## EXISTING GUIDELINES AND RESOURCES

We understand that we are starting from a point where much has already been done in the last few years to fight against hate speech in all its forms and to ensure that multi-sector approaches evolve to effectively tackle adversarial shifts as they happen. For that reason, we want to highlight some of the existing toolkits and resources created to provide guidance, support and best practices for developing, launching, measuring and evaluating work that challenges hate speech and extremism online and offline.

- [Facebook's Counterspeech Hub](#) exists to support voices that are engaged in counterspeech efforts. The site provides overviews of a variety of counterspeech programmatic efforts that Facebook supports around the world and links to open-access resources and analysis to help guide practitioners. The hub helps promote accessibility to Counterspeech tools, resources, and guidance. Above all, it is there to elevate the dialogue beyond the reach of fear, hate, and violence.
- [The Campaign Toolkit](#) is a new and dynamic digital resource for educating, enabling, and empowering the next generation of activists and community organizations as they mobilize to outcompete hate and to promote community cohesion, inclusion and tolerance. Funded by the [Global Internet Forum to Counter Terrorism](#) and managed by ISD Global, the toolkit immerses you in the journey of planning, producing and promoting campaigns for global audiences. It provides a step-by-step guide as well as resources from leading technology companies and civil society. Available in English, French, German and Arabic.
- [European Derad Toolkit](#): As part of the major and sustainable outcomes of the DERAD project, a toolkit with different elements and target groups is developed. Due to the scope of the topics of the project, the DERAD project produced an online platform. Both the platform and the toolkit aim at supporting the work of the main stakeholder groups active in CVE activities.
- [TERRA Toolkit](#) includes a series of training modules. [TERRA Train the trainers](#) program is for teachers, youth workers, police, prison staff, journalists, religious leaders and local and national authorities. This training is developed to support front line workers so that they can undertake prevention work in de-radicalisation, and support disengagement from terrorist and extremist groups.
- ISD Global has developed a number of training programmes and toolkits to help practitioners and activists develop, launch and measure counterspeech. These include the [NGOs Counter Narratives Monitoring and Evaluation Handbook](#), their report about assessing the [Impact of Counter Narratives](#) and the [Counter Narrative Handbook](#).
- The [Online Civil Courage Initiative Help Desk](#) works with civil society organisations to provide tailored support and guidance on developing, managing and measuring online counter-speech. Funded by Facebook and run by ISD Global, the help desk can help civil society groups design a campaign that meets specific needs and objectives, and connects them with organisations who can fill any skills gaps that might be apparent. You can access this resource in English, French or German.

---

# Table of Contents

---

## RESEARCH BRIEFINGS

- |           |  |           |
|-----------|--|-----------|
| <b>01</b> | <b>The European Far-Right in the Year of COVID-19 and Black Lives Matter: Changing Targets, Tactics and Rhetoric</b><br>HOPE not hate                          | <b>06</b> |
| <b>02</b> | <b>Recent Trends and Perspectives: COVID-19 Pandemic Increases Right-Wing Extremism and Hatred in Europe</b><br>Jena Institute for Democracy and Civil Society | <b>20</b> |
| <b>03</b> | <b>‘Join in or else die!’: Siege Culture and the Proliferation of Neo-Nazi Narratives Online</b><br>Center for the Analysis of the Radical Right (CARR)        | <b>36</b> |
| <b>04</b> | <b>Automatic Hate Speech Detection: Challenges and Opportunities</b><br>Swedish Defence Research Agency and Uppsala University                                 | <b>57</b> |

## CASE STUDIES

- |           |   |            |
|-----------|---|------------|
| <b>05</b> | <b>The Kaleidoscope of Counterspeech: From The “Silent Majority” to the “Vulnerable Observer”</b><br>#iamhere           | <b>70</b>  |
| <b>06</b> | <b>Galop and Community Partnership Working to Combat Hate Crime</b><br>Galop  | <b>84</b>  |
| <b>07</b> | <b>Redirect North America: Challenging White Supremacist Extremism on Google Search</b><br>Moonshot                     | <b>93</b>  |
| <b>08</b> | <b>The Online Civil Courage Initiative (OCCI)</b><br>The Institute for Strategic Dialogue (ISD)                         | <b>101</b> |
| <b>09</b> | <b>Together #AgainstOnlineHate in Austria</b><br>Civil Courage and Anti-Racism Work (ZARA)                              | <b>111</b> |
| <b>10</b> | <b>Using AI and Advocacy-driven Counternarratives to Mitigate Online Hate</b><br>Textgain and Media Diversity Institute | <b>119</b> |

---

RESEARCH BRIEFINGS

---

# 01 The European Far-Right in the Year of COVID-19 and Black Lives Matter: Changing Targets, Tactics and Rhetoric

---

by Joe Mulhall and Safya Khan-Ruf

ADDITIONAL RESEARCH:

by David Lawrence, Gregory Davis  
and Patrik Hermansson

EXTRACTS FROM:

HOPE Not Hate

---

This year has seen a series of major international events that have not only begun to reshape elements of European society but have also had a notable effect on the rhetoric, tactics and direction of the European far-right. Simultaneously, we are witnessing the dramatic effect of a global pandemic and the rise of the international Black Lives Matter movement that has impacted Europe as well as America. The purpose of this article is to outline an array of directional trends within the European far-right that have been brought about by a series of events during 2020. While it is, of course, too early to make definitive statements on exactly how these concurrent events will affect the European far-right, there are already some strong indications of the direction of travel that this article will explore.

The article is split into two sections. The first section will explore how the COVID-19 global pandemic has affected the politics and direction of the European far-right. The second section will explore the European far-right's reaction to the rise of Black Lives Matter movements across the continent and the broader societal discussions about race, systemic prejudice and imperial legacies. By highlighting ideological and rhetorical shifts within the far-right we can better plan for the changing nature of the threat they pose both online and offline. Importantly, this year has already seen a number of broad pivots that have affected the targets they attack and the lexicon they employ to do so. Understanding these changes is essential for effective opposition to, and moderation of, content produced by far-right content creators.

It is important to state from the outset that any overview of the European far-right will necessarily talk in broad terms. While this article will explore specific cases and examples, the point here is to use them to highlight wider shifts and effects. This is especially important to understand when talking about the contemporary online far-right. While it remains important to explore changes in traditional far-right organisations such as political parties, the modern far-right is currently undergoing a broader and more fundamental shift; namely the emergence of a transnational and post-organisational threat. The international far-right scene today is a mixture of formalised far-right political parties, such as the Sweden Democrats, Vox in Spain, Lega in Italy and the AfD in Germany, and a series of looser, transnational far-right movements comprised of a disparate array of individuals collectively but not formally collaborating. In the age of the internet we have seen the emergence of disparate movements such as the anti-Muslim 'counter-jihad' movement and the international alt-right. While all these groupings have formal organisations within them, they are often post-organisational. Thousands of individuals, all over the world, offer micro-donations of time and sometimes money to collaborate towards common political goals, completely outside traditional organisational structures. These movements lack formal leaders but rather have figureheads, often drawn from an increasing selection of far-right social media 'influencers'. For most of the post-war period, 'getting active' required finding a party, joining, canvassing, knocking on doors, distributing leaflets and attending meetings. Now, from the comfort and safety of their own homes, far-right activists can engage in politics by watching YouTube videos, visiting far-right websites, networking on forums, speaking on voice chat services like Discord and trying to convert 'normies' on mainstream social media platforms like Twitter and Facebook. The fact that this can all be done anonymously greatly lowers the social cost of activism.

These new movements are best understood as a many-headed hydra. If one prominent activist or leader falls from grace, it is no longer a fatal hammer blow; others will simply emerge and the besmirched are discarded. Of fundamental importance is that these movements are genuinely transnational. While activists will generally be primarily preoccupied with local or national issues, they invariably contextualise them continentally or even globally. Often activists from all over the world come together for short periods to collaborate on certain issues and these loose networks act as synapses passing information around the globe. An Islamophobe in one country outraged by the serving of halal chicken in their local fast food restaurant can post on social media and the story will spread through

the network. If picked up by a ‘supersharer’ (an especially influential activist with a large social media following) that local story will be picked up by like minded Islamophobes all over the world and act as more ‘evidence’ and further convince them of the threat of ‘Islamification’. If we are to truly understand the contemporary far-right, we must therefore change our thinking. We live in a shrinking world: be it in our own community, our own country, continent or globe, we are interconnected like never before. Our ability to travel, communicate and cooperate across borders would have been inconceivable just a generation ago and while these opportunities are by no means distributed evenly, they have opened up previously impossible chances for progress and development. Yet greater interconnectivity has also produced new challenges. The tools at our disposal to build a better, fairer, more united and collaborative world are also in the hands of those who are using them to sow division and hatred around the world. If we want to understand the dangers posed by the politics of hatred and division we can no longer just look at our street, our community or even our country, we must think beyond political parties, formal organisations and even national borders. As such, all of the directional shifts discussed in this article should be understood as occurring to different extents in different parts of the European far-right, meaning both formal far-right organisations and post-organisational movements.

Finally, the variety of individuals, parties and movements discussed below make it necessary to briefly explain what is meant by ‘far-right’ in this article. While ‘far-right’ is a very broad term, those within it are united by a common set of core beliefs. Jean-Yves Camus and Nicolas Lebourg point out in *Far-Right Politics in Europe* that:

Far-Right movements challenge the political system in place, both its institutions and its values (political liberalism and egalitarian humanism). They feel that society is in a state of decay, which is exacerbated by the state: accordingly, they take on what they perceive to be a redemptive mission. They constitute a countersociety and portray themselves as an alternative elite. Their internal operations rest not on democratic rules but on the emergence of “true elites.” In their imaginary, they link history and society to archetypal figures [...] and glorify irrational, non materialistic values [...]. And finally, they reject the geopolitical order as it exists.<sup>1</sup>

Though ‘far-right’ is a useful umbrella term, its broadness makes it necessary to split it further into its constituent parts; the democratic radical right and the extreme far-right. The social scientist Cas Mudde explains that the extreme far-right ‘rejects the essence of democracy, that is, popular sovereignty and majority rule’, while the radical right, ‘accepts the essence of democracy, but opposes fundamental elements of liberal democracy, most notably minority rights, rule of law, and separation of powers’.<sup>2</sup>

---

Far-right movements challenge the political system in place, both its institutions and its values (political liberalism and egalitarian humanism). They feel that society is in a state of decay, which is exacerbated by the state: accordingly, they take on what they perceive to be a redemptive mission.



---

## SECTION 1:

### The Effect of the COVID-19 Pandemic on the European Far Right

The COVID-19 pandemic has dominated the news since the start of 2020, exposing cracks in government policies, causing divisions between politicians and re-establishing national borders within Europe. There have been over a million deaths worldwide, and nearly 47 million cases as of 1 November. Europe alone accounts for 233,000 deaths since the first recorded death in France on 15 February while the United Kingdom accounts for the highest number of deaths in Europe at just over 46,700 deaths.<sup>3</sup>

Citizens across the world have grown increasingly worried about the consequences of the coronavirus. The impact of the ongoing pandemic is only beginning to show itself, with the full predicted economic devastation likely to provide fertile grounds for the far-right. Researchers have pointed to a host of catalysts at individual and societal level that contribute to radicalisation.<sup>4</sup> Factors such as personal loss, the psychological burden and the economic instability created by the pandemic provide ideal grounds for far-right recruitment.<sup>5</sup> In a similar but much more widespread fashion than the 14th century plague, different minorities are being blamed and conspiracy theories about the pandemic abound across both the online and offline world.

It is of course too soon to definitively say what the effect on the far-right will be as things are still in flux. There are also large differences across the continent and the far-right has been impacted differently across Europe, depending on the politics of the country, the government's reaction to the pandemic and the power the far-right group had established before COVID-19. In the short term there have been both positives and negatives but it is true to say the far-right has generally failed to capitalise on the pandemic. Most European governments initially enjoyed a surge in popularity as the public rallies in solidarity and unity in the face of the pandemic. However, as the long term effects of the pandemic become known, and a second spike occurs, it could make some communities more susceptible to the far right in the medium to long term.

Unlike previous European flash points such as the refugee crisis of 2015 or the financial crisis of 2008, the far-right have not as yet enjoyed an immediate rise in popularity over the course of the year and across the continent. Despite the political squabbling over masks and the every-country-for-itself approach to COVID-19 in the first months, the populist and Eurosceptic elements of the far-right often failed to dominate the narrative. Many of the far-right parties failed to respond coherently, or with internal unity and took time to develop a new message. Attempts at rallying support against immigration for example, did not succeed in capturing the public mood. The pandemic has shifted migration rhetoric to include the risk to individual health, but the virus has not spread across Europe through the typical refugee and migratory routes. Instead, while far-right politicians were calling for closing ports in Italy, COVID-19 had already created clusters throughout the country. This has weakened the far-right's message associating safety with refusing immigrants. The fact that European countries did exactly what the far-right has been calling for and shut down borders in March also removed an important rallying point for far-right politicians from that point onwards. Their flailing strategies became more focused as the months went by, however, and researchers have seen a dangerous merging of far-right activity with more mainstream protests against lockdowns, masks and safety measures.

The far-right in Europe has scrambled to stay relevant amidst the pandemic with mixed results as priorities turned away from popular far-right talking points to pandemic-related issues. However, in the second half of the year, anti-mask, anti-lockdown and anti-safety-restriction protests have sprung up across the globe. The protests centre around how compulsory rules – even ones on health and safety – infringe on individual freedoms. Importantly, this has motivated both increased online and offline activity. Social media has played a central role in facilitating the growth of these COVID-skeptic groups and the far-right has consciously engaged in this space.

Several European states also had protests in April but the frequency and size of the protests were still small. By late August, thousands of anti-lockdown protestors filled London's Trafalgar square. On the same day, a rally in Berlin drew 38,000 participants. This was followed by other protests throughout September, ranging from hundreds to thousands of protesters, from Melbourne to Madrid to Montreal.

There is increasing worry amongst security forces and far-right researchers about the influence of far-right extremists within these protests. On 22 August a far-right segment of the protest in Dublin armed themselves with iron bars and batons and clashed with counter protesters. Police believe some of the masked men were part of Generation Identity.<sup>6</sup> Meanwhile at a Berlin protest, hundreds of far-right activists waved the black, white and red flag of the pre-1918 German Empire and stormed through a police barrier to force their way into the German parliament. In Rome in late October far-right protestors clashed with police over a series of nights during demonstrations against coronavirus restrictions.<sup>7</sup> The fact that the far-right shows distrust in government measures is not surprising and fits well with anti-establishment narratives. This includes the sinister theory of the police state and that governments are using COVID-19 to take freedoms away. It has also positioned far right groups at the centre of these protests.

However, these are not far-right only, or even far-right-led protests, and the blurred lines between their demands and the mainstream have enlarged their pool of potential recruits. The French left-wing think tank Fondation Jean Jaurès interviewed 1,000 anti maskers on Facebook and found 50 was the average age and 63% were women. “The epidemic has been gone for months,” one respondent said. “We are just collectively trained to submission,” she maintained.<sup>8</sup> While France’s anti-mask protests have not matched those of Germany or the United Kingdom in numbers, people expressed four reasons for not respecting the law: the mask was judged inefficient in stopping COVID-19 transmission, there was a lack of confidence in the institutions that are pushing these protective measures, a rejection of the elite, and a rejection on the impingement on personal freedom.

Anti-mask groups remain very active on social media, with Facebook groups across the continent attracting thousands of followers. All four of these reasons have been echoed by far-right politicians and online, aiding the positioning of the far-right as the voice of the people. The physical protests have also brought the online movement onto the streets.

## How the far-right in Europe is exploiting COVID-19

Traditional far-right parties across Europe have so far failed to make significant gains in the polls by exploiting COVID-19 fears. However, this does not take into account their success in pushing out hate, spreading disinformation online and exploiting the fear and uncertainty that the pandemic produced. In fact, researchers have warned against the far-right exploiting fears and radicalising the public since the start of the pandemic.<sup>9</sup> The spread of conspiracy theories was aided by the fact that as lockdowns were instated across Europe, more people spent time online. On 8 May, United Nations Secretary-General Antonio Guterres said, “The pandemic continues to unleash a tsunami of hate and xenophobia, scapegoating and scare-mongering” and urged governments to “act now to strengthen the immunity of our societies against the virus of hate.”<sup>10</sup> Politicians and even country leaders have also been guilty of encouraging hate crimes, racism and xenophobia through their rhetoric. Groups across Europe, including France, Germany, the UK, Spain, Greece and Italy, have seized COVID-19 as an opportunity to further nationalistic or anti-immigrant agendas as well as to demonise refugees and opposition groups.

The far-right attempts to rally around and gain supporters during COVID-19 differ greatly depending on whether they were opposition groups or in power. Authoritarian governments such as that of Hungary have exploited COVID-19 to give themselves greater powers and push through non-pandemic related legislation. In Serbia, there are allegations that the right-wing populist government manipulated the number of COVID-19 deaths prior to the June elections.<sup>11</sup>

As the economic and social strain began across Europe, it also offered an opportunity for far-right local movements to showcase their links to the communities and gain support from locals. In Italy, far-right group CasaPound posted photos of activists delivering groceries to the elderly while the extreme far-right group Hogar Social has done the same in Madrid. In Germany, neo-Nazi group The Third Way has been providing food to low-income households, and in the UK, Britain First and For Britain claimed to feed the homeless and volunteer at the NHS. Believers in accelerationism – the concept that Western governments are too corrupt to save and one should speed up their collapse by sowing chaos – have also welcomed the opportunities COVID-19 has created, and even celebrated it. The death tolls and confusion have entrenched these views and discussions around how the virus will bring on civil war and the collapse of society was a popular topic on far-right groups online.

## Rise of Conspiracy Theories

One of the most important and noticeable effects of the COVID-19 pandemic has been the growth of people engaging with conspiracy content online. The growth in the conspiracy theory has opened new doors for far-right recruitment but also encouraged the far-right to adopt more explicitly conspiratorial rhetoric to exploit this rise. Importantly, the online European far-right has also been central to the creation and dissemination of COVID-related misinformation.

Mapping the scale and nature of disinformation that spread during the pandemic is complicated by the staggeringly high levels of it on both mainstream platforms like Facebook and more obscure messaging apps like Telegram. These are often disseminated from the US, but have quickly spread within European networks. A study showed that between January and April, websites hosting disinformation received 80 million interactions on Facebook.<sup>12</sup> The conspiracy theories around the pandemic often focus either on it being a hoax, or that the virus is real, but was created or released intentionally by a host of different actors. Jews, Muslims, George Soros, Jeff Bezos, Bill Gates... all have been accused of designing and spreading COVID-19. A host of other theories surrounding the nature of the virus have also found popularity, such as the virus being caused by snakes, that washing hands is a propaganda by soap companies or that COVID-19 is being spread by Coca-Cola.<sup>13</sup> It is not surprising that inaccurate stories have spread about the virus during the uncertainty surrounding the pandemic, and many of the false or inaccurate theories have a very short half-life. However, the threat and danger of the misinformation caused by other theories cannot be underestimated. Some encourage violence and hate against public figures or minority groups, while others increase mistrust of safety procedures put in place to protect the general public against COVID-19.

The theories often “work” and spread using a few plausible facts that paper over the lies – this has made spreading these theories into the mainstream easier. For example, the conspiracy theory that Bill Gates was involved in creating the virus or knew there was going to be a COVID-19 pandemic focuses on a TED talk he gave in 2015 about the Ebola virus, where he warned the world was not ready for another pandemic. Another variation claims this is part of his plot to vaccinate the world and install microchips that will control people. Anti-vaxxers online have been vocal about the theory of a global vaccine program designed to kill a part of the population prior to COVID-19 and the pandemic has allowed them to repackage this within the current circumstances.

Far-right commentators with large platforms have actively pushed these theories out – often several contradictory narratives simultaneously – but traditional far-right parties have also spread these outside of online platforms. Marine Le Pen in France said that it made sense to ask if COVID-19 was made in a lab,<sup>15</sup> and 40 percent of her party believe that it was.<sup>16</sup> The conspiracy theories have also spread dangerously within the public discourse. In Spain, a poll in April showed nearly half of the respondents believed the virus had been created intentionally.<sup>17</sup> Meanwhile, a HOPE not hate poll showed two-thirds of people in the UK think it is important to seek alternative opinions about the coronavirus. This is not surprising considering mainstream newspapers have given platforms for certain conspiracy theories.

One of the most popular conspiracy theories alleges that China designed the virus in a secret lab in Wuhan. This held a certain level of plausibility for readers since the city of Wuhan also contains a virology institute where bat coronaviruses were being studied. A prominent virologist working at the lab said she was concerned enough by the theory to check that the COVID-19 genetic sequencing did not match any of the viruses studied at the institute – it did not. However, the theory, pushed online in part by a documentary produced by Epoch Times has slipped into the mainstream media.

Another popular theory asserts that China (or Russia or Israel or another country) created COVID-19 as a bioweapon. The US political far-right and even mainstream right are particularly taken with this theory, with even US senators propagating it.<sup>18</sup> This has also been thoroughly disproved through the genetic sequencing of the virus, showing it is of natural origin. The theories are also aided by the derogatory language used by politicians. The decision by President Trump to call the coronavirus the “Chinese virus” on 16th March seems an obvious attempt to stoke outrage and his supporters such as conspiracy theorist Paul Joseph Watson were quick to adopt his language. While Trump no doubt decided to do so in an attempt to deflect criticism around his handling of the outbreak and placing blame elsewhere, it has also stoked anti-Chinese sentiments.

Known far-right conspiracy theorists such as Alex Jones of Infowars have also been busy pushing several theories simultaneously, such as one claiming COVID-19 does not exist but is a fiction spouted by the “global elite” to remove freedoms. Anti-maskers, and far-right political parties have in various forms, taken up this theory across Europe. US president Donald Trump’s assertions that the virus is “no worse than the flu” has also been combined within that theory and taken up by protesters. There is also a direct marketing element to this – Alex Jones for example, sells pills that supposedly cures all diseases – making it in his interest to push other theories such as the virus being a plot by big pharmaceutical companies. Another far-right anti-vaccination and anti-abortion activist is Dr Annie Bukacek who warns viewers on YouTube that COVID death rates are inflated.<sup>19</sup> This theory has been widely taken up by the far-right – and elements of the mainstream – as a reason to ignore social distancing and lockdown measures.

The far-right are not the only source of the coronavirus conspiracies. Anti-GMO activists for example, pushed out that genetically modified crops (GMOs) were responsible for the virus, with Francisco Billota publishing an article in a non-politically affiliated Italian newspaper, asserting the virus was propagating faster due to GMO crops.<sup>20</sup> But even non-far-right conspiracy theories are being adopted by them and this remains a source of danger as they widen their pool of potential radicalisation. A popular adopted theory centres around 5G communications being the source or an accelerator of COVID-19. This has led to telecommunications apparatus being vandalised in Europe and elsewhere. In a similar method to TV networks crossing over two popular series to widen the viewership for each individual series, the crossing over of conspiracy theories and their mainstreaming have exposed a larger segment of the population to ideas that had remained within far-right circles prior to 2020.

## The Emergence of QAnon in Europe

The most formalised development in the area of conspiracy theories is the emergence in Europe of a small but growing movement known as QAnon, a process that has been dramatically accelerated by the COVID-19 pandemic. QAnon is a conspiracy theory that alleges that President Trump is waging a secret war against a cabal of powerful Satanic paedophiles, alleged to be kidnapping, torturing and even cannibalising children on a vast scale. The theory has developed beyond its roots in the intensely hyper-partisan and US-centric right, moving from a niche far-right interest that we have termed orthodox QAnon into a broader, less uniform type we call eclectic QAnon. This development has enabled the theory to gain supporters from across the political spectrum and of diverse backgrounds. As it stands today it is a decentralised, grand and multifaceted phenomenon, at once a conspiracy theory, political movement and quasi-religion, with variants tailored to chime with different subcultures and national contexts.

Its central narrative subverts legitimate concerns about child trafficking and child abuse with fantastical misinformation and antisemitic tropes, fostering a dangerous anger in the process.

Whilst it is important not to overstate the threat of QAnon in Europe, which remains marginal, there are reasons to be concerned about its further spread. Antisemitic tropes are inherent to the theory, and there is scope for the far-right exploitation of the developing UK scene due to significant overlapping narratives. QAnon, which groundlessly alleges that countless authority figures are Satanic paedophiles, has the potential to sow an intense distrust in institutions, including healthcare authorities in the midst of a global pandemic. The theory also risks obscuring genuine child abuse and hampering legitimate efforts to better child welfare. Moreover, whilst it is impossible to know exactly how seriously QAnon followers take their beliefs, and when they will act on them, the highly emotive narratives at the core of QAnon have the potential to inspire individuals towards disruption, harassment and even violence.

Until early in 2020, QAnon was a largely unknown phenomenon outside of the US, and even within it. While some European individuals and groups had been promoting the theory since its earliest days, they were largely looking in from the outside at an explicitly US-centric phenomenon and a narrative with little applicability to the politics of their own nations. While international conspiracies have always formed part of the narrative of QAnon, with the Rothschild family, the House of Saud and George Soros all identified as part of an all-powerful global Satanic elite, the primary narratives have always been centred on the machinations

and minutiae of political developments in Washington DC. However, it was in 2020 that QAnon truly began to spread and take root across Europe, adapting itself to local contexts and interacting with culturally-specific reference points rather than existing as a foreign import. In August, academic researcher Marc-Andre Argentino used a set of criteria to define whether a country had an independent QAnon presence, such as whether it had a specific national QAnon Facebook group and whether local influencers were applying the narrative to domestic issues. He identified such a presence in almost every country in Europe, with only Estonia, Montenegro and Albania being without a movement of their own by early August. Some countries appear to have a significantly larger presence than others when accounting for population size. Lithuania has a dedicated QAnon facebook group with 7,300 members, a remarkable number for a country with just 2.7 million inhabitants. This high engagement has been boosted by the endorsement of prominent figures such as the psychotherapist and owner of the Minfo. It news website Marius Gabrilavičius, who has written numerous articles promoting QAnon on his platform. One of the largest QAnon movements in Europe is that of Germany. The German-language QlobalChange network has 106,000 subscribers to its YouTube channel and a remarkable 122,000 subscribers to its Telegram channel, a huge number of users for that platform and a huge spike from the 20,000 subscribers it had in February. The vast majority of Qlobal-Change's output is translations of videos from popular US QAnon influencers, with no Germany-specific content. The largest pan-European QAnon group was QAnon Europa, which had 20,000 members prior to its removal by Facebook in August 2020. The group was set up by German-speakers and the vast majority of the content was in German, although an accompanying website set up in July now also has content in Russian, Spanish, Italian, Polish, Greek, English, French and Thai.

It is important to understand that while QAnon is not a solely right-wing phenomenon, drawing supporters from across the political spectrum, it has developed pockets of support among the European radical and far-right. Whilst the spread of the theory has so far largely been limited to an individual rather than organisational basis, QAnon has found proponents among a handful of influential online figures, and its narratives are beginning to take hold in far-right Facebook groups and street movements. The significant areas of crossover between the QAnon worldview and pre-existing far-right conspiracy theories and populist narratives has facilitated this spread, and provides opportunities for further cross-pollination.

There are, however, significant shared narratives and concerns that have facilitated the intermingling of QAnon and the European far-right. Conspiracy theories and populism both employ a binary worldview that divides societies between corrupt or evil elites and the pure or unknowing people, a framework that contextualises fears and hardships by personifying them into an identifiable enemy. Right-wing rhetoric has exploited the deep political and cultural divides, and an intense distrust of London-centric political and media “elites”, as well as shadowy “globalists” in the European Union. The turmoil of the COVID-19 pandemic and subsequent government measures has exacerbated this preexisting distrust, and has facilitated an explosion of anti-lockdown, anti-5G and anti-vaccine conspiracy theorising, which has proved popular, as we have reported elsewhere, amongst sections of the far-right.

Also, belief in one conspiracy theory signifies an openness to others. In some ways QAnon is particularly well suited for adoption by right-wing reactionaries, who present themselves as chivalrous “protectors” of the nation and the family, and so have long stoked fears about rapacious - and, in recent decades, south Asian and Muslim - child abusers preying on white children. Children play a symbolic role in nationalist discourse, representing the innocence of the nation as a whole, and so invoking a threat to children is an effective way of mobilising support against a group of people. From age old antisemitic myths, to the exploitation of the grooming gang scandals, such discourse reflects both genuine fears but also a cynical political tactic; presenting an enemy as child molesters, murderers and, at the most conspiratorial end, cannibals is the most effective and unequivocal way to demonise them.

To perhaps a greater degree than any comparable movement, QAnon is a product of the social media era. Aside from the occasional QAnon placards that could be seen at Donald Trump rallies and the emergence in late Summer 2020 of anti-lockdown and #SaveOurChildren street protests, this ideological movement could rarely be seen outside of its home on social media platforms and web forums. Q's reach would have remained fringe, however, if it was limited to 4chan and 8chan. It was the movement's spread onto the mainstream social media platforms - and from there onto the streets - that made this phenomenon into a global concern, one that could do long term damage to the US political environment and an unknown potential for similar harm around the world.

Whatever the future of the core US-centric QAnon narratives, it seems clear that the imported themes will continue to impact on the conspiracy theory milieu across Europe. The extent to which QAnon can be adapted to new national contexts will impact on its ability to implant itself in new locations, but could also lead to utterly distinct variants emerging that can no longer usefully be classified as belonging to the wider movement.

### Asian and Chinese people under fire

Another effect of the spread of the pandemic across Europe has been an upturn in anti-Asian racism and an increased focus on China and the Chinese diaspora by the organised far-right.

Since the pandemic has started there have been increased accounts of anti-Asian assaults, harassment and hate crimes across the globe. This includes verbal aggressions of “go back to China” and “bringing in the virus” to more physical assaults on victims assumed to be Chinese, or even just Asian. The European Union’s Agency for Fundamental Rights has noted a general spike in hate against people of Chinese or Asian descent across the states. The hate has also impacted their access to health services.<sup>21</sup>

In the UK, which has a significant population of Asian origin, figures show that attacks against “Orientals” recorded by the Metropolitan Police rose steeply as the pandemic spread, fell during the lockdown and then, since the easing in May of restrictions, has started to steadily rise again.<sup>22</sup> In February 2020 Dr Michael Ng of a Chinese Association in the UK told the Guardian that hostility against the Chinese community was at the worst level he had seen in 24 years.<sup>23</sup> This hostility translated into hate crimes, and in May it was announced that hate crime directed at south and east Asian communities had increased by 21% during the coronavirus crisis.<sup>24</sup> This pattern was seen across large parts of Europe. In France the hashtag #JeNeSuisPasUnVirus (I’m not a virus) was used by French-Asian citizens facing stigma and attacks.<sup>25</sup> This followed the outrage caused when a local newspaper, *Le Courier Picard*, used the headline “Alerte jaune” (Yellow alert) and “Le péril jaune?” (Yellow peril?), complete with an image of a Chinese woman wearing a protective mask.<sup>26</sup>

In Italy, the NGO Lunaria has collected over 50 reports of assaults, discrimination and bullying by people perceived to be Chinese.<sup>27</sup> The barrage of hateful rhetoric can also be traced to politicians and even parties in power. The governor of the Veneto region of Italy, an early epicenter of the pandemic, told journalists in February that the country would be better than China in handling the virus due to

Italians’ “culturally strong attention to hygiene, washing hands, taking showers, whereas we have all seen the Chinese eating mice alive.”<sup>28</sup> Gianni Ruffin, director general of Amnesty International Italy, spoke out on the issue stating, “Scientifically incorrect information, irresponsible affirmations by politicians and incomprehensible local measures [taken against the virus’ spread] have led to a shameful wave of Sinophobia.”<sup>29</sup> Similar rises of anti-Asian racism were seen across the continent. In Sweden and Poland for example, reports of xenophobia and racist attacks against people of Asian descent were reported.<sup>30</sup> While in Hungary, Asian’s of non-Chinese heritage felt the need to make clear they are not Chinese, with at least two shops in Budapest displaying signs reading “Vietnamiak vagyunk”, meaning “We are Vietnamese”.<sup>31</sup>

However, the hate experienced by Asians due to COVID-19 does not exist in a political vacuum. It would be wrong to say that the upswing in anti-Asian racism this past year has been the result of the far-right alone, not least because there is a broader societal racism at play. However, the far-right has certainly sought to exploit this societal prejudice and in some cases, exacerbate it. However, the emergence of COVID-19 coincidentally coincided with a broader shift towards anti-Chinese politics by the international far-right. Over the last few years, as we have seen the ‘decoupling’ of the US and Chinese economies and a shift towards what some are calling a ‘cold war chill’ between the West and China, nationalist and far-right figures have increasingly targeted China and Chinese people.

Of course, much of the criticism China faces is well deserved. It is an authoritarian state with an abysmal human rights track record, especially in relation to its appalling treatment of the Uyghurs. Many have also rightly complained about the widespread theft of intellectual property that makes international trade ‘unfair’. When it comes to the coronavirus outbreak there are also many questions still to be answered concerning China’s early obfuscation and intimidation of those speaking out. No doubt this will be investigated more thoroughly in time. However, while criticism of the Chinese government is warranted, the continuing development of a new ‘cold war’ is having serious consequences for Chinese and Asians living in Europe and will only be exacerbated as political tensions increase. It will also result in some people blaming Chinese people for anything and everything. This can already be seen with articles like Douglas Murray’s in *The Sun* where he suggests China released COVID-19 on purpose to attack the US economy.<sup>32</sup>

One trend in this direction that is of interest, and could outlast the current pandemic, is prominent far-right figures within the international anti-Muslim movement increasingly targeting China and Chinese people. Since 9/11, and the series of subsequent Islamist attacks in Europe, sections of the far-right have framed their politics as defence of national security in the face of an Islamic threat. For some, the external threat they highlight has now been expanded to include China, which will no doubt also frame Europeans of Chinese descent as a possible fifth column. This process is well developed in America with major anti-Muslim organisations such as Act for America already pivoting towards China, but the same is starting amongst elements of the European far-right. One of the best examples of this so far is the prominent European far-right figure Stephen Yaxley Lennon (aka Tommy Robinson) who has increasingly targeted China, sometimes using racist Chinese caricatures.<sup>33</sup>

The combination of a longer-term trend of anti-Chinese rhetoric compounded by COVID-19-related politics, means we are likely to see a prolonged period of the European far-right focusing more on China and anti-Chinese racism.

---

## SECTION 2

### **The Racist Backlash Against the Black Lives Matter Movement**

The brutal murder of George Floyd by a Minnesota police officer sparked a global response, galvanising a long-brewing resentment and anger at deep-rooted and systemic racism, as well as broader societal anti-Blackness and white supremacy. Inspired by the demonstrations across America, people have taken to the streets across Europe to show solidarity and raise awareness about racial injustice closer to home. Thousands gathered in Paris, London, Berlin and Amsterdam, amongst others, to join in the chants of 'I can't Breathe'. However, like everyone else, the European far-right have followed events in the US closely, seeking to exploit them for their own domestic gain and provide international support to Donald Trump and the US far-right more generally. While the proliferation of continent-wide discussions about race, colonialism and imperial legacies has been a welcome one, it has also been seized upon by elements of the European far-right as an opportunity to talk about race in a more exclusionary and supremacist manner. This has happened in two ways. Firstly, existing racial nationalist activists and organisations, already preoccupied with the concept of race, have used the BLM protests to push their existing political platform to a wider audience. Secondly, some elements of the far-right that had traditionally distanced themselves from open racial politics, promoting instead 'cultural nationalism', have become more willing and open to explicitly racial politics. Whether this shift is permanent will remain to be seen but in the short to medium term we are likely to continue to see cultural nationalism cede ground to racial nationalism within the far-right.

The most obvious manifestation of this phenomenon has been the emergence and spread of the 'White Lives Matter' slogan in response to BLM. First emerging in the US in 2015, it is only really this year that it has been popularised amongst the European far-right.<sup>34</sup> Decontextualized, the slogan is inoffensive and comparable with 'Black Lives Matter'. In context it represents a negation of the structural and systemic racism implicit in the need to highlight the value of non-white lives. It allows the far-right to push a racist agenda via the use of an indisputably true statement, namely that white lives do indeed matter.

The requirement of explanation and context when opposing the use of 'White Lives Matter' is its major advantage for the far-right. For people who understand racism as something that only occurs when there is direct intent, they are more likely to personalise the issue and get defensive. Where there is cognitive dissonance on people's understanding of historical racism's bearing on systemic discrimination today, it is also easier for people to distance themselves from the problems at hand and thus make them more likely to see nothing wrong with the use of the slogan White Lives Matter. However, while some people genuinely but mistakenly believe that BLM movement is being dismissive of white lives, many on the far-right are wilfully misunderstanding the issue for political gain.

In the UK, the slogan has been adopted widely by the domestic far-right. The anti-Muslim organisation Britain First, for example, released numerous images of Lee Rigby, Emily Jones and Charlene Downes - all white murder victims - with text overlaid reading 'White Lives Matter'. The hashtag #WhiteLivesMatter has also trended in the UK, though admittedly much of the traffic is in condemnation of its use. Similarly, the name of Lee Rigby, the British soldier murdered by al-Muhajiroun activists on the streets of London, also began to trend on Twitter. Many on the far-right have sought to draw false equivalency between the two tragedies. Katie Hopkins for example tweeted, 'Outrage. Available in any colour, As long as it is black #leerigby'. For some, this more open discussion of race was something of a departure. Prominent figures and groups such as Stephen Yaxley-Lennon (AKA Tommy Robinson) and Britain First, known primarily for their Islamophobia, switched their focus to race as part of broader plans to 'defend' various statues and memorials, in response to protests about their links to slavery and colonialism. When a Burnley FC supporter was condemned for organising a plane to fly the 'White Lives Matter' slogan over Manchester City stadium, Lennon likewise lent his support. While the likes of Lennon and Britain First were far from moderate in their view prior to this, such a move is clearly worrying to the extent it can normalise more extreme far-right ideas in such a socially divided time.

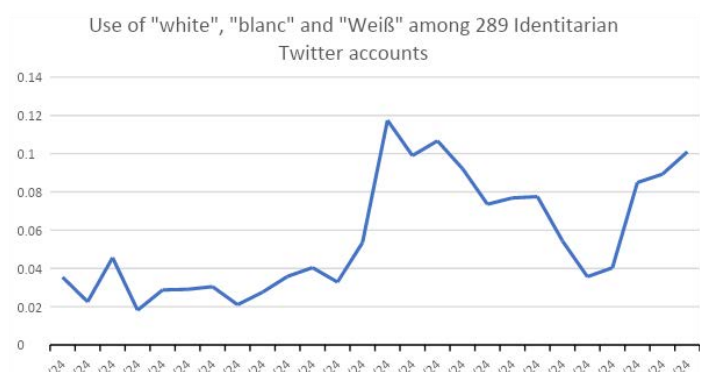
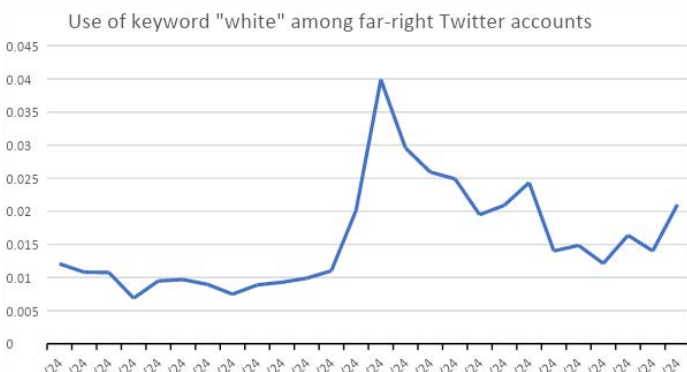
The most sustained use of the slogan White Lives Matter in the UK has come from a new racial nationalist organisation called Patriotic Alternative.<sup>35</sup> Formed in 2019 by Mark Collett, former Head of Publicity for the British National Party, the group has quickly grown to a following of nearly 18,000 on Facebook.<sup>36</sup> PA is a racist far-right organisation with antisemitism at its very core. They aim to combat the "replacement and displacement" of white Britons by people who "have no right to these lands". In this regard PA follows the broader trend in recent years amongst many in the far-right of rebranding white nationalist ideology as a defense of 'indigenous' Europeans against their 'Great Replacement' from non Europeans. On 9 August Patriotic Alternative (PA) held a day of action across the UK to coincide with International Indigenous People's Day (IPD). The event involved repeating, at a national scale, a strategy the group employed on 4 July when they displayed a 'White Lives Matter' banner on the top of Mam Tor, a hill in Derbyshire. The image of the banner atop Mam Tor was intended to stir up controversy and in so doing bait the media and concerned members of the public into giving the marginal group free publicity. Though press coverage was only local, the event attracted attention on social media and was successful in bringing in new supporters to PA. Due to this success they decided to hold the much larger event on IPD. The result was images of roughly 80 locations displaying the slogan, alongside related phrases, from just over 100 activists. There were also a handful of pictures submitted from abroad, including by the fascist groups Nordic Resistance Movement in Denmark and Action Zealandia in New Zealand.

Similar stunts using the White Lives Matter slogan have been seen across the continent in 2020 with reports of banners being unfurled at football games in The Czech Republic, Ukraine, Hungary and the Netherlands. One report by DW showed how "Monkey chants, a Confederate flag, "White Lives Matter" banners and even a call for the release of the policeman charged with the death of George Floyd have all been seen at football grounds in Europe over the past month."<sup>37</sup>



However, one of the most concerted and high profile campaigns in reaction to the BLM movement this year has come from the Identitarian movement across the continent. The international Identitarian movement started in France with the launch of Génération Identitaire (Generation Identity, or GI), the youth wing of the far-right Bloc Identitaire. It has since spread across the continent with affiliated groups, the most prominent of which, in addition to France, are based in Germany, Italy and Austria. At the core of identitarianism is the racist idea of ethnic-separatism which they call ‘ethnopluralism’. Similarly, they also call for ‘remigration’, a coded term for the idea of repatriation of non-white people. Part of the movement’s success has been their ability to take extreme ideas and present them in a way that sounds moderate. They affect public attitudes by promoting a lexicon which, for those unfamiliar with the contemporary far-right, may have less obvious links to extreme, prejudicial and dangerous political ideas and policies. It is for this reason that they have pounced on the White Lives Matter slogan so enthusiastically this year. In June for example, GI activists in France held an anti-BLM counter protest and unfurled a huge banner reading “Justice for the victims of anti-white racism: #WhiteLivesMatters”.<sup>38</sup> Similarly, in Germany, GI activists sought to capitalise on a series of large BLM demonstrations across the country by launching a campaign titled #NiemalsaufKnie (Never on our knees) in response to protestors and politicians kneeling in solidarity with the victims of racial violence.<sup>39</sup>

The increased prevalence of more explicit racial politics and rhetoric is not merely anecdotal. Based on keyword matching in the tweets posted by far-right accounts monitored by HOPE not hate, we observed a notable increase in tweets discussing race during the week of George Floyd’s death, a period that also became a flashpoint in the BLM movement. His death took place on the 25th of May, a Monday. That week and the following week, adjusted for total weekly tweet volume, tweets mentioning the keyword “white” increased fourfold compared to the previous two months. Specifically looking at a set of 289 accounts being part of the European Identitarian movement in mainly the UK, France, Germany and Austria, the same pattern was observed. Although the movement more frequently used the keyword “white” (and it’s French and German counterpart) than the average far-right account overall, the week of Floyd’s death saw the amount of discussion increase by approximately 370%. In both the case of identitarian accounts as well as the whole sample of far-right accounts the relative amount of tweets matching the keywords remained elevated until August 31st, the end of the period measured.



The re-racialisation of the far-right has been notably evident within the UK, though similar tactics have been observed across the European far-right. By using the international discussion of racial injustice that has been spawned by the events in America, the European far-right has worked to deny or downplay the scale and uniqueness of anti-black oppression across Europe and promote their longstanding belief that the true victims of societal racism are actually white people at the hands of multicultural and politically correct elites. Egregiously, many have increasingly sought to co-opt the language of human rights and oppression, with some even publicly identifying with figures such as Martin Luther King, Ghandi or Mandela. More generally though the European far-right has seized the BLM moment this summer and sought to mirror its success and co-opt the claim of being a persecuted minority. Here we see a rhetorical gymnastics that frames far-right activism as a struggle for human rights and equality, shorn of overtly racist or crude epithets. This tactic provides a serious challenge to those opposing the far-right or seeking to moderate their activity on social media as the lexicon ostensibly appears progressive thereby requiring increased levels of context to reveal the reality of the prejudiced politics on display.

---

## CONCLUSIONS

It is important to note that all these trends are so far just indications of a direction of travel within the European far-right. Whether it be a growth in anti-Asian racism, more explicit racial politics or the expansion and exploitation of conspiracy theory growth, it remains too early to say if these will be short term or lasting changes. The spate of recent terrorist attacks in France for example, might contribute to a renewed concentration on Islam and Muslims. It is also the case that both the pandemic and the Black Lives Matter movement are still happening and thus the changes within the far-right will continue to be affected by ongoing events. However, the direction shifts highlighted in this article are, we believe, significant enough to have a period of longevity, meaning that they must be considered when developing and planning strategies to mitigate the harmful effects of the far-right in the coming years, be that on social media or offline.

## CITATIONS

# The European Far-Right

- 1 Jean-Yves Camus and Nicolas Lebourg, *Far-Right Politics in Europe* (Harvard University Press, 2017) 21-22
- 2 Cas Mudde, *The Far Right Today* (Cambridge: Polity, 2019), 7
- 3 <https://www.statista.com/statistics/1102288/coronavirus-deaths-development-europe/>
- 4 <http://www.gdr-elsj.eu/wp-content/uploads/2014/02/doc2-WP4-Del-7.pdf>
- 5 <https://www.nature.com/articles/s41562-020-0884-z>
- 6 <https://www.irishtimes.com/news/crime-and-law/garda-concerned-over-increasing-participation-of-far-right-in-anti-mask-protests-1.4358684>
- 7 <https://www.dailymail.co.uk/news/article-8876275/Riots-Rome-Protestors-clash-police-Italian-capital-second-night-demonstrations.html>
- 8 <https://jean-jaures.org/nos-productions/bas-les-masques-sociologie-des-militants-anti-masques>
- 9 <https://www.theweek.co.uk/105909/how-populists-are-exploiting-the-spread-of-coronavirus>
- 10 <https://twitter.com/antonioguterres/status/1258613180030431233?s=20>
- 11 <https://balkaninsight.com/2020/06/22/serbia-under-reported-covid-19-deaths-and-infections-data-shows/>
- 12 <https://www.isdgglobal.org/wp-content/uploads/2020/05/20200513-ISDG-Weekly-Briefing-3b.pdf>
- 13 <https://www.cmu.edu/ideas-social-cybersecurity/research/coronavirus.html>
- 14 <https://www.lefigaro.fr/politique/coronavirus-marine-le-pen-trouve-legitime-de-se-demander-si-le-virus-ne-s-est-pas-echappe-d-un-laboratoire-20200330>
- 15 <https://www.valeursactuelles.com/clubvaleurs/politique/interview-marine-le-pen-le-gouvernement-est-le-plus-gros-pourvoyeur-de-fake-news-depuis-le-debut-de-cette-crise-117518>
- 16 [https://www.abc.es/espana/abci-mitad-ciudadanos-apuntan-teoria-conspiracion-sobre-origen-virus-202004270220\\_noticia.html#vca=mod-lo-mas-p5&vmc=leido&vso=espana&vli=noticia.foto.espana&vtm\\_loMas=si](https://www.abc.es/espana/abci-mitad-ciudadanos-apuntan-teoria-conspiracion-sobre-origen-virus-202004270220_noticia.html#vca=mod-lo-mas-p5&vmc=leido&vso=espana&vli=noticia.foto.espana&vtm_loMas=si)
- 17 <https://www.thetimes.co.uk/edition/world/curse-of-the-bat-woman-what-went-on-in-wuhan-lab-pxmws0pzt>
- 18 <https://www.nytimes.com/2020/02/17/business/media/coronavirus-tom-cotton-china.html>
- 19 [https://www.youtube.com/watch?v=\\_5wn1qs\\_bBk](https://www.youtube.com/watch?v=_5wn1qs_bBk)
- 20 <https://allianceforscience.cornell.edu/blog/2020/03/covid-crisis-captivates-conspiracy-theorists/>
- 21 <https://www.euractiv.com/section/global-europe/news/covid-19-crisis-triggers-eu-racism-against-asians-rights-agency-says/>
- 22 <https://www.theguardian.com/society/2020/aug/29/far-right-using-coronavirus-as-excuse-to-attack-chinese-and-south-east-asians>
- 23 <https://www.theguardian.com/world/2020/feb/18/chinese-people-uk-targeted-racist-abuse-over-coronavirus-southampton>
- 24 <https://www.theguardian.com/world/2020/may/13/anti-asian-hate-crimes-up-21-in-uk-during-coronavirus-crisis>
- 25 <https://www.euronews.com/2020/02/03/coronavirus-france-faces-epidemic-of-anti-asian-racism>
- 26 <https://www.bbc.co.uk/news/world-europe-51294305>
- 27 <https://www.lunaria.org/>
- 28 <https://www.hrw.org/news/2020/05/12/covid-19-fueling-anti-asian-racism-and-xenophobia-worldwide>
- 29 <https://www.aljazeera.com/news/2020/02/18/coronavirus-prompts-hysterical-shameful-sinophobia-in-italy/>
- 30 <http://www.forintegration.eu/pl/coronavirus-related-incidents-of-xenophobia-in-poland> and <https://www.hrw.org/news/2020/04/06/abused-and-shunned-being-asian-descent-sweden-during-covid-19>
- 31 <https://www.ft.com/content/eeda65ea-4424-11ea-a43a-c4b328d9061c>
- 32 <https://www.thesun.co.uk/news/11531209/douglas-murray-we-need-to-make-china-pay/>
- 33 <https://shop.tr.news/product/official-china-is-asshoe-hoodie/>
- 34 <https://www.splcenter.org/fighting-hate/extremist-files/group/white-lives-matter>
- 35 [https://www.hopenothate.org.uk/wp-content/uploads/2020/08/HnH\\_Patriotic-Alternative-report\\_2020-08-v3.pdf](https://www.hopenothate.org.uk/wp-content/uploads/2020/08/HnH_Patriotic-Alternative-report_2020-08-v3.pdf)
- 36 <https://www.facebook.com/PatrioticAlternative>
- 37 <https://www.dw.com/en/white-lives-matter-banners-take-racism-to-a-new-level-after-burnley-plane-stunt/a-53916597>
- 38 <https://www.ruptly.tv/en/videos/20200613-028-France--Counter-protesters-unfurl--White-Lives-Matter--banner-at-Paris-anti-racism-demo>
- 39 <https://gnet-research.org/2020/06/29/niemalsaufknieen-how-the-identitarian-movement-in-germany-reacted-to-the-black-lives-matter-protests/>

---

RESEARCH BRIEFINGS

---

## 02 Recent Trends and Perspectives: COVID-19 Pandemic Increases Right-Wing Extremism and Hatred in Europe

---

by Matthias Quent, Christoph Richter  
and Axel Salheiser

EXTRACTS FROM:

Jena Institute for Democracy  
and Civil Society

---

---

## EXECUTIVE SUMMARY

The COVID-19 pandemic shows how fragile our social communities have become. It clearly reveals Europe's areas of conflict and social tensions. In the context of these divisions, this article describes the social dynamics as reactions and counter-reactions to social change. Digital transformation has accelerated as a result of this crisis, presenting both a challenge and an opportunity. The focus here is on digital 'takeover' strategies by the far-right during the pandemic and the consequences.

The process of digitalisation influences social debate about equal opportunities and social participation, and can strengthen democratic values. At the same time, there is an increasing danger of polarisation and radicalisation. International right-wing extremism is currently the most dangerous threat. It has become a modernised movement in the digital space and has organised itself into a flexible, wide-ranging and work-sharing network. The COVID-19 pandemic has been accompanied by a 'pandemic' of hate speech and disinformation giving right-wing extremists greater reach on social media. This goes hand in hand with the risks to more vulnerable users. Right-wing extremism and conspiracy theories and narratives are coming together and being reframed on the internet. An example of this is the growing 'QAnon' movement. Fears, insecurities and antagonism related to COVID-19 are being used, especially on the internet, as a door-opener to radicalise larger groups of the population. This also increases the risk of right-wing violence, hate crime and terrorism in Europe.

## Conclusions

Possible courses of action include the consistent and unified prosecution of hate speech and hate violence by platforms and state institutions, as well as the strengthening of democratic online culture. Sustainable solutions require proactive strategies to help shape the digital space as a democratic platform for the common good - involving politicians, lawyers, academics, businesses, security agencies, civil society and marginalised groups.

---

## OVERVIEW

Socioeconomic inequality, educational injustice, lack of health care, addictive behaviour, racism, misinformation and a conspiracy mindset: The COVID-19 pandemic and its psychological, social, political and economic consequences is exacerbating various social and political grievances and conflicts worldwide. At the same time, the pandemic is accelerating digitalisation which makes it possible to counterbalance some of these shortcomings: Millions of people in Europe are discovering new opportunities for working digitally and communicating in this time of need. At the same time, however, large sections of the population who work in factories, workshops, nursing, police departments and fire brigades cannot avoid physical proximity to other people. Social divisions and conflict are intensifying. Digitalisation trends can have both positive and negative effects. As if through a magnifying glass, the pandemic reveals and magnifies social change and social problems, while the boundary between the analogue and digital worlds becomes increasingly permeable. This applies to everyday situations, for example when professional and private roles need to be reconciled when working from home. But it also applies to extreme situations, for example when people, motivated by online-induced hatred and conspiracy ideologies, make electoral decisions or even commit acts of violence against other people and infrastructures. This global crisis has revealed emerging trends and developments that will outlast the pandemic. What can we learn from this situation for our future?

This report provides a comprehensive summary of recent movements, processes and phenomena that challenge Europe's democratic culture and the principle of equality for everyone. Special focus is given to the mechanisms and strategies of parties who use prejudice, hatred, conspiracy ideologies and anti-Semitic, racist or ethnocentric agitation in social networks to act as a link between analogue and digital spaces - exploiting the COVID-19 pandemic for scapegoat politics and, in turn, threatening societal cohesion. Based on this, certain recommendations for action are proposed for politics, social networks and civil society.

---

## DIGITALISED RIGHT-WING EXTREMISM

Social media offers dual possibilities: it can be both a challenge and an opportunity. In recent years, following the digitalisation of right-wing extremism, at least part of the 'dark side' of social media has diversified, professionalised and expanded. The dynamics of attention on social media encourage the spread of shortened and often extreme content which can lead to social closure within radical political environments. The far-right benefits from a combination of technological infrastructure and emotionally-charged mobilisation strategies: The reason for this is that extreme right-wing communication strategies - often based on populist oversimplification, provocation and evocation of apocalyptic threat scenarios - are supported by the functional logic of algorithms, which spread dramatic content more effectively. Through the algorithm's propagation of such pages and media content, more 'casual' consumers are being reached enabling the far right to normalise its ideology with broad impact. In addition, technology can be used to give the impression of far greater digital ratios than in reality (e.g. through the creation of fake accounts, manipulation of membership numbers, strategic use of bots, and coordinated hate campaigns). Social media is a contributing factor that reinforces the emergence of 'echo chambers' for the distribution, amplification and popularisation of extreme right-wing content. Compared to other political groups, the tendency to refer only to one's own political sphere of resonance is four times higher in the far-right spectrum.

With the increasing anonymity of social media communications, the spread of extreme content can potentially gain even more momentum. The multitude of applications and social messaging services and other media tools are used in different ways. This leads to a widespread network of right-wing populist and extreme activists and groups, pursuing different goals and serving different purposes, and sometimes achieving massive reach. Digital media serves and exposes the self-expression of activists more than in earlier forms of right-wing extremism. It mixes private and non-political, partly subcultural content with political issues and financial interests. In some cases, it achieves much wider reach, even into other political spheres. Group chats and group channels are used for symbolic and political self-affirmation together with coordinating, networking and preparing of a range of actions. Another key role is played by 'dark' social services, such as Telegram and WhatsApp. Telegram in particular is a popular communication platform for right-wing extremists and other anti-democrats, because the absence of moderation by platform owners makes repressive measures against users unlikely.

Numerous militant and right-wing terrorist groups used and still use open but also partially encrypted chats for recruitment, communication and coordination.

Following the attacks in Pittsburgh, Christchurch, El Paso and Halle, these new digital forms of communication and organisation for right-wing extremism finally came under scrutiny. The attacks shed light on how image boards (e.g. 4Chan, 8Chan), alt-tech platforms (e.g. Gab) and gaming platforms (e.g. Steam, partly Discord) serve right-wing extremists globally and on a daily basis as places to disseminate far-right content. The language, symbolism and ideology of contempt shared on some image boards and platforms can fuel calls to action. It guarantees terrorists the attention and applause of a global online community. Overall, while the content remains the same, right-wing extremism has modernised itself as 'digital fascism' with the help of digital media. It has formed a transnationally interconnected, decentralised, and multi-layered network with massive reach. The target group-oriented, cross-platform and task-sharing presence of these networks, makes them relatively resistant to selective repression measures such as deplatforming.

### Extreme right-wing (sub)cultures

In October 2020, right-wing extremists streamed a martial arts event with participants from 10 different countries: This was supposed to be a replacement for the originally planned major event 'Battle of the Nibelungs', which had to be cancelled due to increasingly repressive measures by the German authorities due to pandemic-related restrictions. This shows how digital fallback strategies are used to maintain subcultural activities, and develop them further. Extreme right-wing (sub)culture both offline and online spans a sphere of art, entertainment and lifestyle that provides identity- and community-building functions for its members. Subcultural participation and interaction encourage the development and consolidation of anti-democratic, group-related misanthropic attitudes and behavioural orientations. Cultivating content in the areas of music, media, martial arts, gaming, nationalist and conservative customs, etc. is strategically important for radical and extreme right-wing parties to achieve a 'pre-political' influence in all areas of society. They are a reflection of ideological and social cohesion as well as providing institutional and organisational continuity for radicalised communities. However, this is not only about aspects of political propaganda or ideological indoctrination, mobilisation, socialisation and the perpetuation of one's own values and norms. It is also a lucrative source of financial income for earning a livelihood and for political

activity. Paradoxically, the necessary public outcry and partial criminalisation of this phenomenon makes it highly attractive, especially to young consumers and users (the desire to provoke by breaking taboos). Radical and extreme right-wing net culture serves partly as a surrogate for numerous offline activities that are (increasingly) tabooed, restrictively sanctioned or criminalised in democratic contemporary societies and are restricted or prevented by democratic involvement: The internet offers (temporary) 'retreat spaces' or 'rallying points' for authoritarian, regressive and reactionary socio-cultural environments or movements, in which they have reorganised themselves and from where, now strengthened, they are again reaching out into other areas of society. Their common goal remains an anti-modernist 'cultural revolution from the right'. Digital hate cultures and radical or extreme right attempts to achieve discursive hegemony in some parts of the internet - on websites, in forums, messenger services and image boards - are an important part of this 'culture war', which is fought with sharepics, memes, music and videos among other things.

### **Martial arts**

Important transnational current trends in Europe have intensified far-right activities in such fields as martial arts and gaming. Mixed martial arts and kickboxing events, sometimes conducted covertly and with international participation and which are streamed on the internet (during the COVID-19 pandemic), are especially attractive to men. The cult of the body and violence as a central element of fascist ideology and Nazi ideology is mixed here with (youthful) 'sensation seeking', the general trends in sport and fitness in society and the goal of being able to use one's own body as a 'weapon' against people who correspond to radical and extreme right-wing images of the enemy (e.g. migrants, racially discriminated people, and political opponents). Militant right-wing extremists with an affinity for terrorism learn mixed martial arts in addition to firearms training, for example in training camps in Eastern Europe, which are attended by people from all over Europe and the USA.

### **Right-wing gaming subculture**

Computer games with radical and extreme right-wing, historically revisionist, racist, anti-Semitic and other anti-democratic content are intended to normalise, or rather encourage gamers to adopt hate ideologies. A current example from Germany is the game 'Homeland Defender' which is financed by the right-wing extremist hate group 'Ein Prozent für unser Land' (One percent for our country). The game is used to promote the ideology of the 'Identitarian movement'. Gaming and related communication in chat rooms and forums, on image boards and gaming platforms (Steam etc.) are also considered a recruitment strategy of extreme groups or rather a (self-) selection mechanism that is relevant for the radicalisation process of right-wing terrorist assailants (e.g. Munich 2016, Aztec 2016, Christchurch 2019, El Paso 2019, Halle/Saale 2019). Regarding the actual execution of the crime, its online staging and the role model character for potential imitators, this has been referred to as the 'gamification' of right-wing terrorism. The connection between right-wing terrorism and gaming goes back to an 'amorphous network' of different online subcultures. It originates in the 'Manosphere' (an online culture inspired by men's rights activists), the Incel movement (a misogynistic men's movement), radical gaming communities (known through 'Gamergate') and various other boards as communication platforms (including 4chan, 8Kun, among others). Effective expressive stylistic tools (primarily the use of so-called memes or sharepics), infused with provocation, cynicism and taboo-breaking as a humorous principle, were combined early on with narratives and feelings of cultural and sexual discrimination, especially of young and frustrated white men. A toxic mix that opened up a suitable resonance space for the far-right, tapping into the widespread tendency to break taboos as well as propagate misogynistic hatred, victimisation and the conspiracy mentality of a white subculture. Protagonists of the 'alt-right' movement successfully used these factors in a struggle for cultural hegemony, which enabled them to connect with society and gain access to a young target group. Since then, at least some sections of these boards and gaming platforms have become a protected and anonymous space for hateful, misanthropic, sexist, racist and anti-Semitic writing and imagery.

## Hate music

The extreme right-wing music subculture is stylistically diverse and international. It breaks down into sub-scenes that can be considerably different. This applies to the artistic methods they use as well as to their performance and the appearance of their followers. Radical and extreme right-wing parts of the rap subculture represent a relatively new and still little-noticed development, which is closely linked to the 'Identitarian Movement', among others, and tries to make 'provocative', 'fresh' or 'hip'-looking identification offers to teenagers and young adults. Recordings and merchandise from 'right-wing rock' are mainly sold on the internet. Streaming services, download portals and video platforms are also used for audio-visual content. YouTube's recommendation system and its Like mechanic can lead users further and deeper into the world of right-wing hate music. Despite some ideological differences, there is a great amount of international cooperation in today's 'right-wing rock' across stylistic and national borders, e.g. in production and distribution processes and in events or performances. Almost the entire spectrum of blacklisted hate music is available without major barriers via platforms and social media channels.

## Right-wing terrorism

Trends cannot be understood in isolation from the history of phenomena. This also applies to the international increase in hate crime and right-wing terrorism as well to the strengthening of a modernised identitarian right-wing extremism in the last decade. A milestone for these developments was the extreme right-wing-motivated attacks in Oslo and Utøya on the 22nd of July 2011 which resulted in 77 fatalities. Many subsequent killers have referenced the Norwegian Breivik, who in turn drew his political ideology and identity, as well as elements of the conspiratorial strategy of violence, primarily from the international Islamophobic blogosphere. While terrorist violence usually harms the political causes of a movement rather than benefits them in the long run, this was not the case with the Breivik attack. No other right-wing terrorist attack has led to the normalisation and spread of modernised anti-democratic and anti-human concepts, symbols, identitarian self-understandings and programmes as much as this mass murder. The right-wing terrorist attacks on the 22nd of July 2011 in Oslo and Utøya with 77 fatalities marked the beginning of a global series of fatal incidents, characterised in particular by the transnational radicalisation and networking of the attackers via the internet. The global series of attacks ties together previous far-right terror campaigns but also differs from them:

- Terrorists find their positive affirmation communities in the first place through the internet and address them there directly. Through acts of violence, the subjective realities and narratives of these communities find their way into the mass media and the political public in society as a whole. Hate crime, in the form of messages, is a key medium by which inhuman beliefs and communities work their way from the internet into the attention of an even larger audience.
- In radicalised online environments within various platforms, hatred towards the negative reference groups is stirred up, while scapegoats and supposedly legitimate victims are collectively marked. In this context, these are not closed and exclusively self-referential spaces, but rather selection and amplification media for overarching social developments and debates.
- The political radicalisation of far-right hate communities (meso level) can be traced online through mostly intentional digital documentation, but is always influenced by overarching socio-political influences, debates and developments (macro level) and by specific personal conditioning and circumstances (micro level).
- At the individual level, it has been observed that some people with mental illness adopt anti-Semitic and racist ideologies and conspiracy narratives, partly to justify massive acts of violence politically and to give 'meaning' to their own actions.
- Within the (online) community, there is a blending and mixing of right-wing extremist elements with elements that did not begin that way (e.g. cynical humour, meme-fication, gamification, and pop culture references). The anti-Semitic and racist terrorist in Halle (Saale), for example, drew from anime culture, weapons forums, right-wing extremist music and ideologies, and used gaming references. Sometimes, seemingly contradictory and highly individualised right-wing extremists' self-made 'mosaic' identities make it hard for security authorities, educators, family members, friends, social media moderators, politicians and media producers to classify them.
- Terrorists often act alone in the actual execution of the crime, but they are not alone in the cognitive and operational preparation. This complicates the early detection of criminal intentions. It also fits with the purpose of conspiratorial individual action described in detail by Breivik. That this approach can be effective from the terrorists' point of view is confirmed by the fact that many right-wing extremist and right-wing terrorist groups and plans have been exposed in recent years through chat groups with compromising content.



Nowadays, politically-related subcultures connect worldwide with the help of social networks and blend with other communities. On the one hand, this normalises right-wing extremist content; on the other hand, it reconfigures the ideology and habits of right-wing extremists. This is how, for example, commonalities and interdependencies between right-wing extremists, vaccination critics (anti-vaxxers) and conspiracy ideologues have emerged and become apparent during the COVID-19 pandemic. These are mutual dependencies in which the environments benefit from the apparent common strength and logistical and conceptual resources of each other, as well as from the provocation principle of breaking taboos - especially in the 'currency' of public attention.

---

## NEW CHALLENGES DUE TO THE COVID-19 PANDEMIC

These processes threaten to intensify under the influence of the COVID-19 pandemic. Disruption, fear of the future, social isolation, loneliness, frustration, psychological problems and alcohol or drug abuse during the pandemic can reinforce the mechanisms of self-empowerment through scapegoating and the willingness and susceptibility to adopt and support radical narratives. Anti-democratic parties try to exploit increased vulnerabilities and rely particularly on the emotionalising and misleading effects of online communication. Among other things, they try to hijack virulent discussions related to the pandemic using hashtags or through groups, placing racist and anti-democratic content and redirecting users to even more radical channels, such as Telegram or YouTube. The increased vulnerability and intensified activities of anti-democratic parties from various social network movements can take advantage of both ideological insecurities and the massive increase in the use of social media around the world during the pandemic. The general rule here is: more internet use leads to more hate messages and more people affected by hate messages. These harmful dynamics are reinforced by the fact that social work, relationship work and deradicalisation work are interrupted or severely disrupted by the physical distancing measures enforced as a result of the pandemic.

In regards to the attempts by radical and extreme right-wing parties to exploit the pandemic, the picture in Europe has so far been ambivalent: On the one hand, right-wing populists and extremists hope to profit from national border closures, far-reaching critiques of globalisation and demands for the re-nationalisation of the ways in which vital goods are produced (e.g. in the social mainstream - protective masks or vaccinations). Across Europe, far-right factions are

propagating a new nationalism as a response to the coronavirus crisis. The question of whether 'coronationalism' or global solidarity will prevail in the political mainstream is also hanging in the balance across the European Union. In many European countries, extreme right-wing and in some cases violent parties have used the crisis and the drastic political countermeasures as a reason to protest and are trying to establish themselves as the parliamentary arm of the corona-deniers and down-players. Conspiracy and populist channels are gaining popularity, especially on and through social networks. The scattered motives gathered in virtual communities that lead to eccentric statements in street protests attract significant public attention. In Berlin, for example, as many as 40,000 people took part in these protests at the end of August 2020.

However, far-right parties in Europe have predominantly not benefited from the crisis in forecasts and elections so far - their popularity has generally declined since the beginning of the pandemic. An online survey conducted in 15 Western European countries, mostly during the period of the first lockdown, shows: more people are once again supporting the current decision-makers and institutions. Support for the ruling parties, trust in governments and satisfaction with democracy have all increased following the various lockdowns. Consequently, this pandemic could benefit governments - rather than (populist) opposition parties. During the COVID-19 crisis, more people (at least so far) have tended to look to the government of the established factions for support, guidance and leadership. This paradox goes hand in hand with dangers of polarisation and radicalisation that are caused by the mechanisms of social networks. These mechanisms can increase both the human need for ideological coherence as well as the selection of information in favour of propaganda and cognitive isolation over other arguments and opinions, and over scientific findings. Added to this: during the pandemic, while there have been no gains for anti-liberal opposition parties in Europe as a whole so far, right-wing governments, such as those in Hungary and Poland, are using the crisis to diminish democratic rights through emergency regulations. Furthermore, the crisis has seen a global rise in discrimination, prejudice, hate speech and hate crimes related to the virus, with human rights being challenged. A survey in the Czech Republic conducted during the lockdown of early summer 2020 showed: hostility towards immigrants during the COVID-19 crisis has increased. So in the long run, it seems the far-right may still benefit from the pandemic. It is therefore even more important that social networks, media and politicians do not reinforce ethnocentric and nationalistic tendencies in the population, but rather implement appropriate measures to reduce them.

## Anti-Semitism and racism: COVID-19 related hate speech and hate crime

The coronavirus pandemic together with the infodemic of pandemic-related false reports and conspiracy narratives on social networks which resulted in hundreds of deaths worldwide, was and is accompanied by a pandemic of group-specific prejudices: stigmatised groups of people are blamed for the spread of the pandemic. Historic diseases such as the plague, syphilis, the Spanish flu and the HI virus, were also associated with 'the others', i.e. with immigrant and minority groups. Because the COVID-19 virus broke out in China, blame has been placed especially heavily on people who are seen as East Asian. Hate speech against people seen as Asian appear on social networks in significant numbers - as do reports from those affected, who use those networks to draw attention to racist incidents under the hashtag #iamnotavirus. An increase of racist and anti-Semitic hate speech has been seen on social media in the wake of the COVID-19 pandemic, and there has been a rise in COVID-19-related racist and xenophobic incidents in the European Union.

The sometimes serious consequences of hate speech on an individual and social level are becoming a subject of growing research: a representative study conducted by the German Institute for Democracy and Civil Society (IDZ) with over 7,000 respondents from Germany in 2019 shows, among other things, the consequences of hate on the internet. Almost half (47%) of the people asked participate less often in internet discussions because of the fear of hate speech. Many of the internet users surveyed also withdraw completely from certain online platforms because of hate comments. Thus, 16% of respondents said they "used less, or no longer used, one of online services in connection with hate speech on the internet" and a further 40% said they would respond in the same way if they were the target of hate speech. 15% of participants - and for those under 24 even one in four (24%) - have "deactivated or deleted their profile on an online service" because of hate speech; 37% would do so if they felt the need. Taking action against hate speech is therefore also in the economic interest of these platforms. Two thirds to three quarters of respondents support measures to combat hate speech. For example, the creation of designated points of contact between victims, commissioners and/or central investigation units within the police and the public prosecutor's office. Other measures supported by respondents include financial assistance for victims and educational opportunities in schools on the topic.

Comparing across Europe, the European Commission's current evaluation report shows that the most frequent forms of reported hate speech incidents are related to sexual discrimination, followed by xenophobia, racism, antiziganism and anti-Semitism.

In extreme cases, discriminatory threats based on physical appearance led to aggressive behaviour and violent hate crime, which has increased in some countries during the pandemic, despite curfews and lockdowns. It is not only people who supposedly or actually come from East Asia who were and are declared scapegoats in the pandemic. Hate campaigns are also directed against Muslims, Black People and People of Colour, Roma and the LGBTQ+ community in Europe, depicting these groups as unhygienic, irresponsible or better off than the majority of society in the pandemic. Right-wing extremists spread, among other things, inflammatory posts claiming that Muslims are deliberately spreading the virus among non-believers. At the same time, there are calls in right-wing extremist Telegram groups to spread the virus among BIPOC, Jews and liberals.

Antiziganistic discrimination and hate messages have been reported in Romania, Hungary and Slovakia, among other countries. Anti-Semites around the world declared Jews or the State of Israel to be the real source of the virus and set in motion a wave of rumours full of insinuations, accusations and conspiracy speculations with a definite anti-Semitic character. The blame is placed on 'the Rothschilds' or Georg Soros. Moreover, governmental pandemic measures are equated with the dictatorship of National Socialism in Germany and with the persecution of Jews at that time. This serves to relativise the Shoah. This revealed numerous historical parallels with past anti-Jewish prejudices and passed-down cultural ideas, in which Jews were portrayed as 'well poisoners', 'child murderers' and as secret rulers of governments and the media. The 'QAnon' movement is driven by a metaphorical actualisation and collective restaging of this anti-Semitic conspiracy paranoia - exploiting the mechanisms of subcultures and social networks.

Generally speaking, the COVID-19 pandemic did not invent new target groups of hate speech, but rather reactivated historical patterns of devaluation in the context of the respective national traditions. The radical and extreme right, too, has only adapted its identity-forming agenda to the conditions of the pandemic. This applies also to attacks and calls for violence against the established media, democratic politicians and transnational or multilateral organisations, as well as to hostility towards science, which in the pandemic has unfortunately extended to doctors, nurses and researchers, as well as to some extent to the police and other state authorities – i.e. those who bear responsibility in the field for implementing the measures to combat the pandemic. As a result, hate speech and anti-Semitic and racist scapegoat narratives not only divide the integrity of inclusive societies, but also hinder the fight against the pandemic.

#### **Other violent protest phenomena in the pandemic**

The radical and the extreme right in its parliamentary, virtual, cultural, violent and terrorist manifestations are the greatest threats to liberal democracy and social cohesion in Europe as a whole in terms of political movements and parties. But it is not only radical, extreme and populist right-wingers who seek to exploit the pandemic. During this crisis, Islamist groups have increased their efforts and their reach for approaching and radicalising people, especially online, and have recently carried out terrorist attacks in Europe once again. The ‘Islamic State’ is calling for attacks in Europe with simple means during the pandemic to exploit the particular vulnerability of Western states and societies. As the pandemic unfolded, France was once again heavily hit by Islamist terrorism. The UK, Germany and Austria also suffered Islamist terrorist attacks with fatalities during the pandemic. Mutual escalation dynamics and cumulative radicalisation processes between Islamists and right-wing extremists present a particularly serious threat. Islamists’ reaction patterns to the pandemic are also diverse and based on scapegoating. For example, the pandemic is taken as evidence of the truth of an Islamic Orthodox lifestyle (including religious hygiene practices) and abstinence from the alleged Western decadence. As punishment for ‘sins’ such as alcohol and drug consumption, party hedonism, eating pork, homosexuality and promiscuity, the virus spreads mainly among ‘non-believers’.

COVID-19 is a virus of the West against which the Islam is the ‘immunisation’. Islamist narratives also contain anti-Semitic and anti-liberal conspiracy narratives that often are similar to those of right-wing extremism. Apocalyptic interpretations, according to which the pandemic heralds the end of the Western democratic world, are also a part of the narrative. Just as in anti-vaxxer and right-wing extremist environments, the blame for the virus is placed on Israel, malicious ‘globalists’, Bill Gates or the USA. Additionally, Islamists in Europe organise social welfare and social support services.

The radical left’s position in terms of content generally varies greatly within and between countries. Nevertheless, violent factions from the militant left spectrum carried out several attacks in Europe with reference to the COVID-19 pandemic. They were mainly directed against telecommunication infrastructures. Anti-Semitic and conspiracy ideological narratives about the coronavirus are also circulating in comparatively small sections of the anti-capitalist left, as well as the hope that the long awaited demise of capitalism could be accelerated by the pandemic. In virtual communities as well as in street protests, different oppositional movements mix to express their protest against the government’s policy in the crisis. On the one hand, this makes it difficult to classify them; on the other hand, it is accompanied by a cross-movement mixing and spreading of conspiracy theories, and of anti-Semitic and anti-democratic narratives.

#### **Complexity and radicalisation of the protests against the COVID-19 response policy**

A number of more or less heterogeneous protest movements against restrictions or state policies during the pandemic have emerged in Germany, Austria, Switzerland, France, the UK, Ireland and other European countries. In these movements, opinion-forming and mobilisation practices are closely linked to activities in the social media on the internet. The dominant protest motives can be very different and are by no means always right-wing. On the whole, the influence of radical and extreme right-wing players is relatively high. The ‘critical’ conspiracy-ideological, anti-democratic discourses comprise or form only a fraction of the general ‘infodemic’, by which is meant the inflationary spread of pandemic-related information and its effects. Particularly well-known people within the so-called ‘alternative media’ scene, or the leading media of the radical and extreme right, exploit the discourse around the restrictive measures to further their agenda.

From Spain to England to Poland, far-right parties and far-right activists and hooligans are leading or ‘hijacking’ COVID-19-related protests. Demonstrating is a fundamental human right and there are good reasons in this crisis to go out into the streets to protest for social balance and justice, as is happening in many European countries. But anti-Semitism, scapegoating, hatred and right-wing extremism exacerbate social divisions and do not solve any of the serious problems.

As diffuse as the multitude of content disseminated online and offline may seem, distinctive aspects or components can nonetheless be identified: the spreading of misinformation, the creation of uncertainty through speculation and a sometimes fundamental questioning of official reporting of established or state-run and reputable media, the spreading of conspiracy ideologies and myths (see below) as well as esoteric or spiritually inspired beliefs and convictions. The latter can increase the sense of distance from state institutions or institutions perceived as close to the state (such as science and ‘orthodox medicine’), as well as the distrust in their representatives and the rejection of their decisions and policies - among other things because of the emphasis on the irrational. Long before the pandemic, social psychological research showed that adherents to conspiracy ideologies were more likely to reject established prophylactic and curative methods and vaccinations. The fact that even extremely risky ‘alternative healing methods’ are presented and propagated as unproblematic on the internet or in social media can be seen as indicative of how important it is to provide accurate multimedia information and education in the field of public health awareness.

## Conspiracy ideologies

The outbreak of the pandemic and the state countermeasures were followed by a wave of disinformation and the visibility of conspiracy, racist and anti-Semitic narratives. This brought conspiracy ideologies back into the spotlight of public interest. They can undermine trust in democratic institutions and put lives at risk around the world by denying the dangers or existence of the virus or promoting life-threatening alternative treatments. This is often accompanied by a rejection of democratic institutions and the denigration and vilification of minorities, especially those defined as Jewish or Asian. It is only since the middle of the 20th century that such ideologies have increasingly been critically questioned, and since then they have experienced an overall decline. However, a high widespread prevalence of such attitudes must still be assumed today. Despite national differences and little comparability between studies, surveys in countries such as Italy, England, France, Hungary and Slovakia indicate generally a relatively strong prevalence of conspiracy belief in some parts of European societies. In some countries, particularly those with right-wing populist governments such as Hungary, Brazil and the USA, these beliefs sometimes also influence specific government actions - with dangerous, even fatal consequences for affected individuals and population groups. Scientific studies have long confirmed what is evident in the COVID-19 protests on the streets and in social media: democracy scepticism, devaluation of minorities and affinity for violence often go hand in hand with conspiracy ideologies. In right-wing populism, the vertical dimension (anti-elitism) is extended by the horizontal hierarchisation of ‘us’ versus ‘them’ and is therefore based on the same ‘logic’ that is followed by conspiracy ideologies. Across Europe, strong positive correlations are found between believing in conspiracy narratives and the willingness to vote for right-wing populist parties. As the virus spreads, affecting people globally, so does the network and dissemination of conspiracy ideologies. Particularly notable is the ‘QAnon’ conspiracy theory, which alleges a global conspiracy of ‘elites’ with ‘satanic’ and ‘paedo-criminal’ networks using predominantly anti-Semitic argumentation patterns, has become increasingly popular internationally. Donald Trump is revered as a pseudo-religious saviour figure who is going to destroy these networks. Symbols of and content from the ‘QAnon’ movement have since been seen at numerous demonstrations against the pandemic prevention measures across Europe. The risk of violent attacks from this movement is increasing.

---

## **SOCIAL CLEAVAGES, NEW SOCIAL MOVEMENTS AND COUNTER-MOVEMENTS**

Around the world, social movements use social networks for their causes. Counter-movements also react immediately to these. New social movements, typically initiated by social minorities and which usually rely only on limited institutional, human and material resources and socio-political influence or interpretive power, can today use the internet to create favourable opportunity frameworks for association, identity formation, consolidation and expansion. Finally, the internet facilitates inter- and transnational activities and cooperation. The web is particularly important for young people and vulnerable communities such as ethnic, cultural and social minorities. For them, social networks or specific areas of the internet represent important, indispensable or even exclusive protected spaces for interaction.

Social transformation, new social movements and civic groups or initiatives become targets of aggression. Counter-movements are forming and expressing themselves in the social networks and have become an international phenomenon. This applies above all to counter-movements in which strong structural links to digital hate cultures exist or whose supporters promote or advocate hate speech, discriminatory behaviour or other norm violations. These counter-movements aim at and contribute to a cultural backlash, which today is primarily orchestrated via social networks. This cultural backlash also manifests itself outside the internet in the form of a culture war instigated by the far-right and contributes to polarisation and mutual radicalisation. The action-reaction dynamic goes hand in hand with polarisation and radicalisation tendencies along the lines of social division, which increases the likelihood of confrontations - presumably above all around the conflicts outlined below.

### **Migration and asylum**

Intensifying refugee and migration flows to Europe and their consequences since the middle of the past decade have been surrounded by a controversial debate both online and offline. Migratory movements are a part of broad globalisation processes in which the importance of nation-state solutions is diminishing. On the one hand, local and transnational aid and solidarity networks were organised in 2015/16 to support refugees. On the other hand, it was followed by an escalation of nationalist, chauvinist, racist and anti-Muslim hate speech, hate crimes and far-right propaganda. In European countries where governments advocated and implemented the admission of refugees, not only protection seekers, 'asylum-friendly' individuals, NGOs and public institutions but also state representatives were attacked online and offline. This even extended to terrorist individuals who networked via social media.

### **Black Lives Matter and Decolonize**

The anti-racist citizens' movement in the 2020 protests across the USA has been adapted and adopted in Europe, not least by means of the internet. There, it has given new relevance and momentum to discussions about everyday racism, institutional and structural racism, colonial history, neo-colonialism and global relations of exploitation. This triggered competitive posting of racist content, especially on social media. Not only did radical and extreme right-wing politicians and activists play a key role here and presented themselves as 'role models' for 'non-conformists', 'freedom-lovers' or the 'traditionally-minded'; the combative term 'political correctness' has now also been adapted by the so-called 'centre of society' and is now commonly used to reject criticism of racism.

### **The climate and environmental movement**

Hardly any other protest movement has gained as much attention and significance in recent decades as the international climate movement (School Strike for Climate/ Fridays for Future, Extinction Rebellion, etc.). This can without a doubt be attributed to the pressing issues of our time and the large number of committed people, but also in part to the resistance they face, especially through hate messages on the internet. This is how the Swede Greta Thunberg has become, in a relatively short time, not only one of the most famous women in the world, but also one of the most hated (see below). There has been a growth of counter-communities and political networks, which deny human-influenced climate change, try to intimidate activists, politicians, journalists and scientists with hate speech and use disinformation to influence debates in the digital space. It is possible to predict: Conflicts over the ecological transformation will gain significance in the future and counter-movements will follow similar patterns as the resistance against the pandemic policies.

### **(Queer) Feminism, Gender, LGBTQ+ and the Anti-Gender Movement**

The fact that the modern women's rights movement attracted and continues to attract organised opponents is part of its history of more than 100 years. The rejection of allegedly exaggerated or harmful demands for legal and social equality, self-determination and comprehensive social participation of all genders is still expressed with varying degrees of aggression today and can sometimes lead to open hatred of women, which in its turn can also turn into violence. The sexist 'Manosphere' or the social Darwinist-fatalist Incel subcultures are examples of this hateful behaviour. The rejection of gender theories and practices has emerged as an important transnational trend in recent years. The movement known as 'anti-genderism' is directed against (queer) feminist politics and NGO activities, as well as against equality and acceptance of LGBTQ+ people.

In Eastern Europe in particular (e.g. Poland, Hungary and Russia) discrimination is especially strong in part because it is governmentally legalised, legitimised or even imposed. But also in Central, Western and Northern Europe, anti-gender propaganda and opposition to (queer) feminist demands and LGBTQ+ rights is a significant issue. Online activities of the anti-gender movement include misogynist, homophobic and transphobic hate speech, propaganda and coordinated mobilisation for collective action directed against women's rights, LGBTQ+ and the acceptance of gender/sexual diversity.

### **Anti-science hostility**

In the debate on the origins and handling of the COVID-19 pandemic, but also on controversial topics such as climate protection, migration and social inequality, scientists are also facing hate speech, threats and violence. The threats come from different levels. From the government side, academic freedom is attacked and restricted especially in authoritarian regimes. In digital spaces as well as in their lives and work, academics are also targets of organised hate campaigns, insults and threats or even violent attacks. Since 2011, the organisation 'Scholars at Risk' (SAR) has documented hundreds of cases of attacks against scientists and is voicing serious concerns about increased threats during the pandemic.

It is foreseeable that polarisation along these and other socio-economic, political and cultural divisions (including inequality, urban-rural disparities) will be reinforced and accelerated by the progressing digitalisation.

## Progressing digitalisation

The digital revolution permeates all areas of life and social spheres. In particular, profound social innovations and responses are associated with the expansion of universally available high-performance data infrastructure and the increasing importance of machine learning. The global and European roll-out of 5G networks continues this rapid development and is likely to take it to the next level, which comes with significant challenges, risks and threats. As a result, we can observe trends in which the above-mentioned negative tendencies, threats to democratic coexistence in general, as well as the vulnerability of certain groups of people, are drastically intensified as digitalisation progresses.

Firstly, there is the increased speed of spread and reach of fake news, disinformation campaigns, hate speech, anti-democratic propaganda as well as criminal content, including the violation of personal rights (invasion of privacy, spying, stalking, blackmail, doxing, defamation, threats, insults, etc.).

Secondly, the technical prerequisites for even easier and almost unlimited distribution and use of enormous amounts of data - especially audiovisual, interactive content or services - will be achieved in the future. In addition, there is an increasing threat of audiovisual deep fakes. This will pose great challenges not only for content management and the administrative activities of platform operators, but also for the actions of state stakeholders and educational institutions. It will also affect the fields of radicalisation prevention, hate speech and hate crime intervention, as well as research into those areas. All of that requires a great deal of coordination, as cross-platform communication is likely to grow considerably as a result of the further diversification of online products and services. A particular challenge is the shift of radical and extreme right-wing activists to alternative platforms and communication channels. This way they try to evade regulations and react to or anticipate the deletion of their content and accounts. For this reason, intensified cooperation between different platform operators will become more important.

Thirdly, for radical, extreme or criminal activists and hate groups a universally available faster internet opens up an attractive field for action and even better opportunities for networking, advertising, mobilising, radicalising, and preparing for crimes. Innovative web-based forms of interaction are quickly adapted and used or abused by these groups, as recently demonstrated for example by the wave of racist and anti-Semitic Zoom-bombings during the COVID-19 pandemic in the spring of 2020. At the same time, new innovative formats offer activists increased opportunities for mimetic repackaging of content that violates laws and community standards, for camouflaging and disguising the authorship of dangerous content, for circumventing or disguising it, as well as for leveraging existing, increasingly outdated or ineffective prevention and intervention mechanisms.

These mechanisms must therefore undergo a fundamental revision and redevelopment.

---

## OUTLOOK AND RECOMMENDATIONS

The above mentioned trends and online-offline dynamics require action strategies that not only react to existing problems, but also counter the risks prospectively.

- In the course of current developments, tech companies will be even more responsible than before for guaranteeing fundamental rights (such as freedom of expression, press, art and information), for protecting personal rights, intellectual property, human rights and for anti-discrimination.
- Strategic decisions by platform operators and tech companies should be evidence-based and provide a high level of transparency for the general public. This requires the involvement of decision-makers from politics, the judiciary, academia, security authorities as well as representatives of civil society.
- Improved strategies for dealing with social divisive and polarising tendencies together with consistent combating of hate and inflammatory speech can help maintain and regain the trust of users.
- In order to increase the acceptance of social networks' rules, tech companies should institutionalise joint, transparent and science-based procedures to take cross-platform actions against hate organisations, both nationally and internationally, and to communicate decisions, such as profile or page removals, in a comprehensible way. This is especially true for those fields where the legal norms for the articulation of certain social and political interests, freedom of expression, freedom of the press, freedom of art and personal rights are affected, and which in some cases differ greatly from country to country.
- This requires innovative European concepts: The regulatory and innovation gap in the digital transformation process in Europe should be addressed holistically. The current deficit situation offers the opportunity for implementing digital design concepts that would develop into an independent social digital strategy between the two market-dominating positions of digital repression and control (China) and private-sector deregulation (USA). This social digital strategy should focus on strengthening the positive sides of digitalisation and minimising the negative consequences.
- The messenger service Telegram in particular has become a radicalisation accelerator. The service does not regulate itself in practice, which must change as soon as possible. Legal regulations and penal procedures must also be applied to this and similar platforms.
- There is a pressing need to optimise the recognition systems for right-wing extremist audiovisual content as well as internal aesthetics, vernacular and symbolism, some of which are highly nationally and regionally specific. Here, an even closer cooperation of the platforms with civil society and scientific expertise could close some existing gaps.
- It is important to sustainably strengthen pluralist-democratic and emancipatory counter-narratives through active support of information, education and awareness-raising services or infrastructures in these thematic fields. Platform operators should, for example, cooperate more often with educational institutions and NGOs to promote the development and popularisation of interactive formats that would serve as forums for open, constructive and civil debate culture on the internet. Strategies against marginalisation and structural exclusion in online-based social discourses (e.g. through social inequality, ethnocultural identity, age and language barriers) are highly relevant here.
- Participants of the (digital) civil society are the most important actors in the fight against hate and right-wing extremism and are also important cooperation partners for platforms and researchers in identifying current trends at an early stage. The transfer of expertise between platforms, science and civil society must therefore be institutionalised more at a national and international level. Participants in civil society need to make greater use of their local, regional, national and transnational resources to increase the resilience of solidarity-based network communities against hate and inflammatory speech and to strengthen the resistance against conspiracy ideologies. It is also essential to provide members of civil communities with the skills to tolerate and objectively disprove ambiguities and contradictions.



- This includes improving digital media literacy, which enables users to critically examine external information and communication content, and to reflect on their own selection, consumer, participation and sharing behaviour. Governmental and non-governmental educational offers should not only consider questions of personality or data protection or 'netiquette', but should focus on the practical democratic relevance of the online-offline dynamics that are problematic.
- During the COVID-19 pandemic, it is even more important to keep a distance from and oppose anti-Semitism, racism, misinformation, conspiracy thinking, the glorification of violence and anti-democratic coup fantasies. Existential fears and discontent during the crisis do not justify support for hate organisations.
- Politics must - sustainably and with a focus on social balance - mitigate the effects of the pandemic and the associated social crisis dynamics. It must not fall into the trap of nationalism and must always explain drastic measures transparently and objectively. This will also be crucial for future crisis management, such as in combating the consequences of climate change.
- In order to shape the multidimensional processes of accelerated social change and mitigate the negative consequences, social work with and in digital spaces and communities should be supported (digital streetwork).
- The fundamental reorganisation of collective knowledge generation in the 21st century confronts research in the social sciences and humanities with the task of overcoming a technological knowledge gap: It is currently insufficiently prepared for the multiplication and lack of transparency in information flows, network structures and data sources. This applies to a large extent to research on right-wing extremism/right-wing terrorism, radicalisation, discrimination and hate speech. To keep pace with structural and technological changes, research has to rely on funding, technology and data.

## CITATIONS

# Recent Trends and Perspectives: COVID-19 Pandemic Increases Right-Wing Extremism and Hatred in Europe

- 1 Bjola, Corneliu/James Pamment (2019): Countering Online Propaganda and Extremism. London and New York: Routledge
- 2 Freelon, Deen/Alice Marwick/Daniel Kreiss (2020): False Equivalencies: Online Activism from Left to Right. In: Science, Sep 2020, S. 1197-1201
- 3 Fielitz, Maik/Holger Marcks (2019): Digital Fascism: Challenges for the Open Society in Times of Social Media. UC Berkeley: Center for Right-Wing Studies. <https://escholarship.org/uc/item/87w5c5gp>
- 4 Littler, Mark/Benjamin Lee (2020): Digital Extremisms. Readings in Violence, Radicalisation and Extremism in the Online Space. London: Palgrave Macmillan
- 5 Freelon, Deen/Alice Marwick/Daniel Kreiss (2020): False Equivalencies: Online Activism from Left to Right. In: Science, Sep 2020, S. 1197-1201
- 6 Awan, Imran, Hollie Sutch and Pelham Carter (2019): Extremism Online – Analysis of extremist material on social media. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/834369/Awan-Sutch-Carter-Extremism-Online.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/834369/Awan-Sutch-Carter-Extremism-Online.pdf)
- 7 Fielitz, Maik/Karolin Schwarz (2020: Coming): HATE NOT FOUND?! The Deplatforming of the Far-Right and its Consequences. Project report
- 8 Fielitz, Maik/Holger Marcks (2019): Digital Fascism: Challenges for the Open Society in Times of Social Media. UC Berkeley: Center for Right-Wing Studies. <https://escholarship.org/uc/item/87w5c5gp>
- 9 Exif: Der „Kampf der Nibelungen“ 2020: Online-Stream statt Großevent. (The ‘Battle of the Nibelungs’ 2020: Online Streaming in place of a Major Event.) <https://exif-recherche.org/?p=6760>
- 10 Nagle, Angela (2017): Kill all Normies: Online Culture Wars from 4chan and Tumblr to Trump and the Alt-Right. Winchester, UK, and Washington D.C.: Zero Books
- 11 Claus, Robert (2020): Ihr Kampf. Wie Europas extreme Rechte für den Umsturz trainiert. (Their fight. How Europe’s Extreme Right Trains for the Coup.) Bielefeld: Die Werkstatt
- 12 Strobel, Benjamin (2020): From humour to hate. Right-wing ideologies run rampant on the Steam gaming platform. How does this happen? <http://www.neues-deutschland.de/artikel/1143366.computerspiele-von-humor-zu-hass.html>
- 13 Puls, Hendrik: The Gamification of Terror. Ein brauchbarer Begriff um rechtsterroristische Anschläge zu beschreiben? (‘Gamification of Terror’. A Useful Term to Describe Right-wing Terrorist Attacks?) [http://www.nfg-rexdel.de/images/Working\\_Paper\\_3.pdf](http://www.nfg-rexdel.de/images/Working_Paper_3.pdf)
- 14 Dittrich, Miro/Jan Rathje (2019): The Mob is the Movement. Das Netzwerk rechter Onlinekultur von #Gamergate zu „Alt-Right“. (‘The Mob is the Movement’. The Network of Right-wing Online Culture from #Gamergate to the ‘Alt-Right’.) <https://www.antifainfoblatt.de/artikel/das-netzwerk-rechter-onlinekulturen-von-gamergate-zu-%E2%80%9Ealt-right>
- 15 Nagle, Angela (2017): Kill all Normies: Online Culture Wars from 4chan and Tumblr to Trump and the Alt-Right. Winchester, UK, and Washington D.C.: Zero Books
- 16 Reference groups that activists distance themselves from, reject or even fight against are described as ‘negative’. Positive reference groups, on the other hand, are those with which activists identify and on which they base their thoughts and actions (see Waldmann, Peter (2005): Determinanten des Terrorismus. (Determinants of Terrorism.) Weilerswist: Velbrück, S. 86)
- 17 Ackermann, Gary/Hayley Peterson (2020): Terrorism and COVID-19: Actual and Potential Impacts. In Perspectives on Terrorism 13/3, S. 59–73
- 18 Bieber, Florian (2020): Global Nationalism in Times of the COVID-19 Pandemic. In Nationalities Papers, S. 1–13. DOI: 10.1017/nps.2020.35; Ozkirmli, Umut (2020): Coronationalism? <https://www.opendemocracy.net/en/can-europe-make-it/coronationalism/>
- 19 Blais, André/Bol, Damien/Giani, Marco/Loewen, Peter John (2020): The Effect of COVID-19 Lockdowns on Political Support: Some Good News for Democracy? In European Journal of Political Research. DOI: 10.1111/1475-6765.12401
- 20 Bartoš, Vojtěch/Bauer, Michal/Cahlíková, Jana/Chytilová, Julie (2020): COVID-19 crisis fuels hostility against foreigners. <http://vojtechbartos.net/wp-content/uploads/CovidHostility.pdf>
- 21 Cinelli, Matteo/Quattrociocchi, Walter/Galeazzi, Alessandro/Valensise, Carlo Michele/Brugnoli, Emanuele/Schmidt, Ana Lucia/Zola, Paola/Zollo, Fabiana/Scala, Antonio (2020): The COVID-19 Social Media Infodemic. <http://arxiv.org/pdf/2003.05004v1>; Islam, Md Saiful/Sarkar, Tonmoy/Khan, Sazzad Hossain/Kamal, Abu-Hena Mostofa/Hasan, S. M. Murshid/Kabir, Alamgir/Yeasmin, Dalia/Islam, Mohammad Ariful/Chowdhury, Kamal Ibne Amin/Anwar, Kazi Selim/Chughtai, Abrar Ahmad/Seale, Holly (2020): COVID-19-Related Infodemic and its Impact on Public Health: A Global Social Media Analysis. In The American Journal of Tropical Medicine and Hygiene 103 (4), S. 1621–1629. DOI: 10.4269/ajtmh.20-0812
- 22 Velásquez, N./Leahy, R./Johnson Restrepo, N./Lupu, Y./Sear, R./Gabriel, N./Jha, O./Goldberg, B./Johnson, N.F. (2020): Hate multiverse spreads malicious COVID-19 content online beyond individual platform control. <https://arxiv.org/ftp/arxiv/papers/2004/2004.00673.pdf>
- 23 European Union Agency for Fundamental Rights (FRA) (2020): Coronavirus Pandemic in the EU – Fundamental Rights Implications. [https://fra.europa.eu/sites/default/files/fra\\_uploads/fra-2020-coronavirus-pandemic-eu-bulletin-1\\_en.pdf](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-coronavirus-pandemic-eu-bulletin-1_en.pdf)
- 24 Geschke, Daniel, Anja Kläßen and Christoph Richter (2019): Hatred on the Net: The Creeping Attack on our Democracy. Eine bundesweite repräsentative Untersuchung. (Hate on the Internet: The Covert Attack on our Democracy. A Nationwide Representative Study.) <https://www.idz-jena.de/fileadmin/>

- 25 Pankowski, Rafał/Dzięgielewski, Jacek (2020): COVID-19 Crisis and Hate Speech. <https://tandis.odhr.pl/handle/20.500.12389/22640>
- 26 Ebner, Julia (2017): The Rage. The Vicious Circle of Islamist and Far Right Extremism. London, New York: IB Tauris
- 27 Mason, Paul (2020): „Wir befinden uns in einem Bürgerkrieg“. („We are in a Civil War“) <https://www.ipg-journal.de/regionen/global/artikel/wir-befinden-uns-in-einem-buergerkrieg-4733/>
- 28 Islam, Md Saiful/Sarkar, Tonmoy/Khan, Sazzad Hossain/Kamal, Abu-Hena Mostofa/Hasan, S. M. Murshid/Kabir, Alamgir/Yeasmin, Dalia/Islam, Mohammad Ariful/Chowdhury, Kamal Ibne Amin/Anwar, Kazi Selim/Chughtai, Abrar Ahmad/Seale, Holly (2020): COVID-19-Related Infodemic and its Impact on Public Health: A Global Social Media Analysis. In The American Journal of Tropical Medicine and Hygiene 103 (4), S. 1621-1629. DOI: 10.4269/ajtmh.20-0812
- 29 Lamberty, Pia/Imhoff, Roland (2018): Powerful Pharma and its Marginalized Alternatives? Effects of Individual Differences in Conspiracy Mentality on Attitudes Toward Medical Approaches. In Social Psychology, July 2018. DOI: 10.1027/1864-9335/a000347.
- 30 Butter, Michael (2018): Nothing is as it Seems. About Conspiracy Theories. Frankfurt a.M.: Suhrkamp.
- 31 Moreno Mancosu/Salvatore Vassallo/Cristiano Vezzoni (2017): Believing in Conspiracy Theories: Evidence from an Exploratory Analysis of Italian Survey Data. In South European Society and Politics 22:3, S. 327-344.
- 32 University of Oxford (2020): Conspiracy beliefs reduce the following of government coronavirus guidance. <https://www.ox.ac.uk/news/2020-05-22-conspiracy-beliefs-reduces-following-government-coronavirus-guidance>
- 33 Rees, Jonas H./Pia Lamberty (2019): Compelling Truths: Conspiracy Myths as a Danger to Social Cohesion. In Zick, Andreas/Küpper, Beate/Berghan, Wilhelm (Hrsg.): Lost Center - Hostile Conditions Right-wing Extremist Attitudes in Germany (2018/19) Bonn: Dietz
- 34 Thórisdóttir, Hulda/Silvia Mari/André Krouwel (2020): Conspiracy Theories, Political Ideology and Political Behaviour. In Butter, Michael/Knight, Peter (Hrsg.): Routledge Handbook of Conspiracy Theories. London and New York: Routledge
- 35 Quent, Matthias (2019): Civil society: Defending the Global Village: Strategies against the Cultural Backlash on Social Media. In Baldauf, Johannes/Ebner, Julia/Guhl, Jakob (Hrsg.): Hate Speech and Radicalisation Online: The OCCI Research Report. London: Institute of Strategic Discourse (ISD) / in cooperation with Facebook), S. 44-49
- 36 European Union (2020): European Union terrorism situation and trend report 2020. [https://www.europol.europa.eu/sites/default/files/documents/european\\_union\\_terrorism\\_situation\\_and\\_trend\\_report\\_tesat\\_2020\\_0.pdf](https://www.europol.europa.eu/sites/default/files/documents/european_union_terrorism_situation_and_trend_report_tesat_2020_0.pdf)
- 37 ADL (2020): ADL Finds Hateful Content on Telegram Exceeds That on Facebook. <https://www.adl.org/news/press-releases/adl-finds-hateful-content-on-telegram-exceeds-that-on-facebook>
- 38 Kracher, Veronika (2020): Incels: Geschichte, Sprache und Ideologie eines Online-Kults. (History, Language and Ideology of an Online Cult.) Mainz: Ventil
- 39 FRA (2020): EU LGBT survey. European Union lesbian, gay, bisexual and transgender survey. [https://fra.europa.eu/sites/default/files/fra-eu-lgbt-survey-main-results\\_tk3113640enc\\_1.pdf](https://fra.europa.eu/sites/default/files/fra-eu-lgbt-survey-main-results_tk3113640enc_1.pdf)
- 40 <https://www.scholarsatrisk.org/academic-freedom-monitoring-project-index/>



‘SIEGE is to be used as a cookbook and guide,’ wrote Ryan Schuster in his introduction to the explosive neo-Nazi tome by James Mason, *Siege* (first published in 1992).<sup>1</sup> A lifelong racist and neo-Nazi, in *Siege*, Mason created an unflinchingly anti-Semitic, racist, homophobic, misogynistic text that promoted ‘lone wolf’ terrorism in the name of accelerationism, i.e. the execution of mass violence in order to induce a race war that would lead to the downfall of existing democratic multicultural societies.

Schuster’s words have proven prescient—*Siege* is now at the heart of the reinvigorated global neo-Nazi movement, stoking violent rhetoric and solo-actor terrorism. In the roughly four decades since its original publication, Mason’s work has adapted and evolved to reflect changing socio-political landscapes. It is a living thing, having spawned an entire subculture known as *Siege Culture*.

The Internet and social media in particular have been essential in the propagation of *Siege* and the invention of *Siege Culture*. With the free, easy accessibility of the source text, blogs devoted specifically to *Siege Culture*, and innumerable posts on social media praising Mason and encouraging audiences to read *Siege*, this work has radicalised innumerable twenty-first century extremists and terrorists. It is directly responsible for some of the most graphic content found online today and for the deaths of multiple people around the world. It is a large part of what the United Nations has identified as the ‘growing and increasingly transnational threat posed by extreme right-wing terrorism.’<sup>2</sup> Thus, understanding *Siege* and its subculture is essential to countering online neo-Nazi incitement and criminality.

---

## BACKGROUND ON THE PRODUCTION OF SIEGE

Originally, Mason wrote *Siege* as a newsletter for the American-based National Socialist Liberation Front (NSLF) between 1981 and 1986. Understanding the origins and evolution of the document is vital to grasping *Siege Culture* more generally, as it captures the extent to which Mason's writing has developed in reaction to changing social and political landscapes. Ahead of his time in many ways, Mason's promotion of 'lone wolf' terrorism, his denunciation of hierarchical group structures, and his intense admiration for Charles Manson placed him at the fringe of the already fringe elements of the right-wing. While certain scholars have dated the idea of 'leaderless resistance'—more commonly known as 'lone wolf' terrorism today—to an article by Louis Beam in the late 1980s, the reality is that *Siege* called for single actor terror attacks years before; in fact, a chapter of *Siege* features the phrase 'lone wolves.' It is likely that Beam knew of Mason's calls for individualistic violence, if he did not directly draw inspiration from *Siege*.

Despite its limited readership (and popularity) as a newsletter in the 1980s, Mason eventually reformatted his writings into a monograph in 1992. 217 articles within the newsletter became the 'chapters' within *Siege*, coupled with an introduction to the text written by Mason. Organised not like an anthology of articles reliant on chronology, the text features thematic subsections such as "National Socialism" and "Lone Wolves and Live Wires," and runs to 434 pages. First released by radical publisher Storm Books, the book would be reformatted again, with new introductions, and printed with Schuster and Black Sun publication around a decade later, this time with almost 30 pages of new content. This pattern would repeat itself in 2017, when IronMarch released an edition of *Siege* standing at more than 560 pages, while the most recent edition of *Siege* (2018) is over 680 pages.

This growth is not merely a matter of different formatting or page size, nor the inclusion of additional works from the original 1980s newsletter. Rather, appendices prove the largest contribution to the text's expansion. These new appendices discuss twenty-first century realities both within and without the neo-Nazi movement, helping address the fact that the main body of the text remains riddled with 1980s references foreign to younger readers. With new text, scanned newspaper clippings, photographs and sketches, the appendices range in format and messaging, while also all serving the fundamental purpose of marketing Mason's ideas to readers. Notably, the two most recent editions focus heavily on Mason's relationship with Atomwaffen Division (AWD), a terrorist group operational in numerous countries around the world whose members have been charged with multiple violent crimes (see the *Siege Culture* and *Violence* section of this report for further details). In these appendices, photographic evidence places Mason in meetings with members of AWD, written work praises the deeds of alleged violent actors from the group, and ideas are further transmitted via numerous designs and sketches from AWD-favourite artist Dark Foreigner.

This link between international AWD and the appendices written for a more expansive audience also helps Mason expand his audience geographically. The original *Siege* articles focus on the American context, and in addition to the temporally dated references are geographically specific (e.g. analysis of publicised murders in America or circumstances in specific American cities). However, the appendices broaden *Siege's* scope to a global audience. Sketches and photographs related to Germany and the United Kingdom, as well as ruminations on the state of white people around the world, all make *Siege* appear today as the embodiment of an internationally influential ideology.

---

## SIEGE NARRATIVE AND RHETORIC

Since *Siege* initial publication there have been innumerable pieces of writing put forward online and in print that have served to sway the trajectory of twenty-first century neo-Nazism, but it is *Siege* that has had the greatest impact. This impact, arguably, is attributable to the appeal of the rhetoric and messages advanced in the book. It is therefore a worthwhile endeavour to take time to illuminate and interrogate the ideology of *Siege*, as well as to consider the specific (offensive) terminology it employs.

From his first published article in 1981 to the final lines of the appendix of the 2018 edition, Mason has produced work that is unmistakably and unrepentantly white supremacist, anti-Semitic, homophobic and misogynistic. Mason merges traditional discriminatory narratives with new terminology with the results being some of the most graphic, offensive language seen in white supremacist/neo-Nazi writing.

*Siege* rests upon the fundamental idea that Western societies as they exist at present—democratic, multicultural and multifaith—are doomed to fail because their fundamental beliefs are inherently flawed. Ideas of equity, according to Mason, ignore the inherent superiority of whiteness, and all social tension and outbursts of violence ultimately prove the result of forcing unnatural values upon people. Schuster observes how *Siege* characterises liberal democracies as part of “the System” that is a kind of “virulent poison” to people.<sup>3</sup> *Siege* promotes the accelerationist narrative, a declinist one in which all liberal societies will ultimately collapse, and that the best thing for (white) people is to help quicken this process through violence.

Exemplary is a line Mason wrote first in 1984 that finds resonance with twenty-first century audiences of *Siege*: “the country isn’t going but has gone MAD; that the final END of society is accelerating; that the entire foundation itself is thoroughly corroded; and that there is no longer any place to go to hide (save maybe a tent in the North Woods). Now isn’t that the most encouraging thing anyone has reported to you in a long, long time?”<sup>4</sup> While *Siege* warns that the coming race war will be comparable to the Dark Ages and the collapse of the Roman empire, it says that afterwards, the future (free from non-white peoples) would be bright.<sup>5</sup>

It would be bright, according to *Siege*, because of the lack of non-white peoples, which are characterised as ruining the world. As a neo-Nazi text, for instance, *Siege* promotes traditional Jewish conspiracy theories and Holocaust denial. “[It] was indeed a damnable shame that Hitler did not, in fact, kill at least six million Jews during the War,” one chapter states, while many others reiterate this point and argue that the Holocaust is a falsity foisted upon societies by Jewish-run media and government institutions.<sup>6</sup> This blends with his embrace, moreover, of the Zionist Occupation Government (ZOG) conspiracy theory that identifies Jewish people as covertly running all major government institutions, the media, and the banking system.<sup>7</sup> According to *Siege*, Jewish people do this because of a pathological/biological greed and need to exploit non-Jewish peoples, using their intelligence to capitalise on the suffering of other races. As such, within *Siege*, the extermination of all Jewish people is seen as a desirable element of the forthcoming race war that terrorists should attempt to facilitate.

In addition to employing exploitative tactics that harm white people, *Siege* ideology describes the Jewish community as manipulating and exploiting purportedly less intelligent, sub-human non-white races. Mason and *Siege* are particularly prolific in denouncing and ridiculing black peoples, most frequently African Americans. *Siege* propagates a narrative that blames black people for the supposed destruction of American values and societal violence (which black peoples are allegedly predisposed to use). Black Americans are frequently described using racial slurs, whilst being depicted as unintelligent, dirty and inferior to the white race. They are identified as a “source of filth” in American society,<sup>8</sup> and, for example, when pondering the idea of racial equity in light of the successes of black athletes, Mason writes “racial equity or even Black superiority? Don’t make me gag!”<sup>9</sup> Black lives, meanwhile, are identified as “expendable.”<sup>10</sup>

Little attention is paid in the original source text to non-black, non—Jewish peoples—though white supremacist rhetoric elsewhere in the text indicates that Siege ideology does not view them as equals to whites. However, reflective of shifts in global concerns on race and religion, the appendices in recent iterations of Siege focus a great deal more (negative) attention on these groups. Embracing classic anti-immigrant, Islamophobic ideas linked with the most extreme elements of nativism and populism, Siege promotes the idea that white Western societies are under imminent threat of extermination through migration. The ‘Great Replacement’ idea—wherein immigrant communities with high birth rates will eventually outnumber white people in formerly white-majority states and subsequently replace or destroy traditional values—is at play in Siege. There is also the influence of more contemporary vocabulary, with terms such as “white sharia,” part of a wider trend within white nationalist rhetoric to fuse terms linked to Islamic extremism (*sharia, jihad*) with their aims for white-only nation-states.

Another major indicator of the white supremacist ideology promoted by Siege is its violent stance on interracial relationships. Interracial individuals are belittled and, in addition to considering non-white peoples as subhuman and so warranting death, even white people who engage in or accept interracial relationships are threatened with death. The notion of blood purity abounds in Siege, a standard element of neo-Nazism. Articles state that white people “will be COMPELLED to join in or else die!” and that support for multiculturalism, which is deemed an “anti-White conspiracy,” would be “a crime that shall be punished by death.”<sup>11</sup>

Threats of death such as these are central to Siege and the ideology it promotes. As an accelerationist text, Siege promotes the idea that it is only through bloodshed that necessary reforms will occur. Not content even with the idea of forced migration or separatism, Siege outright aims at a ‘race war’ that would genocidally result in the deaths of all non-white people, in essence the establishment of an all-white world.

In pursuit of this, the text encourages individuals to engage in terrorism, and wherever possible plot and perpetrate hate crimes and terrorist attacks that will further the cause. Whilst stating that the other option for devotees is ‘total drop out’ (i.e. withdrawal from society, refusal to acknowledge the government, and adopt a lifestyle that features only other white supremacists/neo-Nazis), the work eventually states that all good white people should take up arms to overthrow the status quo. Anyone who stands in the way of the establishment of white-only states, moreover, is identified as worthy of death, even family members and vulnerable people.

---

The quotations below provide a sense of the level of dehumanising rhetoric and graphic violence promoted by Siege. They are emblematic of countless others scattered across the more than 600 pages of later editions.

---

“Strike hard and strike deep to build the climate for revolution...”<sup>12</sup>

“The Enemy is the Enemy and aliens are aliens. All politicians—high and low—are PIGS in a Pig System. If they weren’t, they wouldn’t be there. From President to dogcatcher, they are all the same bureaucratic, sell-out swine.”<sup>13</sup>

“The gameboard is rigged against us and so we are constrained to kick over the gaming table itself.”<sup>14</sup>

“Terrorism is a two-way street for, as Hitler stated, the only answer to terrorism is stronger terrorism.”<sup>15</sup>



It is worth specifically noting the psychological tactics Siege employs to manipulate readers into embracing its ideology. They are methods that are uniquely appealing to online subcultures and so it makes sense that Siege Culture has attained a cult online following for those on the fringes of society. In *Siege*, Mason targets those feeling disaffected, marginalised and isolated, who might self-identify as outsiders. The work addresses itself to readers who feel like social outcasts, who struggle to fit in with mainstream society, yet who nevertheless seek a community with a purpose. The “Lone Wolves and Live Wires” section is not, as might be expected, devoted to discussing the various ‘lone wolves’ that Mason venerates elsewhere in the text, but rather addresses itself to social outcasts, arguing that they can regain control over their lives and earn the respect of their peers if they engage in terrorism in the name of white supremacy. In plotting and perpetrating violence on their own and not with a group, moreover, Mason tells readers that they will prove themselves to be true, strong, independent men (and it is invariably men Mason calls to).

According to *Siege*, to be an outcast is a sign of mental and moral superiority, attributes which are further validated by embracing neo-Nazi ideology and becoming a terrorist. ‘The man of ill repute today,’ one chapter argues, ‘must one day go on to emerge as the Hero that he is just as our entire Movement must emerge as the saviour of an entire People.’<sup>19</sup> Elsewhere, Mason writes, “What I will describe is the kind of person who—while still very much an individual—stands apart, stands forth against the System. A person who is of such magnitude that his act of standing forth in such a manner is equivalent to whole worlds colliding. That kind of person can never and will never be counted as a victim or as a statistic. And that kind of person, whether his numbers are in the dozens or in the hundreds, is the kind which is making up the Movement of the future.”<sup>20</sup> “I am speaking of people who do not fit into THIS society because of what it IS and what THEY are. To be outside this society is a marked badge of honour,”<sup>21</sup> *Siege* reassures its readers. Couple this with the pages of praise Mason devotes to those who engaged in racist violence, memorialising and canonising them, and *Siege* can clearly be seen as a base text that attempts to manipulate and radicalise vulnerable people who might be lamenting their perceived lack of power and community involvement.

---

“We do however subscribe to the old adage that, in order to kill an ‘ism’ you must kill the ‘ists.’”<sup>16</sup>

“I’ve been told in recent years that we just can’t blow the heads off the powers that be, that we simply cannot call for anarchy. But what these sensitive, conservative types can’t grasp, or else refuse to grasp, is that the alternatives are either fast being removed by circumstances themselves, or they are gone already.”<sup>17</sup>

“You have to be determined to do whatever is necessary in order to win. And it matters not against whom, once they have demonstrated that it is conflict they want. Comrade, friend, family member...it doesn’t matter.”<sup>18</sup>

---

## SIEGE CULTURE WEBSITE

Siege as a monograph is now listed as essential reading for many attempting to join terrorist organisations. However, the book format is not the exclusive means by which Mason's particular iteration of neo-Nazism has been propagated. Individuals inspired by Mason's ideology set about creating an entire subculture (online and off) what is now known as Siege Culture, alternatively spelled Siege Kultur or Siege Kulture. This derivation of Mason's beliefs has been pivotal in the radicalisation of numerous young people, and to the violent actions of many. As such, the rest of this report will focus specifically on the nature of Siege Culture as an online and offline subculture, outlining its features and articulating its real-world implications (both to date and potentially into the future).

The hub of this subculture was, until recently, concentrated on an eponymously-named Siege Culture website, curated by members of AWD.<sup>22</sup> Registered as a website in late 2017 (it has been taken down multiple times and at the time of this report is not operational), for years the site consisted largely of a blog featuring prose and pictorial posts expounding and expanding upon Siege. Sections of the blog interrogated specific themes within Siege and other radical texts while other pages on the website were designed to help fundraise and connect visitors with potential recruiters. According to the site's 'Worldview' page, "What we are creating here is something that James Mason attempted to put into form but because of circumstances it never was implemented [sic] until the year 2017 when Atomwaffen Division discovered and met James Mason...Too long has the movement trapped people into a mindset of chasing their own tail. Those of you who are in here, perhaps, will create history. That is our intention."<sup>23</sup>

Mason's hate-filled ideology and Siege's mixture of dehumanising non-white peoples whilst attempting to draw in disaffected white readers is on full display throughout the various pieces of writing, not least because Mason himself contributed to the site. For instance, visitors to the blog would routinely be met with such statements about racial hierarchies such as these featured (on the right).

---

"...it doesn't matter whether one is White, Black, Yellow, Brown or whatever as 'we are all equal'. This is the greatest lie ever propounded."<sup>24</sup>

"'Equality' and 'the rights of man' bullshit."<sup>25</sup>

"They have a stage play now about Alexander Hamilton full of niggers in the cast. Total falsification of history, Stalinist style..."<sup>26</sup>

"Whites have development and destiny. World domination and high culture as well as science."<sup>27</sup>

"All high civilizations start out as White but die after having assimilated sufficient colored blood."<sup>28</sup>

"The reality is that most of the human types on earth have a lesser proportion of this God-given brain than do some of the rest. And here again, we are regarding Whites. In short, you may rule out the coloreds. Those in whom you can literally SEE the animal in their very countenance... "You can take the nigger out of the jungle but you can't take the jungle out of the nigger."<sup>29</sup>

[Of Muslims] "They hate us for what we are. As long as we survive they must see themselves as what they are. And that is a miserable image which they can't stand. They can do nothing about it. They can't clean themselves up. They can only try to pull us down...The ONLY God is OUR God."<sup>30</sup>

---

## EXCERPTS OF REFLECTIONS

---

Anti-Semitism, including overt Holocaust denial, was also a routine feature on the website:

---

“We today enjoy TONS of material, literature and evidence of our own which literally demolishes the big lie of the so-called ‘Holocaust’...[Jews] WERE the Reds, they WERE the oppressors, the torturers, the overlords of most recent months and years... So much for ‘The Holocaust’, a work left uncompleted.”<sup>31</sup>

“This business of so-called ‘racism’...an epithet created by the Enemy even as he was formulating all the rest of his attack against us. What it defines in reality is no more or less than anormal condition of wanting to be with one’s own fellows, abide by his own laws, traditions and customs and, above all, to BREED TRUE. Anything aside from this natural condition is nothing more than DEPRAVITY and DEGENERACY, some alien form of twisted and wholly artificial, unnatural mind warp. It arose, took hold and came to dominance through alien Jews whose wish to destroy us as a people goes back at least to their murder of Christ.”<sup>33</sup>

“Take away the falsehood of the ‘gas chambers’ and what do you have left? Hitler dared to TOUCH the supposedly UNTOUCHABLE!”<sup>32</sup>

“The foulest degenerates will be elevated to practically ‘hero’ status while the most worthy of genuine national hero types you will never even hear of or, if you do, they will be tar-brushed as ‘haters’ or now even as ‘domestic terrorists’.”<sup>34</sup>

“Trump doesn’t and can’t go all the way with his statements. He can’t because he is surrounded by Jews. It is a JEWISH media that has done this to our people.”<sup>35</sup>

---

## EXCERPTS OF REFLECTIONS

---

Promoting an accelerationist narrative, which is again of a particular declinist bent, the Siege Culture website also repeatedly reinforced negative interpretations of society as it currently exists. Theories about The Great Replacement, racial and religious differences in birth rates, and the erosion of an apparently monolithic (and superior) white culture abound :

---

“[It] is the threat of direct race-mixing which has always been that sentence of death which hangs over our heads. Death because any crossing between White and coloreds results in the end of a White line and absorption into a colored line. Death.”<sup>36</sup>

“So-called ‘democracy’ or, as I like to call it, MANAGED CONFUSION, is death to any society.”<sup>37</sup>

“The formerly White society of this country is rapidly dying and has been replaced by something evil and totally unnatural. All of this to suit an evil and unnatural agenda belonging to the very enemies of humanity itself...When will YOUR turn come? Or that of your children or grandchildren? With this evil force in command, it is only a matter of time.”<sup>38</sup>

More than just employing racial epithets and conspiracy theories, Siege Culture as a blog also pushes for redress by way of violence and the establishment of fascism governing white-only states as the alternative. The promotion of fascism comes not only in the form of promising that it would promote stability and sanity in comparison to the destructive forces of democracy at work in much of the modern West, but also by defending fascist leaders of the past. Several articles praise the Third Reich, defend Mussolini, and welcome the rise of fascist Francoism in Spain, for instance.<sup>39</sup>

Meanwhile, Siege Culture's calls for violence come with a mixture of a sense of burden or necessity and ghoulish glee. "All 'citizens' need to be treated equally. But at the bottom of all this today is the stated agenda of our RACIAL ENEMIES to mix us up with former slaves and biological SHIT-HOOKS of all descriptions in order to eliminate us as a RACE," as one article reads. "Awareness has got to be backed up by ACTION. It is no more, no less than self-defense."<sup>40</sup>

However, self-defence is far from what Siege Culture advocates support. Rather, blog posts praise state-sanctioned killings including forced euthanasia and President Duterte's extrajudicial policies regarding drug dealers. However, the text goes further, stating that in the idealised future, interracial relationships, homosexuality, and other alleged deviant behaviour would be punishable by death. Human rights are disavowed, the reasoning being: "We, as National Socialist revolutionaries, are not concerned with "legalities". We realize that the only issue is RACE!"<sup>41</sup>

---

Examples of the pro-violence bent in their writing includes, but is by no means limited to:

---

"I just love to recount the actions of the Philippine President as he sanctions the open killing of the dope addicts over there."<sup>42</sup>

"Biblically queerism carried the death penalty. It also carried the death penalty in the Third Reich. Actually, the Third Reich was most humane. They established their concentration camps to place these types into—every sort of anti-social type—first to protect the general society and, secondly, to see whether there was in existence any hope of individual salvation."<sup>43</sup>

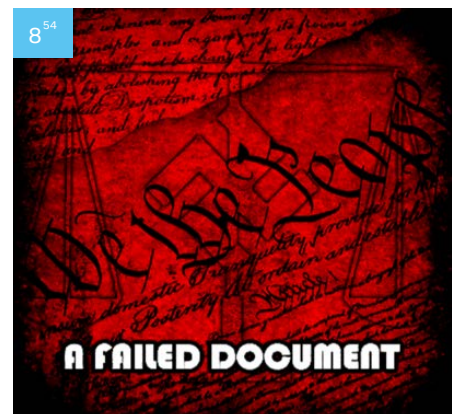
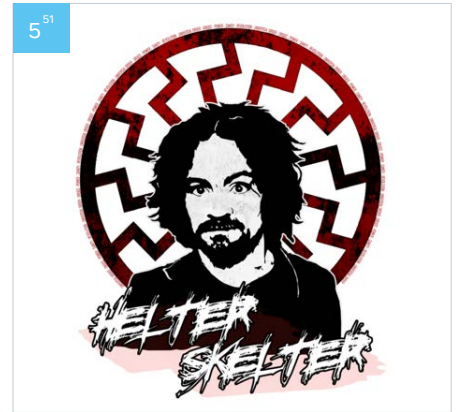
"To the dopers, no more endless waste of civil service time in resuscitating them. Let 'em die. For those incarcerated on dope charges, apart from actual dope DEALING, cut 'em loose out the front door with the warning that if they are caught again, it'll mean hanging publicly. No appeals. Colored are ordered out, period. No concerns over "legalities" or anything else. Just COLOREDS OUT! Then, after a time, should any colored face be seen within our renewed White society, it would mean another public hanging...Rights? As Hitler said, there is but ONE right as well as ONE duty and that is to keep the race pure."<sup>44</sup>

Throughout, the essential element of pushing readers to engage in violence is achieved by ubiquitous references to the idea that society is presently at a pivotal moment, and that the actions of readers could tip the scales, so avoiding the destruction of the white race and civilisation itself. A favourite phrase in Siege Culture is that of the ‘climax.’ Phraseology like “the climax is upon us,”<sup>45</sup> that “a climax approached,”<sup>46</sup> and “Any man who is a man will have wanted for a certain climax to come. Well, here it is.”

Another significant element to the website is its visual appeal to visitors, which will be discussed here before moving into an overview of Siege aesthetics more broadly. Blog posts are often accompanied by sketches and photographs. Beyond and in light of the maxim that a picture is worth a thousand words, these images warrant examination, not merely for observable patterns within a radical context, but also to consider what they are actually communicating to onlookers.

Perhaps the best example of this might be found in an article entitled ‘Strategy,’ published in September 2017:

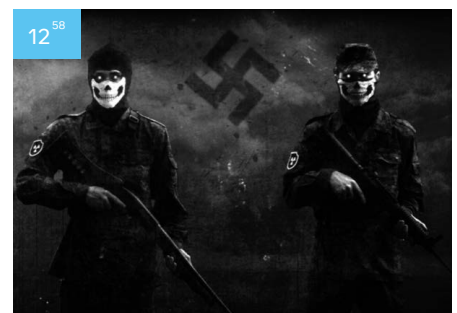
“We need to be told, just as Goebbels told the German people: ‘People arise and let the storm break loose!’...Good and brave men sacrificed everything in the name of sparking a fighting revolution in this country. In the name of the Fourteen Words they sacrificed and died. And God bless them all.”<sup>48</sup>



The website also delves into what is designed to be humorous or meme-inspired imagery (see Images 9 and 10). At times, such images feature images of celebrities, edited and marked with hate symbols, particularly if the person in question is non-white. Donald Trump, meanwhile, also features heavily in content. Relatedly, caricatures of alleged enemies are used, typically with those depicted carrying out acts of violence or otherwise playing into negative stereotypes associated with the group (for instance, Jewish people in Orthodox attire counting money or assaulting women, or black Americans with exaggerated features engaging in aggressive actions).

Finally, another scenario played out in the imagery found on the website is that of the violence to which these groups aspire. Not only are there several images of mushroom clouds denoting a kind of nuclear warfare, but there are many showing people engaged in violent beatings, riots, or other violent assaults. Meanwhile, there is a frequent use of figures in masks (typically skull masks) brandishing weapons (see Images 11 and 12). Image 12 in particular also shows the symbol of AWD on the arm-patches of two figures.

Such images advance the narrative of an emboldened, well-armed movement not content to believe in racist ideas in isolation, but actively preparing for or engaging in violence in the name of neo-Nazi principles.



Other patterns in the presentation of these images are also worth mentioning. There is, for example, a heavy reliance upon red, black, and white as the colour scheme (invoking the colours of the Third Reich flag). Meanwhile, sketches and retouched photographs overlaid with symbols are also given a kind of filter that makes them appear grainy or pixilated (see Images 13-15).

At times, this effect even makes the image difficult to decipher at first. However, this is done for dramatic effect. Such distorted images are striking, they require viewers to look closely to take in and decipher all the various components, and they intend to evoke a kind of dark, powerful energy. They are a manifestation or corollary of the dark, edgy, clandestine movement of which the creators see themselves as being a part.

Thus, when viewing the Siege Culture website, visitors experience an immersive hate-filled realm, curated by Mason and those within the movement who help him publish such material online.

---

### SIEGE CULTURE AND ITS EMBRACE BY RADICAL ACTORS

'Siege Culture' is not limited just to Mason's new writings nor to the specific Siege Culture website alone. Groups and individuals inspired by his works contribute their own takes on the belief system, publishing their interpretations on public and private message boards, on end-to-end encrypted chats, and on their own websites.

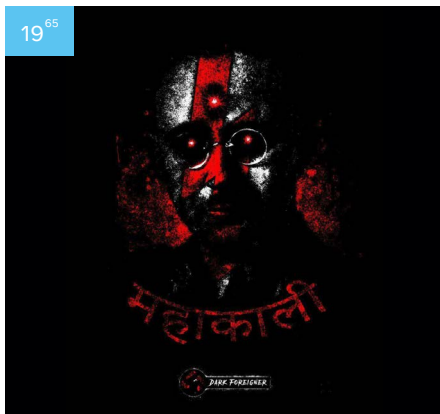
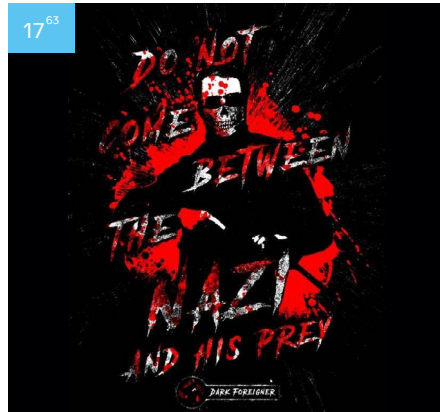
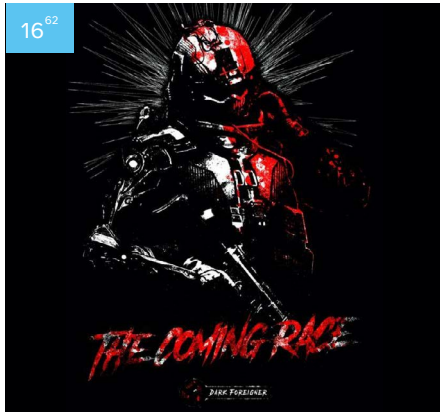
To begin with, in the realm of images, this can be seen in the graphic content produced by Dark Foreigner, whose art is widely distributed among individuals and groups who embrace Mason's ideas. This imagery can also be found in the new edition of Siege the monograph. Like the earlier images, Dark Foreigner's work again relies heavily upon a dark colour scheme, typically black, white and red, and his work tends to feature figures (sometimes famous politicians or public figures, sometimes masked men) with brief phrases accompanying them.





These images, coupled with written messages that present figures as promoting extreme violence, are only a sampling of the dozens found online (others include Osama bin Laden, Mason, Hitler, Dylann Roof, and Ted Kaczynski aka the Unabomber).

Dark Foreigner's images can be found scattered across the Internet where they are used by numerous neo-Nazi groups to recruit. AWD in particular uses his images to recruit. As Images 20 and 21 show, Dark Foreigner's images are employed in AWD propaganda, where they utilise a mixture of popular-culture references, Nazi symbols, and violent rhetoric.



AWD has had the strongest ties to Mason and Siege itself, as noted earlier in this report. It has also been the most notorious, violent international neo-Nazi organisation operating in recent years. Founded approximately half a decade ago, the American-based neo-Nazi, accelerationist group is arguably the largest and most notorious international, violent accelerationist group in operation. Within two years of its creation, it had members in almost half of the American states and more than two-dozen chapters, and quickly expanded to found cells in various European countries, as well as spin-off groups in Australia such as the Antipodean Resistance in Australia.<sup>68</sup> At their founding, AWD declared, “We are a very fanatical, ideological band of comrades [sic] who do both activism and militant training... keyboard warriorism is nothing to do with what we are.”

Additional examples of their take on issues in light of their Siege Culture ideology are statements such as: “Dead faggots couldn’t make us happier! Hail AIDS!”; “I hate hearing about ‘innocent people.’”<sup>69</sup> “There are no innocent people in this disgusting modern world.”; and “Bulldozing bodies into mass graves is the obvious solution. But in all seriousness; what re-education [sic] doesn’t fix, the sword will.”<sup>70</sup> Estimates of the number of active AWD members have been as high as 100, though its online forums have received thousands more visitors on a regular basis. Recently, however, arrests of prominent members have drastically reduced the number of known members of AWD and Mason officially disbanded the group in March 2020. However, as observers have noted, this has not led to the end of AWD, but rather its reconstitution under a new name.

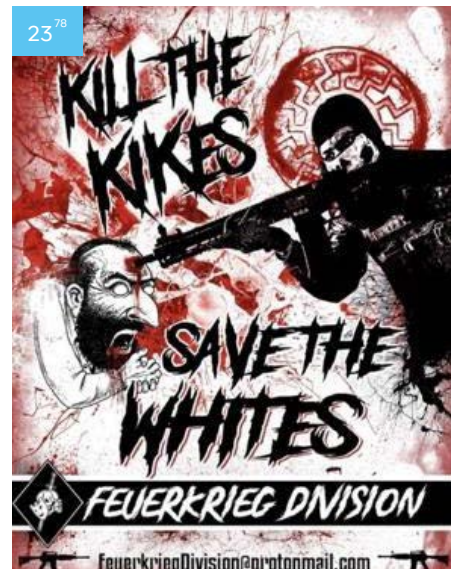
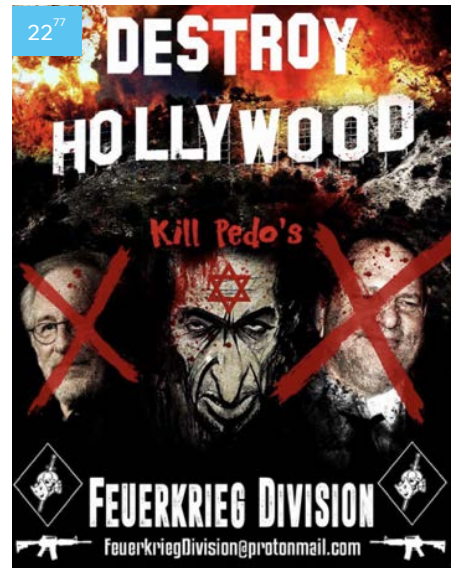
Similarly, founded in 2018 with ties to AWD, Sonnenkrieg Division (SKD) likewise has relied upon Siege Culture for its founding principles. While observers oscillate between considering SKD the European branch of AWD and its own group, the stark reality is that those professing a commitment to SKD have been convicted of serious terrorism-related crimes, largely in relation to encouraging violence. SKD is a proscribed hate group in the United Kingdom, where it is mostly known for producing an image of the Duke of Sussex being killed due to his decision to marry his biracial wife.<sup>71</sup>

Another Siege Culture-inspired group is the recently-defunct Feuerkrieg Division (FKD), another terror group proscribed in the United Kingdom.<sup>72</sup> Although allegedly led by a 13-year-old Estonian boy,<sup>73</sup> FKD had thousands of followers on their online accounts, who at times translated their discussions into real-world violence. Beyond merely believing in white supremacy,<sup>74</sup> FKD purports to support a more extreme view of racial politics. FKD has stated that it aspires to create not just an ethno-state or ethno-states, but rather an “ethnoworld” [sic].<sup>75</sup> It is easily apparent that the group is inspired by Mason, not only from their frequent posting images and quotations by or about him, but also because of statements such as the following, by a representative of the group: “we believe in the teachings of Siege and we just try to apply it in our own lives as much as possible.”<sup>76</sup>

Images such as those to the right, were also part of FKD's online presence, with individual posts on their media accounts receiving thousands of views, and their Telegram channel having more than 1,000 subscribers. Similar to AWD, following a series of arrests in late 2019 and early 2020, the group announced its official disbandment. However, all of their previous content on sites such as Telegram remain accessible. Moreover, as the UK Home Office has noted, 'it is assessed that the group and its members remain active through other channels.'<sup>80</sup>

Finally, The Base is another major Siege Culture-inspired group. Created in 2018, the group is known to have members in the US, UK, Canada, South Africa and Australia, though it has engaged in a widespread online recruitment campaign that may mean members in countries not yet known to law enforcement. As the Anti-Defamation League has observed, The Base 'advocates for some of the most extreme and violent tactics promoted within the white supremacist movement...The neo-Nazi group disseminates instruction manuals which detail specific tactics used in warfare and urban settings, including sniper attacks.'<sup>81</sup> However, for the last several months the group has frayed due to law enforcement efforts, particularly in the United States, and the leader of The Base was arrested in October 2020.<sup>82</sup>

The fraying groups, in particular AWD, are coming together in a new group with similar ideas. According to reports, the National Socialist Order was founded in July 2020 by many of the remaining radicals of these increasingly besieged groups.<sup>83</sup> Meanwhile, the harder number to gauge in relation to the influence of Siege Culture is the exact kind of radical Mason wanted to inspire: the 'lone wolf'. With Siege easily accessible online (full pdf can be found simply by typing the title and author), with even defunct groups' forums still viewable, and with no means by which to easily track who engages with Siege Culture content on end-to-end encrypted services, there are likely thousands, even tens of thousands, of individuals engaged with or who have been inspired by Siege Culture. Again, it should be noted that the Internet has been central to the proliferation of Siege Culture. Content that has fed the movement, such as that of Dark Foreigner and that written on blogs on Siege Culture or IronMarch, have brought Mason's 1980s ideas into the 21st century and to thousands more people. The Internet has made neo-Nazism accessible and made the radicalisation of people, in particular young vulnerable people, significantly easier. It has helped these groups communicate with each other and organise despite being spread across multiple states and countries, and made the transmission of not only ideology, but actual tactics (e.g. 'how to build a bomb' manuals and advice on how to stage a mass shooting) quick and easy.



It should also be noted that, while efforts have been made to de-platform obvious Siege Culture content and Siege-inspired groups from mainstream social media sites, the content is still there. The reality is that it is simply slightly more challenging for individuals to access incidentally. Searching for 'James Mason,' or 'Siege' on Facebook, for instance, does not yield results linked to neo-Nazism, as would previously have been the case. However, search for more niche terminology, such as 14 Words (a term in white supremacism), reveals a number of public and private groups and pages displaying neo-Nazi images and publishing radical content. Meanwhile, to avoid easy detection, racist messages are hidden using special symbols or images not known to content regulators. On other sites, even simple searches of hashtags of terms like 'Mason' and 'Siege' will instantly turn up messages in praise of the work, in some cases with evidence of users referring inquirers to read it or on how to further radicalise after reading the work.

---

## SIEGE CULTURE AND VIOLENCE

It is critical to understand the degree to which Siege, and the entire political philosophy of Siege Culture it has spawned, has caused real-world violence. Far from simply words on the screen or offensive images, Siege Culture has played an active role in serious violence and terrorism around the world.

For example, more than just promoting a violent political philosophy online, AWD members have been implicated or convicted in a series of violent crimes, including five murders. Both the murders and the plots law enforcement officers have been able to successfully foil to date evidence how Siege Culture is mobilising people to bloodshed. In 2017, a former AWD member murdered his two roommates, also AWD members, after his conversion to Islam.<sup>84</sup> During the course of the investigation, police officers not only discovered a fourth AWD member collecting explosive materials but also heard Arthurs confess to AWD's plans to attack synagogues, infrastructure sites, and civilian-dense locations. That same year, a 17-year-old said to have been influenced by AWD propaganda and Siege allegedly killed his girlfriend's parents after they expressed their discontent at his neo-Nazi views.<sup>85</sup> In 2018, police officers arrested a 20-year old man for the death of a gay, Jewish teenager who was stabbed 19 times in the neck.<sup>86</sup> The suspect had previously spoken with members of AWD and, following the murder, they praised his actions.<sup>87</sup>

Several other cases in 2018 and 2019 saw the convictions and/or charging of several visitors to AWD forums on charges of lying on federal background checks while attempting to acquire firearms or with possession of illegal weapons, ammunition, and drugs.<sup>88</sup> Additionally, an AWD member known to have spread content promoting violence against religious and ethnic minorities, as well as women, was arrested on child pornography charges.<sup>89</sup>

For years, it has been known that AWD has made efforts to ensure that its members can properly execute the violence they plot and praise online. For example, they managed to successfully recruit an active-duty United States Marine Lance Corporal, with video footage allegedly placing him amidst the violence of the 2017 Unite the Right rally in Charlottesville. They have also orchestrated Hate Camps, i.e. training camps, for AWD.<sup>90</sup> Law enforcement managed to disrupt these to a certain extent upon the issuing of an Extreme Risk Protection Order against one organiser, wherein they also seized several firearms based on the belief that he possessed an "imminent threat to harm others."<sup>91</sup>

2020 has particularly revealed the serious threat of AWD and its Siege Culture ideology. Massive operations by American investigators have led to the arrest and charging of several high-ranking members of AWD for plotting terrorist attacks and other acts of violence, harassment, and threatening behaviour. Alleged AWD leader Joh Cameron Denton, aka 'Rape,' was among those arrested, on charges related to 'swatting,' i.e. a dangerous harassment technique that entails calling the authorities to falsely report an emergency in the hopes that a SWAT team will arrive.<sup>92</sup> Denton and four others now face an array of federal crimes, while another four face additional charges related to harassing and threatening journalists and members of the Anti-Defamation League.<sup>93</sup> While these cases appear to have significantly hindered the hierarchy of the group and forced it to announce its disbandment, again, the NSO means that group members are not leaving the neo-Nazi movement.

Operating largely out of Europe, SKD's members have likewise been convicted of serious terrorism-related crimes, largely in relation to encouraging violence. In June 2019, courts convicted two SKD supporters of promoting terrorism in light of a threatening image of the Duke of Sussex that declared him a "race traitor."<sup>94</sup> Several months later, prosecutors filed 12 criminal charges against another member, including fundraising and promoting terrorism, in addition to owning "materials useful to terrorists."<sup>95</sup>

While nowhere near as prolific in its actualised violence relative to groups such as AWD, FKD has been linked to the plotting of serious crimes in recent months as well. A member in Lithuania has claimed to have attempted to bomb a building, though the device did not detonate.<sup>96</sup> In the United States, two alleged FKD members, Conor Climo and American Army Specialist Jarrett William Smith, have been charged and convicted of crimes linked to extremism; the former pled guilty to possessing bomb-making components,<sup>97</sup> and the latter pled guilty on two counts of 'distributing information related to explosives, destructive devices and weapons of mass destruction.'<sup>98</sup> Climo, though, claimed that he left the group because of its "inaction."<sup>99</sup> The Dutch chapter of FKD also published information about the travel plans of a member of the Green Party, with the idea that a member would use it to plot violence against him.<sup>100</sup>

Finally, in January 2020, several members of The Base were charged with crimes including conspiracy to commit murder, vandalism and intimidation, as well as gun-related offenses.<sup>101</sup> Previously, a member of The Base faced charges stemming from what was known as 'Operation Kristallnacht' (a reference to attacks on Jewish properties in Germany in 1938), a vandalism campaign that crossed US state borders.<sup>102</sup>

These events likely represent only a fraction of those crimes planned and executed as a result of influence from Siege Culture. Encouraging small and large-scale acts of violence against minorities, vulnerable communities, and influential individuals, Siege Culture contributes to the radicalisation of many people. Still readily available online and growing, it will continue to promote ideas that threaten individuals and societies, galvanising the next generation of 'lone wolves.'

---

## CONCLUSION

When Mason sat down to write the National Socialist Liberation Front newsletter, *Siege*, in 1981, he likely did not predict that it would serve as the basis for international neo-Nazi terrorist ideology propagated online in the last decade. Nevertheless, it is unmistakably a driving force behind the increasing violence on display in recent years. *Siege Culture* represents some of the most radical content available online, promoting terrorism and genocide to create patriarchal white ethno-states, or even an ethno-world. With numerous groups and individuals lauding and sharing his work in blogs, videos, and posts, Mason's influence has the potential to continue to grow (particularly during lockdown, as individuals have more time online and less time to be reminded of the benefits of living in multicultural societies).

To counteract and prevent the proliferation of Mason's message and broader *Siege Culture* ideology there needs to be a combination of efforts by the public and private sector both to shield vulnerable people from exposure to his messaging and in education that exposes the hate-filled version of the world *Siege Culture* paints as false and dangerous. Social media sites can play a critical role in this, helping stop the publication of violent images and ideas on their websites. It will require considerable effort, as *Siege Culture* proponents constantly change their behaviour and rhetoric to avoid detection and de-platforming (just as Mason adapted his work with the times). However, through assistance from experts on *Siege Culture*, such as the proposed image guide linked with this research briefing to help uncover image-based radical content as well as further investigative work on the existence of radical closed groups, it is possible to address the spread of this hate-filled ideology.

## CITATIONS

# 'Join in or else die!': Siege Culture and the Proliferation of Neo-Nazi Narratives Online

- 1 Ryan Schuster, 'Introduction,' in *Siege*, Fourth Edition, 34. \* All references to *Siege* will be taken from the Fourth Edition unless otherwise stated
- 2 United Nations Security Council Counter-Terrorism Committee, 'Member States Concerned by the Growing and Increasingly Transnational Threat of Extreme Right-Wing Terrorism,' CTED Trends Alert (April 2020), [https://www.un.org/sc/ctc/wp-content/uploads/2020/04/CTED\\_Trends\\_Alert\\_Extreme\\_Right-Wing\\_Terrorism.pdf](https://www.un.org/sc/ctc/wp-content/uploads/2020/04/CTED_Trends_Alert_Extreme_Right-Wing_Terrorism.pdf)
- 3 Schuster, 'Introduction,' in *Siege*, 14
- 4 James Mason, 'Saturation Point' [July 1984], in *Siege*, 213
- 5 James Mason, 'The Poison and the Rot' [April 1985], in *Siege*, 214
- 6 James Mason, 'It Couldn't Have Happened To a Sweeter Bunch,' [Unknown], in *Siege*, 461. See also, James Mason, 'The Most Deadly Misconception We Face,' [October 1980], in *Siege*, 158-159; James Mason, 'Dark Age,' [April 1985], in *Siege*, 226-231; James Mason, 'Thrill Kill,' [July 1985], in *Siege*; James Mason, 'Night Of The Buck Knives,' [Unknown], in *Siege*; James Mason, 'Circumstantial Constraints and Karma,' [June 1983], in *Siege*.
- 7 Mason, 'What Movement, Whose Movement?,' [April 1982], in *Siege*, 78; Schuster, 'Introduction,' in *Siege*, 13.
- 8 James Mason, 'The NSLF One-Man Army,' [January 1981], in *Siege*, 72
- 9 James Mason, 'The Whole Is Greater Than The Individual,' [October 1981], in *Siege*, 220
- 10 James Mason, 'After the Fact,' [Unknown], in *Siege*, 305
- 11 James Mason, 'Strike Hard, Strike Deep' [February 1982], in *Siege*, 75; James Mason, 'Big Brother, the System & the Establishment' [June 1982], in *Siege*, 224
- 12 James Mason, 'Strike Hard, Strike Deep' [February 1982], in *Siege*, 75
- 13 James Mason, 'What We Can Dispense With' [June 1984], in *Siege*, 187
- 14 James Mason, 'By Accident or Design' [July 1982], in *Siege*, 432
- 15 James Mason, 'Terrorism Redefined' [November 1981], in *Siege*, 97
- 16 James Mason, 'Big Brother, the System & the Establishment' [June 1982], in *Siege*, 223
- 17 James Mason, 'The Simplistic Society' [July 1982], in *Siege*, 224
- 18 James Mason, 'The Enemy is Anyone Who Attacks' [January 1984], in *Siege*, 205
- 19 James Mason, 'Men of Ill Repute' [July 1982], in *Siege*, 415
- 20 James Mason, 'Of Victims and Statistics' [October 1985], in *Siege*, 243-244
- 21 James Mason, 'Alienation' [June 1982], in *Siege*, 339
- 22 To be found at the now defunct link: <https://siegekultur.biz>
- 23 Hatewatch Staff, 'The racist "alt-right" is killing people.,' Southern Poverty Law Center, <https://www.splcenter.org/hatewatch/2018/02/22/atomwaffen-and-siege-parallax-how-one-neo-nazi's-life-s-work-fueling-younger-generation>
- 24 James Mason, 'Identity,' *Siege Culture*, undated
- 25 James Mason, 'Eugenics and Euthanaisa,' *Siege Culture*, undated
- 26 James Mason, 'A New Media,' *Siege Culture*, 11 October 2017
- 27 James Mason, 'If This Were A Hundred Years Ago,' *Siege Culture*, undated
- 28 James Mason, 'Islam,' *Siege Culture*, undated
- 29 James Mason, 'If This Were A Hundred Years Ago,' *Siege Culture*, undated
- 30 Mason, 'Islam,' *Siege Culture*
- 31 James Mason, 'An American Nazi Looks at "The Holocaust",' *Siege Culture*, 29 September 2017
- 32 James Mason, 'Why Does the System So Hate "Fascism?",' *Siege Culture*, 13 September 2017
- 33 James Mason, 'Don't Make the Mistake,' *Siege Culture*, undated
- 34 James Mason, 'A New Media,' *Siege Culture*, 11 October 2017
- 35 Mason, 'Islam,' *Siege Culture*
- 36 James Mason, 'Don't Make the Mistake,' *Siege Culture*, undated
- 37 Mason, 'Islam,' *Siege Culture*
- 38 James Mason, 'If This Were A Hundred Years Ago,' *Siege Culture*, undated
- 39 For a defense of fascism in long-form see, James Mason, 'Why Does the System So Hate "Fascism?",' *Siege Culture*, 13 September 2017
- 40 James Mason, 'It's Not Too Late to Hate,' *Siege Culture*, undated
- 41 James Mason, 'Trump,' *Siege Culture*, 25 July 2017
- 42 James Mason, 'Those Who Want to Live,' *Siege Culture*, undated
- 43 Mason, 'Those Who Want to Live,' *Siege Culture*
- 44 Mason, 'Those Who Want to Live'
- 45 James Mason, 'We're All in the Same Boat Now,' *Siege Culture*, undated
- 46 Mason, 'Why Does the System So Hate "Fascism?",' *Siege Culture*, 13 September 2017
- 47 James Mason, 'Whole Lotta Race Mixing Goin' On,' *Siege Culture*, undated
- 48 The Fourteen Words refer to the infamous 14 words central to white supremacism, 'We must secure the existence of our people and a future for white children.' James Mason, 'Strategy,' *Siege Culture*, 20 September 2017
- 49 James Mason, 'It's Not Too Late To Hate,' *Siege Culture*, undated
- 50 James Mason, 'That's Mighty White of You,' *Siege Culture*, undated
- 51 SIEGE, 'Pyramid Upside-down: The Passing of Charles Mason,' *Siege Culture*, undated
- 52 Vincent Snyder, 'No Response?,' *Siege Culture*, undated
- 53 Vincent Snyder, 'The Movementarian Meance,' *Siege Culture*, undated
- 54 James Mason, 'The Constitution: A Failed Document?,' *Siege Culture*, undated
- 55 James Mason, 'The Media,' *Siege Culture*, 30 August 2017
- 56 James Mason, 'Me Three,' *Siege Culture*, undated
- 57 Mason, 'That's Mighty White of You,' *Siege Culture*
- 58 James Mason, 'Russia,' *Siege Culture*, 03 August 2017
- 59 James Mason, 'Don't Make the Mistake,' *Siege Culture*, undated

- 60 SIEGE, 'American Futurism Workshop,' Siege Culture, undated
- 61 SIEGE, 'A New SIEGE Culture,' Siege Culture, undated
- 62 Dark Foreigner, 'The Coming Rage,' <https://www.deviantart.com/thedarkforeigner/art/The-Coming-Race-726805693>
- 63 Dark Foreigner, 'Don't Come Between the Nazi and His Prey,' <https://i.4pcdn.org/pol/1529329625286.jpg>
- 64 Dark Foreigner, 'Beyond Humanity,' <https://i.4pcdn.org/pol/1522936898689.jpg>
- 65 Dark Foreigner, 'Untitled,' <https://www.pinterest.co.uk/pin/686306430695012717/>
- 66 Atomwaffen Division, 'Siege the Fucking System, Introduce a Little Anarchy,' [https://www.reddit.com/r/PropagandaPosters/comments/bnryri/poster\\_for\\_the\\_neonazi\\_terrorist\\_organization/](https://www.reddit.com/r/PropagandaPosters/comments/bnryri/poster_for_the_neonazi_terrorist_organization/)
- 67 'Atomwaffen Division (AWD),' Anti-Defamation League, <https://www.adl.org/resources/backgrounders/atomwaffen-division-awd>
- 68 Intel Brief, 'Examining Atomwaffen Division's Transnational Linkages,' The Cypher Brief, 20 May 2020
- 69 'Atomwaffen Division,' Southern Poverty Law Center, <https://www.splcenter.org/fighting-hate/extremist-files/group/atomwaffen-division>
- 70 'Atomwaffen Division,' Southern Poverty Law Center, <https://www.splcenter.org/fighting-hate/extremist-files/group/atomwaffen-division>
- 71 Home Office, 'Proscribed Terrorist Organisations,' [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/901434/20200717\\_Proscription.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/901434/20200717_Proscription.pdf)
- 72 Home Office, The Rt Hon James Brokenshire MP, and The Rt Hon Priti Patel MP, 'Order to proscribe far-right terrorist group comes into force,' 28 February 2020, <https://www.gov.uk/government/news/order-to-proscribe-far-right-terrorist-group-comes-into-force>
- 73 Daniel De Simone, 'Neo-Nazi group led by 13-year-old boy to be banned,' BBC News, 13 July 2020, <https://www.bbc.com/news/uk-53392036>
- 74 The Anti-Defamation League defines as believing in the genetic and cultural superiority of the white race, and thereby the right of white people to either live in racially homogenous societies or rule over non-white people. Anti-Defamation League, 'White Supremacy,' <https://www.adl.org/resources/glossary-terms/white-supremacy>
- 75 Feuerkrieg Division \*\*Official\*\*, Telegram post, 12 October 2019
- 76 Subcomandante X, 'Feuerkrieg Division Member Talks About New Group on Far-Right Podcast,' 08 January 2019, <https://medium.com/americanodyssey/feuerkrieg-division-member-talks-about-group-on-far-right-neo-nazi-podcast-314086068acf>
- 77 Feuerkrieg Division \*\*Official\*\*, Telegram post, 19 November 2019
- 78 Feuerkrieg Division \*\*Official\*\*, Telegram post, 11 December 2019
- 79 Feuerkrieg Division \*\*Official\*\*, Telegram post, 9 August 2019
- 80 Home Office, 'Proscribed Terrorist Organisations,' [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/901434/20200717\\_Proscription.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/901434/20200717_Proscription.pdf)
- 81 'The Base,' Anti-Defamation League, <https://www.adl.org/resources/backgrounders/the-base>
- 82 Nicholas Bogel Burroughs, 'F.B.I. Arrests Michigan Men Tied to White Supremacist Group,' The New York Times (29 October 2020), <https://www.nytimes.com/2020/10/29/us/fbi-arrests-the-base-michigan.html>
- 83 Ben Makuch, 'Neo-Nazi Terror Group Atomwaffen Division Re-Emerges Under New Name,' Vice, 05 August 2020, <https://www.vice.com/en/article/wxq7jy/neo-nazi-terror-group-atomwaffen-division-re-emerges-under-new-name>
- 84 Jonah Engel Bromwich, 'Man in Florida Told the Police He Killed Neo-Nazi Roommates for Disrespecting His Muslim Faith,' The New York Times, 24 May 2017, <https://www.nytimes.com/2017/05/24/us/neo-nazi-roommate-murder.html?login=email&auth=login-email>
- 85 Justin Jouvenal, 'Va. teen accused of killing girlfriend's parents to be tried as an adult,' The Washington Post, 25 September 2019, [https://www.washingtonpost.com/local/public-safety/va-teen-accused-of-killing-girlfriends-parents-to-be-tried-as-an-adult/2019/09/24/3e628fae-af13-11e9-a0c9-6d2d7818f3da\\_story.html](https://www.washingtonpost.com/local/public-safety/va-teen-accused-of-killing-girlfriends-parents-to-be-tried-as-an-adult/2019/09/24/3e628fae-af13-11e9-a0c9-6d2d7818f3da_story.html)
- 86 Emily Shapiro, '1 year after Blaze Bernstein's killing, parents look to turn alleged hate crime into 'movement of hope,' ABC News, 30 December 2018, <https://abcnews.go.com/US/year-blaze-bernstains-killing-parents-turn-alleged-hate/story?id=59754707>
- 87 A.C. Thompson, Ali Winston, and Jake Hanrahan, 'Inside Atomwaffen As It Celebrates a Member for Allegedly Killing a Gay Jewish College Student,' Pro Publica, 23 February 2018, <https://www.propublica.org/article/atomwaffen-division-inside-white-hate-group>
- 88 'Atomwaffen Division (AWD),' Anti-Defamation League, <https://www.adl.org/resources/backgrounders/atomwaffen-division-awd>
- 89 'Atomwaffen Division (AWD),' Anti-Defamation League, <https://www.adl.org/resources/backgrounders/atomwaffen-division-awd>
- 90 'Atomwaffen Division (AWD),' Anti-Defamation League, <https://www.adl.org/resources/backgrounders/atomwaffen-division-awd>
- 91 'Atomwaffen Division (AWD),' Anti-Defamation League, <https://www.adl.org/resources/backgrounders/atomwaffen-division-awd>
- 92 'Sweep of arrests hits US neo-Nazi group connected to five murders,' The Guardian, 06 March 2020, <https://www.theguardian.com/world/2020/mar/06/neo-nazi-arrests-deals-blow-us-group-atomwaffen-division>
- 93 'Sweep of arrests hits US neo-Nazi group connected to five murders,' The Guardian
- 94 Richard Ford, 'Terror ban for Sonnenkrieg Division neo-Nazis who branded Prince Harry a race traitor,' The Times, 24 February 2020, <https://www.thetimes.co.uk/article/terror-ban-for-sonnenkrieg-division-neo-nazis-who-branded-prince-harry-a-race-traitor-09zf3l3tb>
- 95 'Atomwaffen Division (AWD),' Anti-Defamation League, <https://www.adl.org/resources/backgrounders/atomwaffen-division-awd>
- 96 'Feuerkrieg Division Exposed: International Neo-Nazi Terrorist Network,' Eugene Antifa, <https://eugeneantifa.noblogs.org/post/2020/02/24/feuerkrieg-division/>
- 97 'Las Vegas Man Pleads Guilty To Possession Of Bomb-Making Components,' United States Department of Justice, U.S. Attorney's Office, District of Nevada, 10 February 2020, <https://www.justice.gov/usao-nv/pr/las-vegas-man-pleads-guilty-possession-bomb-making-components>
- 98 Phil Helsel, 'Soldier who discussed attack in U.S. pleads guilty to distributing bomb instructions,' NBC News, 11 February 2020, <https://www.nbcnews.com/news/us-news/army-soldier-who-discussed-attack-u-s-pleads-guilty-distributing-n1134571>
- Case 2:19-cr-00232-JCM-NJK Document 1 Filed 08/09/19 US District Court, <https://www.courtlistener.com/recap/gov.uscourts.nvd.139164/gov.uscourts.nvd.139164.1.0.pdf>
- 99 Intel Brief, 'Examining Atomwaffen Division's Transnational Linkages,' The Cipher Brief, 20 May 2020, [https://www.thecipherbrief.com/column\\_article/examining-atomwaffen-divisions-transnational-linkages](https://www.thecipherbrief.com/column_article/examining-atomwaffen-divisions-transnational-linkages)
- 100 Neil MacFarquhar and Adam Goldman, 'A New Face of White Supremacy: Plots Expose Danger of the 'Base',' The New York Times, 22 January 2020, <https://www.nytimes.com/2020/01/22/us/white-supremacy-the-base.html>
- 101 'The Base,' Anti-Defamation League, <https://www.adl.org/resources/backgrounders/the-base>



---

RESEARCH BRIEFINGS

---

# 04 Automatic Hate Speech Detection: Challenges and Opportunities

---

by Lisa Kaati, Swedish Defence Research Agency,  
Nazar Akrami and Amendra Shrestha,  
Department of Psychology, Uppsala University

EXTRACTS FROM:

Swedish Defence Research Agency  
and Uppsala University

---

Hateful and harmful messages are widespread in social media platforms and include content that, for example, incites violence, expresses direct hate towards individuals and groups, sexual exploitation of children, extremist propaganda, as well as content that promotes self-harm or suicide. For the target, such content may cause everything between emotional distress and inspiring individuals to commit mass murder.<sup>1</sup> The problem of hate speech is not specific or limited to a certain territory or culture – it is widespread all over the internet.

---

## INTRODUCTION

Politicians and policymakers around the world now and then call for measures to regulate the content of the internet, especially what individuals and groups write/post online. The idea of controlling what is written online is brought to the fore in connection to deadly terror attacks,<sup>2</sup> violent riots<sup>3</sup> and expressions of hate speech that reaches the media.<sup>4</sup> While the idea of control is supported by some, others believe that a regulation obstructs basic fundamental rights in full democracies including the freedom of speech.<sup>5</sup>

Since it is impossible to manually monitor everything that is said online, the research community is dealing with this issue by exploring technical solutions to detect hateful and harmful communication. However, there are several challenges in detecting hateful and harmful messages, henceforth denoted as hate speech (see Defining Hate Speech, below). One challenge is that people perceive hate speech differently depending on their backgrounds and knowledge and depending on the context in which a message is expressed or disseminated. Another challenge is that hate speech can be expressed in many, sometimes innovative, ways, for example, by avoiding certain words, by alternative spellings, and/or by replacing letters with numbers/symbols.

Despite the complexity and difficulties to detect, there have been several attempts to develop effective technical solutions for automatic detection of hate speech. These attempts include the use of technologies such as machine learning (ML) and natural language processing (NLP). In many situations, automatic solutions are necessary due to the massive amount of data available. Automatic solutions are used by a number of actors. For example, social media companies may use automatic methods to detect messages that violate their policies or that violate the law, while law-enforcement authorities may use similar technologies to detect threats towards the safety of individuals and groups in society. It should be noted that automatic methods are not welcomed by everyone and concerns about the utility of such solutions are expressed on a regular basis. It is, for example, argued that these solutions can be easily deceived, that they lack validity and reliability, and that they might reinforce biases.<sup>6</sup>

The aim of this paper is to provide a brief review, discuss key issues, and identify some of the challenges and opportunities in automatic hate speech detection. Specifically, the paper will focus on aspects that we find important for the study of hate in online environments and when developing technologies for automatic hate speech detection. The aspects that we will focus on are:

- 01 Definitions of hate speech
- 02 Antecedents of hate speech
- 03 Technology for detecting hate speech
- 04 Challenges in hate speech detection

---

## DEFINITIONS OF HATE SPEECH

One of the main issues when dealing with hate speech is the lack of a common international definition of what is considered to be hateful. The term hate speech is considered to be vague, broad, and there are a variety of different definitions among scholars. Also, classifications of hate speech are sometimes controversial, subject to dispute, and provoke moral reactions among legal authorities, policymakers as well as the general public.<sup>7</sup> Thus, disagreement about definitions of hate speech is common but definitions have been converging more recently. One of the most comprehensive definitions is that by the United Nations, defining hate speech as “any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor”. Similarly the European Union defines hate speech as: “all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin”. Interestingly, definitions of hate speech are also consistent with definitions of related phenomena, specifically prejudice and biases: “An antipathy based upon a faulty and inflexible generalization. It may be felt or expressed. It may be directed toward a group as a whole, or toward an individual because he is a member of that group”. Ashmore identified four basic points of agreement common to most definitions of prejudice. These are: (a) Prejudice is an intergroup phenomenon. (b) Prejudice is a negative orientation. (c) Prejudice is bad. (d) Prejudice is an attitude. These basic points of agreement hold also for hate speech. Indeed, hate speech is considered to be a key factor in prejudice and intergroup hostility.

While most definitions of hate speech share some common elements, alternative terminology has been introduced to either broaden or to narrow the definition. Examples of such terminology are abusive language, toxic language, and dangerous speech. Social media companies and online platforms have developed their own definitions of hate speech to moderate user-generated content. In Facebook’s policy against hateful content, hateful content is defined as “content that directly attacks someone because of what we call protected characteristics - skin color, ethnicity, national origin, religious affiliation, sexual orientation, caste, gender, gender identity, and serious illness or disability”. Twitter’s rules for hateful conduct states; “You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease”. It is also common that definitions are updated to capture new forms of hate speech. For example, in August 2020 Facebook updated their policy against hateful content to include Jewish conspiracy theories and caricatures of people of African descent in the form of “blackface”.

---

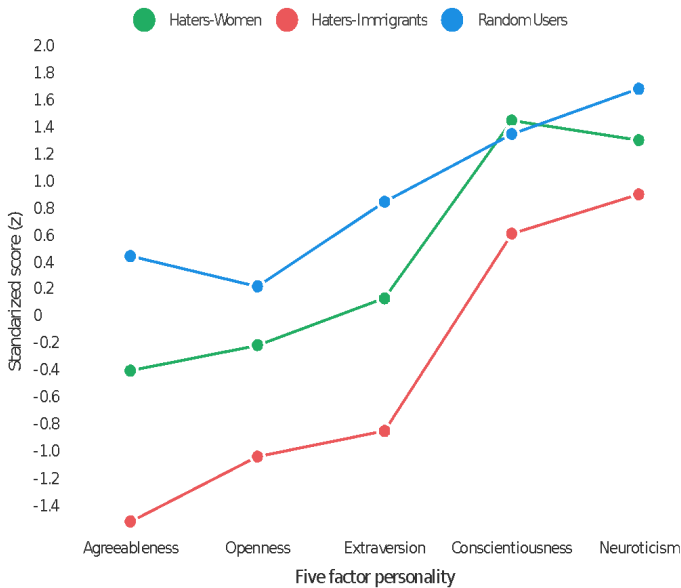
## ANTECEDENTS OF HATE SPEECH

While there are many studies exploring the definition, detection, and consequences of hate speech, very few have paid attention to the antecedents of the phenomenon. Why do some people engage in hate speech? Before answering the question we would need to go back to the similarities of the definitions of prejudice and hate speech we mentioned earlier – the link between prejudice and hate speech. The idea of this link is not new. Some scholars suggest that hate speech is the equivalent term of antilocution. The term which was coined by Allport and is described as follows: “Most people who have prejudices talk about them. With like-minded friends, occasionally with strangers, they may express their antagonism freely. But many people never go beyond this mild degree of antipathetic action”. Thus, Allport considered hate speech as a way of acting out prejudice. Moreover, explaining the link between prejudice, discrimination, and violence Allport wrote “When antilocution reaches a high degree of intensity, the chances are considerable that it will be positively related to open and active discrimination, possibly to violence”. Now, instead of asking why some people engage in hate speech, we could ask; why some people engage in prejudice? The answer to the question is much easier considering the large body of literature dealing with the matter.

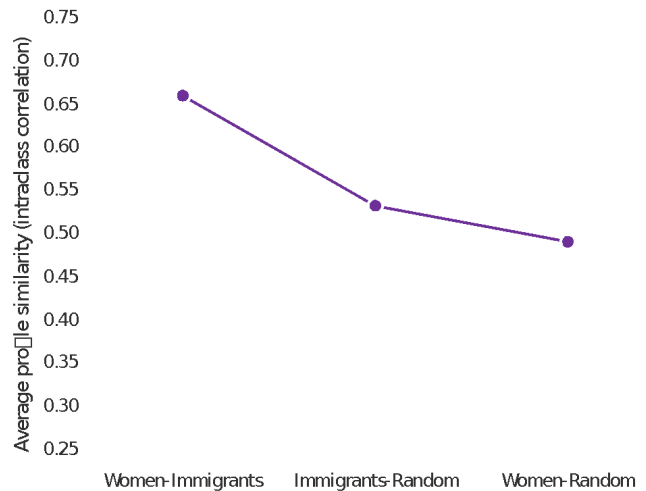
Psychological research on the antecedents of prejudice provides two major explanations – one referring to social factors and the other to individual factors. Research dealing with social factors explaining prejudice has identified group factors such as group membership and group identity and situational factor such as social norm and social threat as key predictors of prejudice. The research on individuals factors shows that core personality traits like agreeableness and openness to experience explain why some individuals engage in prejudice while others not. Research has also found that the dark personality variables (narcissism, Machiavellianism, psychopathy) are predictors of prejudice beyond the above-mentioned personality factors.

While this research above has been conducted focusing on experimental and survey data and traditional expressions of prejudice, some recent studies have examined the antecedents prejudice (/hate speech) by studying online activities. For example, Sorokowski and colleagues, surveyed users who wrote hateful comments online and compared these with a control group and found that the former scored significantly higher on the psychopathy. Another study examining deviant online behavior found that trolling was significantly associated with higher sadism, psychopathy, and Machiavellianism, with sadism having the

most robust associations. A more recent study by Akrami and colleagues examined writings by users on the largest discussion forum in Sweden. First, we selected all posts on the discussion forum that expressed hate towards women and immigrants. Next, the 30 most hateful users from each category (women & immigrants) were compared with 30 randomly selected users as to their mean scores on the personality factors agreeableness and openness to experience and similarities in profile (assessed by intraclass correlation) on the big-five personality factors (openness to experience, conscientiousness, extraversion, agreeableness, neuroticism). The hateful messages and the personality were detected by machine learning models. Figure 1 (below) shows the average personality scores of the three different groups. As can be noted in Figure 1, the two groups of haters (towards immigrants and women) scored significantly lower on two personality factors, namely agreeableness and openness, compared to the control group. Figure 2 (below) shows the similarity between the three different groups. Interestingly, the two groups of haters (toward women and towards immigrants) were significantly more similar to each other, compared to their similarities to the control group. Thus, on average, users expressing hate toward women had similar profiles as these expressing hate toward immigrants.



**Figure 1** The personality of users expressing hate toward women and immigrants compared to a randomly selected group.



**Figure 2** Average profile similarity (measured by intra-class correlation) between the users expressing hate toward women and the users expressing hate toward immigrants and a set of randomly selected users.

Finally, we also combined the data users from all three groups and conducted multiple regression analyses to examine what personality factors explain the average hate level. In line with studies on prejudice, we found agreeableness and openness to be the only significant predictors.

Together, these findings reported above reveal not only some explanations to online deviant behavior like hate speech (and trolling), but also connect to the traditional literature of prejudice and its explanations to that of hate speech. However, despite their coherence with previous research on prejudice and personality, these findings should be considered with care due to the uncertainty when using machine learning models for measuring personality.

---

## TECHNOLOGY FOR DETECTING HATE SPEECH

With the constant increase of data on social media platforms and the internet generally, automated approaches to detecting hate speech have become an absolute necessity. The majority of automated approaches for identifying hate speech try to classify a piece of text as hateful or not hateful. In some cases, the text is classified into a specific type of hate speech such as misogyny, antisemitism, xenophobia, abusive language, or threats. Most automated hate speech detection technologies rely on natural language processing or text mining technologies. The simplest of these approaches are dictionary-based methods, which involve developing a list of words that if a word from the dictionary is present in a text – the text is considered hateful. Similarly, if no word from the dictionary is present, the text is considered to be not hateful. The dictionaries generally contain content words, including insults and slurs. Some dictionary-based methods also include linguistic rules to increase performance.

Dictionary-based methods are easy to understand, but there are several drawbacks to such approaches. One is that the meaning of words can be context-dependent, which means that words may have several different meanings depending on the context, which was noticed by Mehl, Robbins, & Holleran. Another criticism of dictionary-based approaches is that the dictionaries are defined a priori, without any consideration of the properties of the actual data. This makes the detection sensitive to vocabulary variation that is introduced by slang words, different spellings, domain-specific terminology, and that the same word can have different meanings in different environments. Yet another

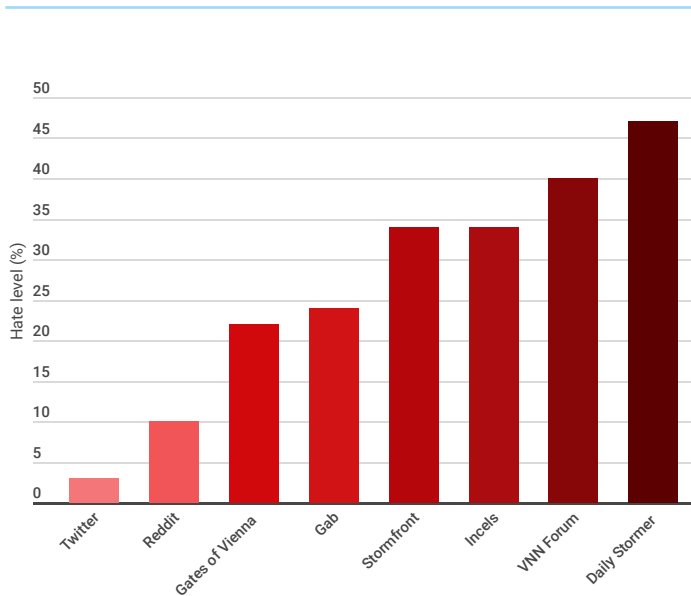
issue is that dictionary-based approaches cannot detect sarcasm or humor. Moreover, the dictionaries require constant updates since new terminology and slang tend to develop quickly in social media.

State-of-the-art hate speech detection techniques involve supervised text classification tasks, which means that we need to provide relevant training examples of what is considered hate and what is not. The training examples are used to learn a machine learning model to recognize hate speech automatically. Several different algorithms can be used to create a machine learning model that can be used for hate speech detection. The algorithms use either predefined features or learn features automatically. Features could either be keyword-based as in or based on the training data with more or less sophisticated representations that include word n-grams, syntactic features, and distributional semantics. With the recent advancement in text analysis, the use of deep learning-based technologies has become a leading approach for automatic hate speech detection. Deep learning methods use advanced linguistic features that are extracted automatically from the training data. To train a machine learning model for hate speech detection high quality training data is needed. It is well-known that both the quality and the content of the training data highly affect the performance of the algorithms. The choice of datasets is one of the key factors in the development of automatic hate speech detection technologies. There are a number of datasets that have been used for developing technology for hate speech detection. Hate speech data provides a list of more than 60 different datasets on various languages that are available for training hate speech models.

To illustrate how technology for hate speech detection can be used in studies of online environments, we have used a machine learning model described in. The model is based on a technique called transfer learning and can recognize hate speech with accuracy slightly above 80 percent and classify hundreds of texts per second. We used the model to determine the hate level in eight different online environments: Twitter, Reddit (one of the most visited places on the internet), Gates of Vienna (a platform for the counter-jihad movement), Gab (freedom of expression friendly social network where users can write messages of up to 300 characters), Stormfront (one of the most well-known and oldest white supremacy discussion forums), Incels (the largest active digital environment for individuals that lives in involuntary celibacy), VNN Forum (a white

supremacist discussion forum) and Daily Stormer (a white supremacy and anti-semitic news site). For each digital environment, we selected a representative sample of posts, sized to achieve a margin of error below 1% and a confidence level above 99%. Each post was classified as hate or not hate, and the hate level is the percentages of hateful posts in the total sample set. The hate level allows us to compare several different environments, even if the measurement method is not fully reliable.

The hate level for each environment is shown in Figure 3. As can be noted, the hate level on Twitter is around 3%, and on Reddit, the level is just above 10%. For the other environments, the hate level varies between 22% and 48%. Daily Stormer has the highest hate level – almost half of the site’s posts are hateful according to our measurement. The differences in the level of hate in the different platforms probably depend on the user rules for the platform (and the level of moderation), the users’ characteristics, and the topics discussed.



**Figure 3** The level of hate in a set of different digital environments.

## CHALLENGES IN HATE SPEECH DETECTION

### Domain transferability

One of the challenges with automatic hate speech detection is to understand what we can expect when it comes to the performance of the technologies. Most automatic hate speech detection technologies are trained, evaluated and tested on similar data. When the performance of hate speech detection algorithms is described we can only draw conclusions on the performance when we use the algorithms on data that is very similar to the data that algorithm was designed to operate on. Gröndahl and his colleagues showed that several hate speech detection approaches work well when training and testing are based on the same dataset but the results are not transferable across datasets. This means that when the techniques are used on new data or data from a different domain, the performance may differ significantly from what we expect.

To understand how different algorithms can be used to identify and assess hate speech there is a need to test and evaluate the developed methods not only on realistic unseen data but also data from different domains. To illustrate the problem of performance of algorithms we used a machine learning model that was created to detect hate speech in Swedish. The model was trained on data that was annotated by a set of experts – in this case represented by 12 final-year psychology students. The 12 students were selected with the assumption that their training made them more qualified than random people to judge if a comment was hateful or not. The choice of expert annotators was done with the aim to achieve high reliability. The definition of hate that we used was broad and the students were instructed to score a given text according to the following scale:

- 3** = Aggression or disgust toward an individual, a group, an organization, or a cause
- 2** = Malice or insults toward an individual, a group, an organization, or a cause
- 1** = Dislike of a specific individual, a group, or an organization
- 0** = Texts that were not hateful or applicable to any of categories above

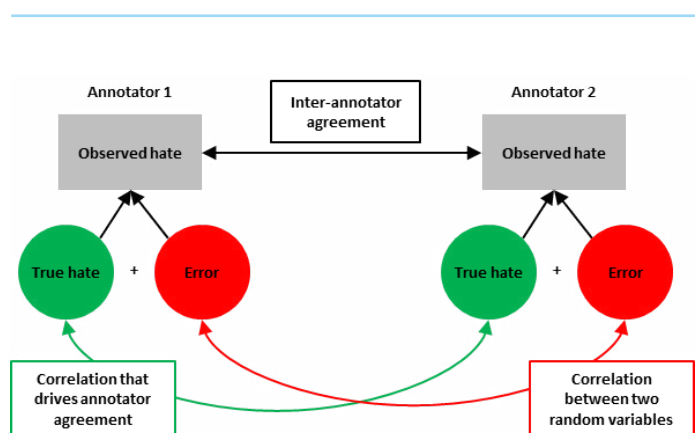
After some considerations to deal with reliability shortcomings, texts marked with 3 and 2 were considered being hate speech while texts marked with 1 and 0 were considered not hateful. At least two students annotated each text. The final inter-annotator agreement when merging the classes/response categories were calculated using the Krippendorff's Alpha and was .90, which can be considered exceptionally high (for a discussion on the reliability of hate speech data, see Ross et al., 2016). We used around 3000 texts to train a machine learning model to recognize hate speech. An evaluation of the model showed that over 72% of the hateful comments were correctly classified as hate, and 87% of the non-hateful comments were correctly classified as not hateful. This outcome is inline with previous work on hate speech detection. However, when testing the model on a dataset consisting of 600 comments from the largest discussion board in Sweden, the result showed that the performance of the model decreased significantly. Slightly over 50% of the hateful comments were correctly identified as hateful, and 75% of the non-hateful comments were correctly classified as not hateful. This shows that the performance decreased by more than 20% when it comes to recognizing hate speech and by 12% for recognizing non-hateful comments. The significant decrease shows that using machine learning models on new data impaired the performance of the model. The possible differences in performance when using a pre-trained model on new data should be taken into consideration when a method for automatic detection of hate speech is selected.

Despite some challenges, there are also some benefits when using domain-specific models. Algorithms that are specifically designed for detecting certain kinds of hate on specific domains could be a technical complement to assist subject matter experts in their analysis. Such algorithms already exist and are used in specific domains, for example, in the case of child sexual abuse material where hashes are used to fingerprint material. A hash is created by a mathematical algorithm that transforms data of any size into much shorter fixed-length data. The hashes are stored in a shared database, and new material can quickly be matched against hashes in the database to determine if it is child sexual abuse material.

## Reliability

While some of the challenges with automatic hate speech detection are already mentioned, there is yet another challenge that needs to be considered – reliability. Since moststate-of-the-art methods for detecting hate speech need training data, one of the challenges with improving the methods is obtaining high-quality training data. The reality is that, now and then, we encounter calls for tools to detect online hate speech as an aftermath of observation of hate speech when models fail. One way of facilitating the process of improving automatic hate speech detection is to put more focus on the aspect of reliability, or inter-annotator agreement. Regardless of what machine learning technique is used, models seem to be hampered by datasets that lack reliability. If humans can't agree on what should be classified as hate speech, we can't expect an algorithm to make the right decision on what should be classified as hate speech.

We illustrate the problem of reliability by using classical test theory (also True Score Theory). Classical test theory is a theory of measurement assuming that every observed score (measured by measuring tape or assessed by an annotator) constitutes of a true score component and an error component (see figure 4). The theory also assumes that the error component can include a random and a systematic error component. The theory can help in understanding the implication of reliable measurement but also identify sources of error in the context of hate speech detection.



**Figure 4** An illustration of the correlation between scores of two annotators as a function of true and random scores. Double-arrows lines denote correlation while single-arrow lines denote contribution to a score.

The basic idea is that agreement measures result from the degree of overlap between the observed scores of, for example, two annotators. This overlap is underpinned by the two components (true and error) of each observed score. However, as two random components (variables) cannot correlate systematically, the overlap is a reflection of the true scores, observations of true hate speech in our case. The key question here is to identify and decrease the random error and boost the true component. By doing so, we would be able to improve reliability. While some aspects of the classical test theory are mathematical, the process of identifying sources of error is theoretical. It is important to emphasize that our aim here is to provide a general framework and some examples where errors can be reduced. The idea is that a part of social media communication includes hate speech, and the task is to identify that amount of hate speech and to do so with high reliability.

Starting with the definition, we argue that the definition is a major source of unreliability. The definition (see the definition section above) provides room for individual differences in how the hate is perceived. For example, take the reference to communication that “attacks or uses pejorative or discriminatory language”. What is perceived as an attack or provocation by one individual can be seen as simple criticism by another. Thus, it is necessary to provide the annotators with precise definitions of all included terms to minimize error due to different perceptions. Another source of error, misunderstanding, or disagreement lies in the character of human language. Human language is a highly creative and complex system of expression with endless variations. The same system that enables us to hurt others by saying a single word also provides us with ways to generate the same hurt feeling with seemingly a positive expression. Annotators would need training but also share and discuss examples of various ways of expressing hate in order to synchronize their assessments and thus avoid error. Moreover, in their everyday life, people are guided by moral/ethical as well as political/ideological principles. While some consider expressing what they think about a group of people to be ethically motivated, others engage in action against hate speech with reference to their moral principles. Also here training is necessary in order to avoid subjective judgments and thereby different interpretations of the same term/definition and contribute to error. Thus, it is generally beneficial to set boundaries for the definition and include terminology by providing training and examples in order to minimize sources of error.

Several studies mention the problem of creating training data with high reliability of annotations of hate speech. Since we can't expect any algorithm to perform better than humans the reliability of human annotations becomes a higher bound for the performance of the algorithm. Ross et al. examined the reliability of annotations of hate speech by letting two groups of internet users determine whether a text was hate speech or not. One of the groups was shown a definition of hate speech before their annotations which lead to that they partially aligned their opinion with the definition. Still, Ross et al. found annotator agreement in the range of 0.18 to 0.29, as assessed by Krippendorff's alpha, which is very low considering the recommended level of 0.8 (or 0.667 in cases where some uncertainty is allowed, see Krippendorff, 2004). Ross et al. suggests that hate speech should be seen as a continuous rather than as a binary problem and that detailed instructions for the annotators are needed to improve the reliability of hate speech annotation. Similar conclusions are drawn by Laaksonen and her colleagues who noted that annotators own knowledge of the issue influenced their annotations. These findings indicate that domain expertise among annotators is important and that using subject matter experts to annotate data would most likely improve the quality of data.

### Languages, expressions, and how to avoid detection

Apart from the challenges with the expectations and reliability mentioned above, there are yet some other challenges when developing technologies for identifying hate speech. One of them is languages. While there are several technical approaches for detecting hate speech in English, there are very few approaches for low resource languages such as Swahili, Tagalog, Somali, Vietnamese, and Swedish. There are some crowd-sourced multilingual dictionaries that can be used for hate speech detection (e.g. Hatebase and the Racial Slur Database), but there is still a lack of high-quality datasets in multiple languages that can be used to develop more sophisticated methods for hate speech detection.



Language is not the only problem in hate speech detection – the specific language and expressions among subcultures, movements, and groups introduce challenges too. Hate speech can be expressed very differently depending on the domain. Some groups and subcultures have developed a very domain-specific jargon and use alternative spellings and words to express hate. The jargon is complex and difficult to understand and interpret for others than subject matter experts. Therefore, some approaches to detect specific types of hate speech in different domains have been developed. This included detecting jihadist hate speech, anti-Muslim hate speech, anti-black hate speech, misogynistic hate speech, and anti-immigrant hate speech. When developing technologies for detecting specific types of hate speech, it is valuable to let subject matter experts assist in the development. By using experts to annotate data and/or create dictionaries that are designed to work on a particular domain, the algorithms will most likely perform better.

Yet another issue with hate speech detection is that many of the existing algorithms are easy to deceive. Gröndal et al. showed that by introducing spaces and misspellings, it is possible to fool several existing hate speech detection algorithms. The results from Gröndal and his co-authors indicate that an algorithm that flawlessly detects all kinds of hate speech is still out of reach.

---

## CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have described some of the challenges and opportunities with automatic hate speech detection. We have focused on the aspects that we find important for the study of hate in online environments and developing technologies for automatic hate speech detection. These aspects are: the definitions of hate speech, why do people express hate, technological approaches for hate speech detection, and the challenges that we encounter with automatic hate speech detection.

While it is clear that technologies can assist us in detecting different forms of hate speech and violent extremist propaganda, it is important to recognize the limitations of technology and the need for human interpretations. An algorithm that provides an optimal result in detecting hate speech is probably out of reach for the moment. However,

useful technologies already exist in particular if the results are combined with manual analysis. The requirement for the accuracy of the technologies depends on what the technologies should be used for. For example, when the objective is to observe hate trends over time, the absolute amount of false positives and negatives is perhaps less relevant, as long as their ratio does not change.

Another way of using existing technologies is to develop specialized algorithms for hate speech detection. The algorithms can then be trained to recognize specific types of hate speech (e.g., dehumanization, antisemitism, and misogyny) on specific domains. This might increase the performance of the algorithms, and the algorithms can be used to detect content that needs reviews from subject matter experts. This approach was also suggested by Alrhoun and colleagues when it comes to detecting different kinds of terrorist propaganda. The algorithms may then be used as an additional layer of automatic analysis to aid human moderation teams since it is more effective than manually going through an enormous amount of available data.

Despite the challenges, there is no doubt that technologies for assessing material from the internet are a necessity in our digital society. To succeed in developing automatic methods to detect hate speech, it is essential to have a comprehensive understanding of the problem's nature. We need to, for example, recognize the complexity of reaching an operational definition of hate speech, the many ways in which hate can be expressed, individual differences in hate speech perception. We need to make sure that the technologies work as we expect them to and encourage researchers from different disciplines to collaborate in the creation of our future algorithms. As we see it, the next breakthrough in hate speech detection is the outcome of multi-disciplinary work.

## CITATIONS

# Swedish Defence Research Agency and Uppsala University

- 1 See for example: Benesch, S. (2019) Proposals for Improved Regulation of Harmful Online Content, Paper for the Israel Democracy Institute (2019) and Ravndal, JA. (2013) Anders Behring Breivik's use of the Internet and social media. *Journal EXIT-Deutschland - Zeitschrift für Deradikalisierung und demokratische Kultur* (2)
- 2 Satariano, A. (April 7, 2019) Britain Proposes Broad New Powers to Regulate Internet Content, *New York Times*
- 3 Romm, T. (January 8, 2021) Facebook, Twitter could face punishing regulation for their role in U.S. Capitol riot, *Democrats say*, *Washington Post*
- 4 BBC News (July 27, 2020) Wiley: Anti-Semitism row prompts 48-hour Twitter boycott
- 5 George, C. (2015). Hate speech law and policy. *The International Encyclopedia of Digital Communication and Society*, 1-10
- 6 See for example Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018) All You Need is "Love": Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISeC '18)*. Association for Computing Machinery, New York, NY, USA, 2-12. and Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R., (2017) Deceiving google's perspective API built for detecting toxic comments, *CoRR*, vol. abs/1702.08138, 2017. [Online]. Available: <http://arxiv.org/abs/1702.08138> and Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019) The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. pp. 1668-1678. [Online]. Available: <https://www.aclweb.org/anthology/P19-1163>
- 7 E.g. Brown, A. (2017) What is hate speech? Part 1: The Myth of Hate. *Law and Philos* 36, 419-468 <https://doi.org/10.1007/s10982-017-9297-1>
- 8 United Nations. (2019). *United Nations Strategy and Plan of Action on Hate Speech*
- 9 European Union (2016). *Code of Conduct on Countering Illegal Hate Speech Online*. [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online\\_en](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en)
- 10 Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley, p.10
- 11 Ashmore, R. D. (1970) *Prejudice: causes and cures*. En B. E. Collins (ed.), *Social psychology*. Massachusetts: Addison-Wesley, Reading
- 12 See also Duckitt, J. H. (1992). *The social psychology of prejudice*. New York: Praeger
- 13 Mullen, B., & Leader, T. (2005) Linguistic Factors: Antilocutions, Ethnonyms, Ethnophaulisms, and Other Varieties of Hate Speech. In J. F. Dovidio, P. Glick, & L. A. Rudman (Eds.), *On the nature of prejudice: Fifty years after Allport* (p. 192-207). Blackwell Publishing. <https://doi.org/10.1002/9780470773963.ch12>
- 14 Facebook (2020). Facebook community rules. 12 Hateful content. [https://www.facebook.com/communitystandards/hate\\_speech](https://www.facebook.com/communitystandards/hate_speech)
- 15 Twitter. (2020). Our rules. Hateful conduct. [https://about.twitter.com/en\\_us/safety/enforcing-our-rules.html](https://about.twitter.com/en_us/safety/enforcing-our-rules.html)
- 16 There are some studies that have considered this, see for example: Sorokowski, P., Kowal, M., Zdybek, P., & Oleszkiewicz, A. (2020). Are Online Haters Psychopaths? Psychological Predictors of Online Hating Behavior. *Frontiers in psychology*, 11, 553. <https://doi.org/10.3389/fpsyg.2020.00553>
- 17 For example: Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41, pp 3-33, Mullen, B., & Leader, T. (2005) Linguistic Factors: Antilocutions, Ethnonyms, Ethnophaulisms, and Other Varieties of Hate Speech. In J. F. Dovidio, P. Glick, & L. A. Rudman (Eds.), *On the nature of prejudice: Fifty years after Allport* (p. 192-207). Blackwell Publishing. <https://doi.org/10.1002/9780470773963.ch12>, and Leader, T., Mullen, B., & Rice, D. (2009). Complexity and valence in ethnophaulisms and exclusion of ethnic out-groups: What puts the "hate" into hate speech? *Journal of Personality and Social Psychology*, 96(1), 170-182. <https://doi.org/10.1037/a0013066>
- 18 Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.
- 19 Allport, 1954, p. 14
- 20 Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley, p. 51
- 21 See for example Akrami, N. (2005). Prejudice: The interplay of personality, cognition, and social psychology. *Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 5. Uppsala, Sweden: University Library and Pettigrew, T. F. (1958) *Personality and sociocultural factors in intergroup attitudes: A cross-national comparison*. *Journal of conflict resolution*, 2(1), 29-42
- 22 E.g. Reynolds, K. J., Turner, J. C., Haslam, S. A., & Ryan, M. K. (2001) The role of personality and group factors in explaining prejudice. *Journal of Experimental Social Psychology*, 37(5), 427-434. <https://doi.org/10.1006/jesp.2000.1473>
- 23 Akrami, N., Ekehammar, B., Bergh, R., Dahlstrand, E., & Malmsten, S. (2009). Prejudice: The person in the situation. *Journal of Research in Personality*, 43(5), 890-897. <https://doi.org/10.1016/j.jrp.2009.04.007>
- 24 Duckitt, J., & Fisher, K. (2003). The impact of social threat on worldview and ideological attitudes. *Political Psychology*, 24(1), 199-222
- 25 See for example Akrami, N., Ekehammar, B., & Bergh, R. (2011). Generalized prejudice: Common and specific components. *Psychological Science*, 22(1), 57-59. <https://doi.org/10.1177/0956797610390384>, and Sibley, C. G., & Duckitt, J. (2008) *Personality and prejudice: A meta-analysis and theoretical review*. *Personality and Social Psychology Review*, 12(3), 248-279
- 26 Paulhus, D. L., & Williams, K. M. (2002) The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of research in personality*, 36(6), 556-563
- 27 Hodson, G., Hogg, S. M., & MacInnis, C. C. (2009). The role of "dark personalities" (narcissism, Machiavellianism, psychopathy), Big Five personality factors, and ideology in explaining prejudice. *Journal of Research in Personality*, 43(4), 686-690

- 28 Sorokowski, P., Kowal, M., Zdybek, P., & Oleszkiewicz, A. (2020). Are Online Haters Psychopaths? Psychological Predictors of Online Hating Behavior. *Frontiers in psychology*, 11, 553. <https://doi.org/10.3389/fpsyg.2020.00553>
- 29 Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, 67, 97-102. <https://doi.org/10.1016/j.paid.2014.01.016>
- 30 Akrami, N., Kaati, L., & Shrestha, A., (2021). Personality and online hate speech. Manuscript in preparation
- 31 Costa, P. T., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders*, 6(4), 343-359. <https://doi.org/10.1521/pepi.1992.6.4.343>
- 32 The machine learning models are described in: Akrami, N., Fernquist, J., Isbister, T., Kaati, L., & Pelzer, B. (2019) Automatic Extraction of Personality from Text: Challenges and Opportunities. 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, pp. 3156-3164, doi: 10.1109/BigData47090.2019.9005467 and Fernquist, J., Lindholm, O., Kaati, L., & Akrami, N. (2019) A Study on the Feasibility to Detect Hate Speech in Swedish, 2019 IEEE International Conference on Big Data, Los Angeles, USA, 2019, pp. 4724-4729, doi: 10.1109/BigData47090.2019.9005534
- 33 Akrami, N., Ekehammar, B., & Bergh, R. (2011). Generalized prejudice: Common and specific components. *Psychological Science*, 22(1), 57-59. <https://doi.org/10.1177/0956797610390384> and Sibley, C. G., & Duckitt, J. (2008) Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review*, 12(3), 248-279
- 34 E.g. Isbister, T., Sahlgren, M., Kaati, L., Obaidi, M., & Akrami, N. (2018) Monitoring Targeted Hate in Online Environments. In proceedings of the second workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)
- 35 E.g. Pelzer, B. Kaati, L. & Akrami, N (2018) Directed Digital Hate, 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), Miami, FL, 2018, pp. 205-210, doi: 10.1109/ISI.2018.8587396
- 36 Mehl, M., Robbins, M. & Holleran, S. (2013) How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *J. Methods Meas. Soc. Sci.* 3.
- 37 Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016) Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web (WWW '16). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 145-153. DOI: <https://doi.org/10.1145/2872427.2883062>
- 38 Sood SO., Churchill, EF., & Antin, J. (2012) Automatic identification of personal insults on social news sites. Automatic identification of personal insults on social news sites. *J. Am. Soc. Inf. Sci. Technol.* 63, 2 (February 2012), 270-285
- 39 As in Nobata et al, 2016
- 40 See for example: Mishra, P., Yannakoudakis H., & Shutova E. (2018) Neural character-based composition models for abuse detection. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pages 1-10. Association for Computational Linguistics, Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., and Leontiadis, I. (2019) A unified deep learning architecture for abuse detection. in Proceedings of the 10th ACM Conference on Web Science, Badjatiya, P. Gupta, S., Gupta, M., and Varma, V. (2017) Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759-760, Zhang, Z., Robinson, D., and Tepper, J. (2018) Detecting hate speech on twitter using a Convolution-GRU based deep neural network, in European Semantic Web Conference. Springer, pp. 745-760. [Online]. Available: [https://link.springer.com/chapter/10.1007/978-3-319-93417-4\\_48](https://link.springer.com/chapter/10.1007/978-3-319-93417-4_48) and Berglind, T., Pelzer, B., & Kaati, L., (2019) Levels of Hate in Online Environments. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada, 2019 pp. 842-847
- 41 Mackenzie, A. (2017). *Machine Learners: Archaeology of Data Practice*. Cambridge, MA: The MIT Press
- 42 [hatespeechdata.com](http://hatespeechdata.com)
- 43 Berglind, T., Pelzer, B., & Kaati, L., (2019) Levels of Hate in Online Environments. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada, 2019 pp. 842-847
- 44 Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018) All You Need is "Love": Evading Hate Speech Detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISec '18). Association for Computing Machinery, New York, NY, USA, 2-12
- 45 See for example: Fernquist, J., Lindholm, O., Kaati, L., & Akrami, N. (2019) A Study on the Feasibility to Detect Hate Speech in Swedish, 2019 IEEE International Conference on Big Data, Los Angeles, USA, 2019, pp. 4724-4729, doi: 10.1109/BigData47090.2019.9005534 and Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018) All You Need is "Love": Evading Hate Speech Detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISec '18). Association for Computing Machinery, New York, NY, USA, 2-12
- 46 Berglind, T., Pelzer, B., & Kaati, L., (2019) Levels of Hate in Online Environments. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada, 2019 pp. 842-847
- 47 See for example Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC
- 48 Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887
- 49 Crocker & Algina, 1986
- 50 See for example Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting Tweets against Blacks. In AAAI'13 Proceedings of the 27th AAAI Conference on Artificial Intelligence (pp. 1621-1622). Bellevue, WA: AAAI Press Pioletto, F., Stranisci, M., Sanguinetti, M., Patti, V., & Bosco, C (2017) Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In Basili, R., Nissim, M., & Satta, G. (Eds.), Proceedings of the Fourth Italian Conference on Computational Linguistics CLIC-it 2017, Accademia University Press
- 51 Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016) Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis, in Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, page 6-9
- 52 Laaksonen SM., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2020) The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring *Frontiers in Big Data*, 3
- 53 [www.hatebase.org](http://www.hatebase.org)
- 54 [www.rsd.org](http://www.rsd.org)
- 55 Kaati, L., Omer, E., Prucha, N., Shrestha, A. (2015) Detecting Multipliers of Jihadism on Twitter 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 954-960
- 56 Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018) The effect of extremist violence on hateful speech online. *arXiv.org*. <https://arxiv.org/abs/1804.05704>

- <sup>57</sup> Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting Tweets against Blacks. In AAAI'13 Proceedings of the 27th AAAI Conference on Artificial Intelligence (pp. 1621-1622). Bellevue, WA: AAAI Press
- <sup>58</sup> Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5), 4743-4752
- <sup>59</sup> Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016) Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis, in Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, page 6-9
- <sup>60</sup> Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018) All You Need is "Love": Evading Hate Speech Detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISeC '18). Association for Computing Machinery, New York, NY, USA, 2-12
- <sup>61</sup> Alrhoun, A., Maher, S., Winter, C. (2020) Decoding Hate: Using Experimental Text Analysis to Classify Terrorist Content. *Global Network on Extremism & Technology*, September 2020

## REFERENCES

- Akrami, N., Fernquist, J., Isbister, T., Kaati, L., & Pelzer, B. (2019) Automatic Extraction of Personality from Text: Challenges and Opportunities. 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, pp. 3156-3164, doi: 10.1109/BigData47090.2019.9005467
- Akrami, N. (2005). Prejudice: The interplay of personality, cognition, and social psychology. *Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 5. Uppsala, Sweden: University Library
- Akrami, N., Ekehammar, B., & Bergh, R. (2011). Generalized prejudice: Common and specific components. *Psychological Science*, 22(1), 57-59. <https://doi.org/10.1177/0956797610390384>
- Akrami, N., Ekehammar, B., Bergh, R., Dahlstrand, E., & Malmsten, S. (2009). Prejudice: The person in the situation. *Journal of Research in Personality*, 43(5), 890-897. <https://doi.org/10.1016/j.jrp.2009.04.007>
- Akrami, N., Kaati, L., & Shrestha, A., (2021). Personality and online hate speech. Manuscript in preparation
- Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley
- Alrhoun, A., Maher, S., Winter, C. (2020) Decoding Hate: Using Experimental Text Analysis to Classify Terrorist Content. *Global Network on Extremism & Technology*, September 2020
- Ashmore, R. D. (1970) Prejudice: causes and cures. En B. E. Collins (ed.), *Social psychology*. Massachusetts: Addison-Wesley, Reading
- Badjatiya, P. Gupta, S., Gupta, M., and Varma, V. (2017) Deep learning for hate speech detection in tweets. in Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759-760.
- BBC News (July 27, 2020) *Wiley: Anti-Semitism row prompts 48-hour Twitter boycott*
- Benesch, S. (2019) *Proposals for Improved Regulation of Harmful Online Content*, Paper for the Israel Democracy Institute (2019)
- Berglind, T., Pelzer, B., & Kaati, L., (2019) Levels of Hate in Online Environments. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Vancouver, BC, Canada, 2019 pp. 842-847
- Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41, pp 3-33
- Brown, A. (2017) What is hate speech? Part 1: The Myth of Hate. *Law and Philos* 36, 419-468 <https://doi.org/10.1007/s10982-017-9297-1>
- Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, 67, 97-102. <https://doi.org/10.1016/j.paid.2014.01.016>
- Costa, P. T., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders*, 6(4), 343-359. <https://doi.org/10.1521/pedi.1992.6.4.343>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.
- Duckitt, J. H. (1992). *The social psychology of prejudice*. New York: Praeger.
- Duckitt, J., & Fisher, K. (2003). The impact of social threat on worldview and ideological attitudes. *Political Psychology*, 24(1), 199-222
- European Union (2016). Code of conduct on countering illegal hate speech online. [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online\\_en](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en)
- Facebook (2020). Facebook community rules. 12 *Hateful content*. [https://www.facebook.com/communitystandards/hate\\_speech](https://www.facebook.com/communitystandards/hate_speech)
- Fernquist, J., Lindholm, O., Kaati, L., & Akrami, N. (2019) A Study on the Feasibility to Detect Hate Speech in Swedish, 2019 IEEE International Conference on Big Data, Los Angeles, USA, 2019, pp. 4724-4729, doi: 10.1109/BigData47090.2019.9005534
- Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., and Leontiadis, I. (2019) A unified deep learning architecture for abuse detection. in Proceedings of the 10th ACM Conference on Web Science
- Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5), 4743-4752
- George, C. (2015). Hate speech law and policy. *The International Encyclopedia of Digital Communication and Society*, 1-10
- Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018) All You Need is "Love": Evading Hate Speech Detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISeC '18). Association for Computing Machinery, New York, NY, USA, 2-12
- Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC
- Hodson, G., Hogg, S. M., & MacInnis, C. C. (2009). The role of "dark personalities" (narcissism, Machiavellianism, psychopathy), Big Five personality factors, and ideology in explaining prejudice. *Journal of Research in Personality*, 43(4), 686-690

- Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R., (2017) "Deceiving google's perspective API built for detecting toxic comments," CoRR, vol. abs/1702.08138, 2017. [Online]. Available: <http://arxiv.org/abs/1702.08138>
- Isbister, T., Sahlgren, M., Kaati, L., Obaidi, M., & Akrami, N. (2018) Monitoring Targeted Hate in Online Environments. In proceedings of the second workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)
- Kaati, L., Omer, E., Prucha, N., Shrestha, A. (2015) Detecting Multipliers of Jihadism on Twitter 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 954-960
- Krippendorff, K. (2004) Reliability in Content Analysis: Some Common Misconceptions and Recommendations. HCR, 30(3):411-433
- Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting Tweets against Blacks. In AAAI'13 Proceedings of the 27th AAAI Conference on Artificial Intelligence (pp. 1621-1622). Bellevue, WA: AAAI Press
- Laaksonen SM., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2020) The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring *Frontiers in Big Data*, 3
- Leader, T., Mullen, B., & Rice, D. (2009). Complexity and valence in ethnophaulisms and exclusion of ethnic out-groups: What puts the "hate" into hate speech? *Journal of Personality and Social Psychology*, 96(1), 170-182. <https://doi.org/10.1037/a0013066>
- Mackenzie, A. (2017). *Machine Learners: Archaeology of Data Practice*. Cambridge, MA: The MIT Press
- Mehl, M., Robbins, M. & Holleran, S. (2013) How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *J. Methods Meas. Soc. Sci.* 3
- Mishra, P., Yannakoudakis H., & Shutova E. (2018) Neural character-based composition models for abuse detection. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pages 1-10. Association for Computational Linguistics.
- Mullen, B., & Leader, T. (2005) Linguistic Factors: Antilocutions, Ethnonyms, Ethnophaulisms, and Other Varieties of Hate Speech. In J. F. Dovidio, P. Glick, & L. A. Rudman (Eds.), *On the nature of prejudice: Fifty years after Allport* (p. 192-207). Blackwell Publishing. <https://doi.org/10.1002/9780470773963.ch12>
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016) Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web (WWW '16). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 145-153. DOI: <https://doi.org/10.1145/2872427.2883062>
- Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018) The effect of extremist violence on hateful speech online. arXiv.org. <https://arxiv.org/abs/1804.05704>
- Paulhus, D. L., & Williams, K. M. (2002) The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of research in personality*, 36(6), 556-563
- Pelzer, B. Kaati, L. & Akrami, N (2018) Directed Digital Hate, 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), Miami, FL, 2018, pp. 205-210, doi: 10.1109/ISI.2018.8587396
- Pettigrew, T. F. (1958) Personality and sociocultural factors in intergroup attitudes: A cross-national comparison. *Journal of conflict resolution*, 2(1), 29-42
- Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., & Bosco, C (2017) Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In Basili, R., Nissim, M., & Satta, G. (Eds.), *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017: 11-12 December 2017, Rome*. Torino: Accademia University Press. doi:10.4000/books.aaccademia.2448
- Ravndal, JA. (2013) Anders Behring Breivik's use of the Internet and social media. *Journal EXIT-Deutschland - Zeitschrift für Deradikalisierung und demokratische Kultur* (2)
- Reynolds, K. J., Turner, J. C., Haslam, S. A., & Ryan, M. K. (2001) The role of personality and group factors in explaining prejudice. *Journal of Experimental Social Psychology*, 37(5), 427-434. <https://doi.org/10.1006/jesp.2000.1473>
- Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016) Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis, in Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, page 6-9
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019) The risk of racial bias in hate speech detection," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics. pp. 1668-1678. [Online]. Available: <https://www.aclweb.org/anthology/P19-1163>
- Satariano, A. (April 7, 2019) Britain Proposes Broad New Powers to Regulate Internet Content, *New York Times*.
- Sibley, C. G., & Duckitt, J. (2008) Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review*, 12(3), 248-279
- Sood SO., Churchill, EF., & Antin, J. (2012) Automatic identification of personal insults on social news sites. Automatic identification of personal insults on social news sites. *J. Am. Soc. Inf. Sci. Technol.* 63, 2 (February 2012), 270-285
- Sorokowski, P., Kowal, M., Zdybek, P., & Oleszkiewicz, A. (2020). Are Online Haters Psychopaths? Psychological Predictors of Online Hating Behavior. *Frontiers in psychology*, 11, 553. <https://doi.org/10.3389/fpsyg.2020.00553>
- Twitter. (2020). Our rules. Hateful conduct. [https://about.twitter.com/en\\_us/safety/enforcing-our-rules.html](https://about.twitter.com/en_us/safety/enforcing-our-rules.html)
- United Nations. (2019). United nations strategy and plan of action on hate speech. <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf>
- Unsvåg, E. F., & Gambäck, B. (2018) The effects of user features on Twitter hate speech detection. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) (pp. 75-85). Stroudsburg, PA: Association for Computational Linguistics.
- Zhang, Z., Robinson, D., and Tepper, J. (2018) Detecting hate speech on twitter using a Convolution-GRU based deep neural network, in European Semantic Web Conference. Springer, pp. 745-760. [Online]. Available: [https://link.springer.com/chapter/10.1007/978-3-319-93417-4\\_48](https://link.springer.com/chapter/10.1007/978-3-319-93417-4_48)

---

CASE STUDIES

---

## 05 The Kaleidoscope of Counterspeech: From The “Silent Majority” to the “Vulnerable Observer”

---

EXTRACTS FROM:

#iamhere

---

This case study aims to review the efficacy and effectiveness of the #iamhere methods in reducing hate speech and misinformation online and specifically on Facebook comment fields. The #iamhere methods, which target a kaleidoscope of readers – from the “silent majority” to the “vulnerable observers” – have been created to:

- give a voice to those who for different reasons do not speak out (“the silent majority”), making sure people are not intimidated into self-censorship and silence
- help and relieve the targeted from the oppression of hate – individuals such as public personalities, activists, journalists and politicians, indeed, anyone who is exposed to hatred and threats so that they can feel safe on social media (“the victims”)
- give more strength and protection to those already participating, supporting each other and engaging in constructive dialogue (“the allies”)
- instil doubts in those prone to hate speech and misinformation who are not yet participating in the spread of hatred, but might in the future (“the vulnerable observer”)

With the mission to make sure that the voices of those who stand up for democracy, human rights and inclusion will never be silenced.

#### **This case study will:**

- Explain the #iamhere organization and methods (section 2 and 3)
- Focus on the results and the strengths of the methods through already available research (section 4)
- Reveal the potential for development of the #iamhere actions in the current evolution of social media, in order to define future areas of progression (section 5)

In summary, the case study will help to show the relevance of the #iamhere methods when we come to address societal challenges such as hate speech and disinformation on social media.

---

### **I AM HERE INTERNATIONAL: THE NETWORK AND THE ORGANIZATION**

I Am Here International is an apolitical, non-religious, non-profit international organization which represents the world’s largest, citizen-driven, anti-online-hate movement, countering hate speech and misinformation online.

The movement is active all over Europe, in Australia, the United States and Canada. One hundred and fifty thousand citizens participate, taking action every day of the year on social media (mainly Facebook, but also Twitter and YouTube) against hate speech and disinformation. Members of the network counter hatred, protect the targeted and inspire people to speak out to defend human rights and freedom of speech.

#### **The methods used by #iamhere have been developed specifically to:**

- stop hatred and politically motivated hate speech
- support the targeted, as silence only strengthens the oppressor, never the victim
- strengthen and support the silent majority in reacting and speaking out when others are exposed to hatred and threats
- reduce polarization by engaging in discussion in a constructive, factual, nuanced and reasonable manner
- prevent the dissemination of disinformation by providing sources of factual information

The #iamhere's network is active in 15 national Facebook groups, in 10 different languages, with hands-on interventions on social media and via educational events, training programs and conferences.

This is the "call to action" in the description of all the Facebook groups of #iamhere, which embodies the main motivations fueling the courage of any member of the network.

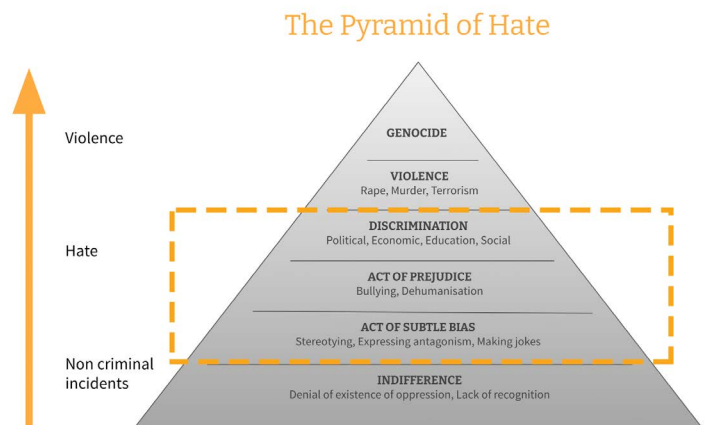
Inside the 15 private groups, each member is engaged with defined common rules (pls refer to section 3 for a more extensive explanation of our methods) to take organized actions in the comment sections of the Facebook platform. The members of #iamhere are emboldened to practise counter speech together in toxic comment sections in a collective act of courage against hate and disinformation.

Riding the waves of reactions – as the methods have been developed around the platform's architecture – these collective organized actions, made of top-level comments, second-level comments and likes/reactions, can push the hate further down the thread (section 3).

As several research papers have demonstrated, (section 4) this creates the so-called "contagion effect" that transforms the spiral of hate to a spiral of civility, opening the doors to positive participation on Facebook comment sections, fostering the freedom of each individual to speak up without fear. Organised actions take place every day, several times per day, in all the groups, with an average of 8.000 actions per year, in Europe alone, that rises to more than 10.000 actions yearly when combined with the Australian and the Canadian groups.

From Utoya to the Christchurch massacre, from Munich to El Paso attacks, all the terrorists declared that they received motivation and validation online and on social networks. Some of them used social networks for their proclamations and manifestos and even to broadcast their attacks live on social media. As political or religious terrorism and extremism do not arise in a vacuum, #iamhere works constantly to stop and prevent the biases, the prejudices and the discriminations from being translated into violent acts and even genocide.

"Would you dare to share your opinion more frequently if you knew that others in the same thread supported you? Are you already taking action to disrupt those who are spreading hate on the Internet, and do you wish you had the support of others who are doing the same?"



Those thousands of actions strike at the base and in the middle of the Pyramid of Hate, against acts and words of subtle bias, prejudice, and discrimination.



## THE #IAMHERE METHODS: A KALEIDOSCOPE OF COUNTER-SPEECH

The members of #iamhere gather in Facebook groups organized by nationality and/or language. Each group counter-speaks in their local language in the comment sections of public posts where hate, threats and disinformation appear. Sometimes, all the groups join together in a so-called “international action” that usually takes place in English in the comment fields of any national media.

The basis of counter-speaking in #iamhere is to write comments focusing mainly on the readers, thus reaching a kaleidoscope of targets. The intention is not primarily to reply directly to haters. Studies (section 4) demonstrate that it is very rare that a hater or a troll changes their mind and behavior, therefore, to reduce hate and misinformation it is instead key to prevent haters from occupying the public debate, and from frightening other citizens, deterring them from practising open, democratic, free speech.

The #iamhere methods, followed consistently in all the groups, rely on simple rules and repetitive actions taken inside and outside the groups.

## THE KEY STRUCTURES OF THE METHODS

### A. To observe and detect hate speech daily, preventing extremism at the early stages

- By being diligent sentinels monitoring hate speech. Every day, in all the #iamhere groups, the moderators regularly scan Facebook pages of news outlets, important public figures, and influencers’ profiles, to detect hate, attacks, harassments or disinformation in the comment sections. The members of each group are invited to do the same with the morning post called the “Fire Extinguisher”.
- By distinguishing hate from opinion. On the Fire Extinguisher post (right), anyone can share links to a comment section where hatred appears, proposing an action. As the mission of #iamhere is to promote lively debate and healthy discussion online, ensuring that no one is silenced, it is of major importance not to intervene in any constructive discussion or stop any free expression of any kind of opinion. Actions are only taken when there’s hate speech, or disinformation, when a person or a group of people is being harassed, attacked or threatened. Often these attacks are organised, targeting public figures and/or minorities.


**Helen Nightingale**  
Moderator  
· 16 October at 08:36 · 🌐

**Fire Extinguisher**

Good Morning Everyone, happy Friday to you all,  
“It does not matter how long you are spending on the earth, how much money you have gathered or how much attention you have received. It is the amount of positive vibration you have radiated in life that matters,”  
— Amit Ray  
Radiant [Jan Mayor](#) and me are your guides today.  
Have a happy day you lovely lot.

Let us know below if you see anything on Facebook in which [#iamhere](#) can get involved. Members can then comment on it and we may also call a stand-alone action.  
Please tell us:  
- the medium,  
- what the post is about,  
- what is going on and  
- make sure to copy and paste the Facebook link.  
By posting in the comment section of this thread you do not need to wait for us to approve your post; this way other members can come to your aid faster - it always makes things a bit easier when you don't feel so alone out there. We're in it together.

**Fire extinguisher Friday**




 **iamhere UK**

[View insights](#) 375 post reach >

👍❤️ 11 45 comments

- By organizing collective actions in places with high visibility. Moderators or administrators launch actions every day in each group, encouraging members to support each other in a specific comment section of a public post to counter hate speech. Usually those actions take place on newspapers' or public figures' pages, inside posts with high traffic and high visibility, where the hatred might be more dangerous.
- By ensuring consistency in any action undertaken by #iamhere. The format of the calls for actions, of the fire extinguishers, of the "love squads" or "love storms" is streamlined within the various groups, for the purpose of consistency (right).

 Karin Kay shared a link. Admin · 25 September at 14:14 · 🌐

**!!! TIME FOR ACTION !!!**

NOTE: Read this entire post for information and instructions. Click "View More" to see the full post.

Another attack that just happened in Paris. We don't know much yet, except this has started next to Charlie Hebdo's former offices. Comment sections are full of speculations. Let's provide some balance. We are posting three links. Simply copy and paste your comment into each thread.

BBC News  
[https://www.facebook.com/bbcnews/posts/10158181900472217?\\_\\_cft\\_\\_\[0\]=AZVhfotUip7sdNsaLoAAQrQ4pOxmglrPavSoZK37BNXd78nfs6UhJgLy\\_LvL1h3ky0IT4MsWBGx1UmpPpnib-fYa7-KBSq33ntZiQC9TtdorNI5X\\_GpRMGJTImt9DuxqFNUTPnW-ZCF1zhebZMr7aLaiFJ2AL6iqCjPQOOZ3xeUG9nPbSA8KrHo44yt3mqScDl4&\\_\\_tn\\_\\_=%2CO%2CP-R](https://www.facebook.com/bbcnews/posts/10158181900472217?__cft__[0]=AZVhfotUip7sdNsaLoAAQrQ4pOxmglrPavSoZK37BNXd78nfs6UhJgLy_LvL1h3ky0IT4MsWBGx1UmpPpnib-fYa7-KBSq33ntZiQC9TtdorNI5X_GpRMGJTImt9DuxqFNUTPnW-ZCF1zhebZMr7aLaiFJ2AL6iqCjPQOOZ3xeUG9nPbSA8KrHo44yt3mqScDl4&__tn__=%2CO%2CP-R)

The Daily Mail  
<https://www.facebook.com/164305410295882/posts/6166123950113968/?extid=krWVzqvDcCnurMeE&d=n>


Sky News  
<https://www.facebook.com/164665060214766/posts/4102688053079094/?extid=ZcijH6ZmXJw8hvbz&d=n>

INSTRUCTIONS:

- Enter the comment fields we link to above and post comments there. Feel free to tag #iamhere
- Like, respond and respond in support of other good comments to lift them in the fields. All efforts are equally important. The more answers, reactions and likes a comment gets, the higher it is raised. In this way we lift our comments and push down hateful comments.
- Avoid responding (eg with angry emoticon) and writing many responses to hateful comments, as this raises them higher in the fields. Like and respond to already existing good answers
- Write what you think and think yourself. But keep in mind that as members of #iamhere, we never spread hatred, prejudice, slander, gossip or rumors. We also do not comment on other people's spelling or writing. We always stay factual.
- Keep a good tone! We never express condescending, contemptuous, ridiculous, or insulting other people. Instead, with our word choices, we show that we stand for openness, respect and good conversation. This is true both in the comment fields we link to and here in the group.
- In order for us to be able to change and make a difference, it is outside the group, in the comment fields we link to, that debate and discussion about the issues should take place, and not here within the group. It is out there that we stop the hatred and nuance the debate. #iamhere is an action group - not a debate group.
- However, small talk about the campaign, such as pepping, support, tips and advice is OK to vent here within the group, in this thread. Also, please post in this thread if you commented (C), liked (L), or responded (R), and in which fields.

#jagärhär #iamhere #jesuisla #ichbinhier #osonoqui #somtu #jsmetu #olentäällä #teztujestem #vierher

# BREAKING NEWS


 NEWS.SKY.COM

**Four wounded in knife attack near former Charlie Hebdo offices in Paris**

[View insights](#) 349 post reach >

## B. To counter-speak with specific rules of engagement

- By countering hate and replacing it with respectful conversation. Members of #iamhere are always invited to intervene by counter-speaking hate in every action. They're always reminded to be empathic and respectful, to base statements on facts and reason, not to "preach" to the other participants in the conversation, to stay on topic and not to engage with trolls or react to provocations.
- The methods suggest writing new stand alone comments. This is preferable to replying directly to haters' comments, which gives them more visibility. The methods also recommend interacting with each other's comments in order to make them top-level comments, setting the agenda with a constructive tone in the conversation (internal or external to the groups). By flooding the comments section with civil and constructive comments, the hatred is pushed down from the algorithm until it is no longer visible.

 Curly Caroline shared a post. Admin · 24 September at 16:40 · 🌐

**!!! TIME FOR ACTION !!!**  
NOTE: Read this entire post for information and instructions. Click "See More" to see the full post.

Megan Rapinoe's acknowledgement as one of the THE 100 MOST INFLUENTIAL PEOPLE OF 2020 is bringing out some vile comments.  
<https://www.facebook.com/10606591490/posts/10157943269136491/?extid=Of8F76bP0xk2rbh8&d=n>

"Rapinoe's impact goes far beyond the pitch. In an era where many demand that athletes "stick to sports," Rapinoe—a proud feminist and an out gay advocate—refuses to be silenced. In the past year, her activism for gender pay equality, racial justice and LGBTQ rights has become as iconic as her fabulous pink hair. Some critics are threatened by her boldness and power. Millions of her fans around the world are inspired.


Megan Rapinoe fearlessly uses her voice to make the world a more equal place. No matter your politics, ethnicity or gender, that's something we should all celebrate."

**INSTRUCTIONS:**


- Enter the comment fields we link to above and post comments there. Feel free to tag #iamhere
- Like, respond and respond in support of other good comments to lift them in the fields. All efforts are equally important. The more answers, reactions and likes a comment gets, the higher it is raised. In this way we lift our comments and push down hateful comments.
- Avoid responding (eg with angry emoticon) and writing many responses to hateful comments, as this raises them higher in the fields. Like and respond to already existing good answers
- Write what you think and think yourself. But keep in mind that as members of #iamhere, we never spread hatred, prejudice, slander, gossip or rumors. We also do not comment on other people's spelling or writing. We always stay factual.
- Keep a good tone! We never express condescending, contemptuous, ridiculous, or insulting other people. Instead, with our word choices, we show that we stand for openness, respect and good conversation. This is true both in the comment fields we link to and here in the group.
- In order for us to be able to change and make a difference, it is outside the group, in the comment fields we link to, that debate and discussion about the issues should take place, and not here within the group. It is out there that we stop the hatred and nuance the debate. #iamhere is an action group - not a debate group.
- However, small talk about the campaign, such as pepping, support, tips and advice is OK to vent here within the group, in this thread. Also, please post in this thread if you commented (C), liked (L), or responded (R), and in which fields.

#jagärhär #iamhere #jesuisla #ichbinhier #iosonoqui #somtu #ismetu#olentäällä #teztujestem #vierher

An assignment listing Kenosha gunman Kyle Rittenhouse as a 'modern-day hero' was given to seniors in an English class at W.T. White High School in Dallas, Texas.

 TIME · 24 September at 14:00 · 🌐

Kirsten Gillibrand on Megan Rapinoe: "Rapinoe—a proud feminist and an out gay advocate—refuses to be silenced. In the past year, her activism for gender pay equality, racial justice and LGBTQ rights has become as iconic as her fabulous pink hair."



TIME.COM  
Megan Rapinoe: The 100 Most Influential People of 2020

[View insights](#) 665 post reach >

- By supplying facts, ideas and new ways of thinking, to bring a constructive voice. Members are never told what to write, and this is essential to the non-partisan philosophy of the groups. The point is to show that it is possible to discuss sensitive matters and express opinion gently - respecting others' freedom of speech, using rationality and facts. We encourage members to back up statements and present credible pieces of evidence, fresh arguments and new ideas.

Curly Caroline shared a link. ...

Admin  
· 10 October at 22:36 · 🌐

!! 🌍 INTERNATIONAL JOINT ACTION 🌍 !!

While we're not a political group, Greta Thunberg has chosen to endorse Biden in the US election for his concern about the environment.

Needless to say, she has been subject to personal attacks and unkind comments.

Let's go in there and support her right to speak out without fear of abuse.


<https://www.facebook.com/Reuters/posts/3742071065813222>

NOTE: Read this entire post for information and instructions. Click "View More" to see the full post.

INSTRUCTIONS:

- Enter the comment fields we link to above and post comments there. Feel free to tag [#iamhere](#)
- Like, respond and respond in support of other good comments to lift them in the fields. All efforts are equally important. The more answers, reactions and likes a comment gets, the higher it is raised. In this way we lift our comments and push down hateful comments.
- Avoid responding (eg with angry emoticon) and writing many responses to hateful comments, as this raises them higher in the fields. Like and respond to already existing good answers
- Write what you think and think yourself. But keep in mind that as members of [#iamhere](#), we never spread hatred, prejudice, slander, gossip or rumors. We also do not comment on other people's spelling or writing. We always stay factual.
- Keep a good tone! We never express condescending, contemptuous, ridiculous, or insulting other people. Instead, with our word choices, we show that we stand for openness, respect and good conversation. This is true both in the comment fields we link to and here in the group.
- In order for us to be able to change and make a difference, it is outside the group, in the comment fields we link to, that debate and discussion about the issues should take place, and not here within the group. It is out there that we stop the hatred and nuance the debate. [#iamhere](#) is an action group - not a debate group.
- However, small talk about the campaign, such as pepping, support, tips and advice is OK to vent here within the group, in this thread. Also, please post in this thread if you commented (C), liked (L), or responded (R), and in which fields.

[#jagärhär](#) [#iamhere](#) [#jesuisla](#) [#ichbinhier](#) [#iosonoqui](#) [#somtu](#)  
[#smetu](#) [#olentäällä](#) [#teztujestem](#) [#vierher](#)



REUTERS.COM

**Climate activist Greta Thunberg shows support for Biden in rare political tweet**

[View insights](#) 307 post reach >

### C. To be consistent, cohesive and safeguard other members

- By using the hashtag on our posts, liking members comments, commenting on other members' top-level comments. The method envisages a series of actions that give more visibility to the #iamhere comments. Members are encouraged to add the #iamhere hashtag in their comments, to help other members identify them. Since not everyone is confident enough to post comments, after someone else has written one, #iamhere recommends that others like, sub-comment, or somehow interact with the original TL comment. This way, respectful and factual comments are lifted to the top, and not least, members of the group are never alone facing hatred: there's always a team behind to give strength and courage.

“I am here”, says Mina Dennert, the founder of #iamhere and creator of the methods, “means ‘this is where I am, and I need help here in this comment field,’ and then we call for each other to help out. It also has the meaning ‘I am present, I can see what you (hater) are doing here and I don’t agree. I am here too.’ ”

### D. To create links and bridges with all the readers

- By breaking filter bubbles and organizing actions wherever hate is. Generally we take action on media outlets or Facebook groups that are closer to extremism. Counter speaking on these pages allows groups to reach a wider target and those readers who might be prone to hate, offering them a different point of view and a more respectful narrative. The main aim of counter speaking is to stop hate, speaking your own mind, and whenever possible planting a seed of reflection and doubt. It's about taking a stand for human rights and hopefully reaching the “vulnerable observer” (see section 4 for definition of “vulnerable observer”).

- This is a call to action in the #jesuislà group, with a link to a CNEWS article about degradation of a kosher restaurant. CNEWS is often criticized for tolerating hate speech on its programs, and on its Facebook page. In this post, the moderator explains what the group members are likely to expect in the comment field (antisemitism, anti-muslim hate, glorification of nazism...).

Neringa Luko a partagé une publication. Modérateur · 2 octobre, 19:58

**ACTION DU SOIR !**

Bien chers vous tous,

L'actualité du soir qui déclenche concurrence victimaire, blagues à deux balles et antisémitisme décomplexé, c'est ce restaurant fast-food casher du 19ème arrondissement de Paris. Des croix gammées taguées partout, un « Hitler avait raison », « sale juif »... Nous pensons qu'il serait bon d'aller mettre un peu de bon sens et de bienveillance dans tout cela, voire rappeler que certaines « opinions » sont des délits...

Vous pouvez aller commenter... Afficher la suite

**CNEWS** 2 octobre, 19:08

Les inscriptions laissées par les auteurs du saccage font froid dans le dos

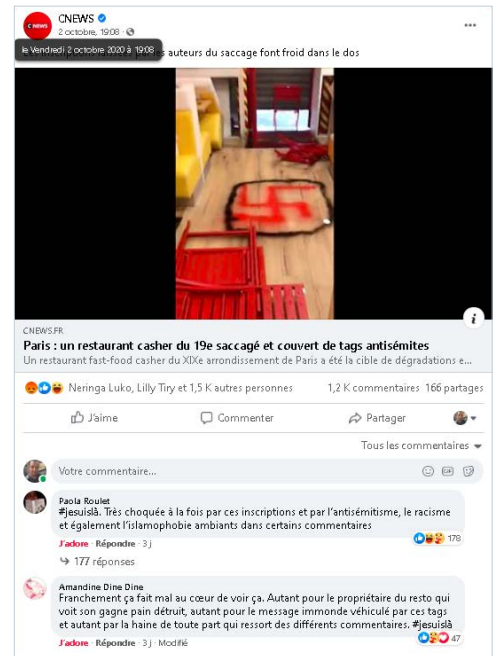
**Paris : un restaurant casher du 19e saccagé et couvert de tags antisémites**

Voir les statistiques Couverture de la publication : 1,4 K

25 70 commentaires

She wrote: “#jesuislà. Very shocked both by these writings, and by antisemitism, racism and also ambient anti-muslim hate in certain comments”. There is a lot of conversation under her comments, she is attacked by trolls, and several other members are helping her. Other members also have their comments moving up to the top of the section.

→ This is the comment section under the CNEWS article. There are 1200 comments, and the comment of one #jesuislà member appears on top of the comment section, with the most likes.



- By not leaving the hate unchallenged. The method is always to ignore and block hateful comments, never to feed the trolls, reporting illegal content to the platform or the police. However, should a “vulnerable observer” start a discussion below a member’s TL comment, then connecting and answering them is a means to avoid leaving the hate unchallenged, as long as any illegal content isn’t shared.



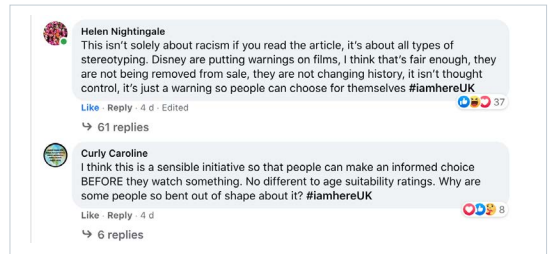
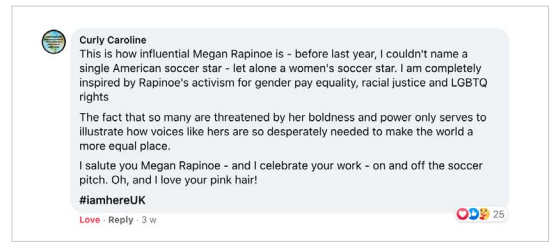
- An anti-masker debating with one of our members and finally understanding and admitting why we wear masks for COVID-19 and were not wearing them for the flu - because there is a vaccine for the flu and not for the COVID.

An internet user described Sophie Pétronin, the French hostage in Mali, who was liberated, as a “bobo”. “Bobo” stands for Bohemian Bourgeois, but it is usually used as a derogatory word to identify “leftists” that have money and/or education and know nothing about “real life”. When asked by a member, he says he still disagrees with her declarations that she wants to go back to Mali, but recognizes that his use of the word “bobo” was a “a big stupidity” from him.



↑ He says to her “OK, I concede. I can’t wait for the anti-COVID vaccine to arrive. Have a good night”

- By using respect and a low arousal approach as the key values in the interactions. Calm against aggression, objectivity and a mild approach to matters, focus on the issue and being pleasant and respectful towards anyone is key in any #iamhere action. Once again, we interact solely with “vulnerable observers” not trolls or haters, only interacting with legal content published under a constructive TL comment. The tone of voice and the approach are therefore immediately recognizable even by a distracted reader and can determine the final success of an action, making hatred invisible on everyone’s timeline.



## TALKING TO READERS, NOT WITH HATERS: WHY IT WORKS

### A. Talking to the readers

#iamhere’s methods, as explained above, have been created to encourage people to speak up and to engage in online debate; to strengthen journalists and politicians and opinion formers by making people react and speak out when others are exposed to hatred and threats; to stop polarization by engaging in political discussion constructively and factually, and to prevent the spread of disinformation by providing sources of factual information.

To reach all those objectives the method relies on one of the axioms of counter-speech: writing and talking to the readers is much more effective than replying to haters (Miskolci et al. 2018 ; Schieb and Preuss, 2016 ; McGrath, 2017 ; Kolbert, 2018 ; Lazarus and Clifford, 2018 ).

Obviously, the more readers a counter speaker can engage with, the more effective the counter speech.

Thanks to the diversity within the membership, #iamhere groups’ actions are able to catch the attention of a kaleidoscope of readers, representing all the different shades and colors of psychological attitudes towards hate speech and disinformation. The majority of readers on social media are the ones that don’t agree to the hateful content but do not care enough, don’t feel they have the time, don’t think it is important enough or don’t have the courage to counter hate alone or simply express themselves because they are reduced to silence for fear of attracting hatred (the “silent majority”).

Readers, usually public figures or minorities (the “victims”) can also be the actual targets of hate and threats, exposed daily to harassment . We may also reach those who are already acting to stop hate and disinformation (the “allies”), giving them strength and protection to keep on counter speaking. Finally, the readers we want to engage with using our methods are the people who feel attracted to hate speech and misinformation, but not yet have started trolling or writing illegal content (the “vulnerable observer”).



## B. The Scientific Background explaining why it works

There are already several studies that demonstrate how and why the #iamhere method is positively working on those very different categories of readers active on social media.

The Department of Psychology of the University of Gothenburg (Moa Lanngren, Managing online harassment – public figure’s perception of online social support) analyzed whether online support could buffer the negative consequences of online harassment for some Swedish public figures who had experienced prolonged abuse and received social support from #jagärhär (the Swedish #iamhere group).

The study, using a technique called “Interpretative phenomenological analysis” to analyze the data from the interviews of the “victims”, shows that all the participants in the study itself experienced online social support as something “overwhelmingly positive”. They outlined that the support was important personally to them, but also for public conversation in general.

Some of the participants in the study even stated that the online support from the #iamhere members was a “prerequisite for their continued work”. Others stated that the personal and intimate experience of the harassment was not affected by the fact that they received encouraging comments online and that they had to create their own processes to manage online harassment.

The conclusions of the study underlined that “certain forms of benign disinhibition seemed to be useful to lessen the negative effects of toxic disinhibition. That means that online social support is important and initiatives like #iamhere are needed. It might not work as a single solution to solve the problem, but together with other solutions (...) progress will be made”.

Protecting the targeted is essential for freedom and democracy: online hate is a major problem by itself, but when it is public figures that are targeted and intimidated to silence, we’re not facing a single crime against an individual. It’s not only a threat to freedom of speech for this person, but it jeopardizes the very essence of democracy.

Enlarging the target readers to include the “silent majority” and the “vulnerable observers”, the Dusseldorf Institute for Internet and Democracy inquired whether or not #ichbinhier (the German #iamhere group) counter-speech has a measurable positive effect on the quality of the follow-up discussion. If it can improve the quality of online debate

perhaps it has the potential to reduce the harmful effects of hate speech and incivility? The question to consider is whether the “spiral of incivility” can be transformed into a “spiral of civility” if comments are respectful, rational, constructive, and reciprocal.

The study by Marc Ziegele, Pablo Jost, Dennis Frieb, and Teresa Naab was based on a quantitative content analysis on all the group actions for 3 months, between November 2017 and January 2018, on a total of 14,104 top-level comments (and their first five replies) on 167 news articles from 21 Facebook pages of German media outlets. The final data set contained 641 top-level comments and 2.928 replies (second level comments).

This study sheds a definitive light on the fact that the #iamhere members’ comments, through their presence alone, increased the average level of online discussion for rationality, constructiveness, civility and politeness. That those comments also motivated other users to have a better follow up discussion was quite clear: “collective civil moderation already contributes to a more balanced, less hateful and uncivil discussion; high-quality comments were also related to increased quality in the subsequent discussion.

Our results do not imply that the #iamhere comments will convince or silence hateful commentators altogether, but they may motivate the quiet readers to contribute with civil, reasoned, and constructive comments themselves, thereby achieving a more pluralistic and democratic culture of online communication. Nevertheless, the personal courage it takes to engage in hatred online discussion, even as a part of a collective, should not be underestimated”.

The so-called “contagion effect” of #iamhere counter speech in boosting collective personal courage against hate speech has been explored and analyzed extensively in a new study, in publication, from researcher Cathy Buerger of the Dangerous Speech Project.

The study used semi-structured ethnographic interviews to examine why group members got involved with #jagärhär (the Swedish #iamhere group), how they strategize while counter speaking, what challenges they face in their work and what keeps them engaged.



### The study analyzed:

- how #iamhere uses the Facebook architecture to their advantage to amplify positive speech
- how the movement can address to the “moveable middle” exposing them to a positive tone
- how #iamhere draws new counter-speakers into conversations, lessening the exposure to hatred for any new member, protecting them, letting them feel braver, with the hashtag perceived as a “shield”
- how #iamhere methods work to keep the members engaged, giving a sense of safety, belonging and happiness.

Finally, the study proved “the value of collective action in fighting against hateful speech online (...) This study is the first to consider the psychological impact of engaging in counter-speech as part of a group. Group members report feeling braver and more willing to enter difficult conversations. Additionally, they describe many aspects of the #jagärhär/#iamhere model that may prevent burnout, a major obstacle to sustainability for many social justice initiatives.”

The study also nuances and focuses on what is considered an effective and successful counter-speech campaign - not changing anyone’s mind, but writing for the larger audience, which we name the kaleidoscope of readers. “Success for #jagärhär members, therefore, isn’t measured by how many hateful comments exist in a conversation, but by how much space has been created for alternative viewpoints. Are there any new voices of people who now feel safe enough to express their opinions in a civil way? If so, then that is a success.”

### As one member described:

“In the end, it’s about democracy, it’s about the debate, it’s about freedom of speech that people will have the courage to say what they think. If you have lots of hate comments, maybe you are afraid, and you don’t want to say what you think. But if we are 10-20 people arguing against the hate then I imagine that others will also want to do so, so that not only the people screaming the highest can say their opinion.”

---

## CHALLENGES, CONCLUSIONS AND TAKE-AWAYS

The #iamhere methods have been shown to be efficient and effective in speaking to a kaleidoscope of readers to activate the “silent majority”, to support and strengthen the “allies”, to protect the “victims” and to plant seeds of doubts in the “vulnerable observers”. However without the normative help of institutions and a major investment in human moderation by platforms and media outlets, it appears very unlikely that the hate and misinformation that are threatening our democracies will slow their growth.

It is not just a matter of supporting the groups of counter-speakers, but of working together in a long-term coordinated collaboration between citizen movements, social media platforms and media outlets. It is a matter of implementing stronger anti-hate and misinformation measures on platforms. Trolls and haters demonstrate every day that they are more coordinated: from the QAnon movement, to the Italian “La Bestia” organized by populist parties, and the COVID-19 negationist crusade. We fear that no matter how much effort is put into encouraging the silent majority or supporting the allies and the victims, if these atrocities are not stopped or the hate speech or the disinformation is not removed from the platforms, the trolls and haters are more likely to reach the “vulnerable observers” than counter speakers. They are more organized, more interconnected and have strong political and economic forces behind them. And it is infinitely easier and faster to make up and spread the lies that people want to hear, than to counter them with facts and research.

Battling hate speech and misinformation every day on social media is tiring and sometimes exhausting. It is difficult to motivate #iamhere members to sustain the battle when they do not see help and positive feedback from platforms and institutions. The burnout is even more evident in administrators and moderators as they’re more exposed to hatred and they’re on the frontline. The self-care measures activated inside the groups are helping to reduce the burden of the volunteers who tirelessly counter hate but voluntary-based civil society organisations can only do so much.

To support and defend free speech, human rights, and ultimately our democracies, is a global responsibility. No-one can deny accountability for it: civil society can operate up to a certain point, then the work needs to be taken over by institutions and media companies.

The #iamhere methods have demonstrated that they can reduce the collective and personal burden of hatred and open new spaces for participants in social media where they can express themselves, expanding freedom and diversity of speech. Unfortunately, this alone is insufficient to stop hatred and misinformation and to fully transmute the toxicity in the comment fields.

Without collective ownership and responsibility for the battle against hate and disinformation, this threat to democracy will not be stopped, in either the short or the long term. Civil rights movements, organized groups of counter speakers, institutions, and not least, companies owning social, digital, and general media must create a coalition against hatred and fake information. They must pass to each other the torch of democracy, starting with the cultural foundation of an inclusive, open, respectful, fact grounded society. This can be developed by civil movements, with indispensable parameters of regulation and the decisive economic investment of the companies involved. Hate and disinformation will be defeated. Only together.

---

## CITATIONS

# #iamhere

---

- <sup>1</sup> J.Miškolci, L Kováčová, E.Rigová “Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia” - Social Science Computer Review, SAGE Journals, 2018
- <sup>2</sup> Benesch et Al. “Considerations for successful counterspeech. Dangerous Speech Project.” 2016
- <sup>3</sup> J.Miškolci, L Kováčová, E.Rigová “Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia” - Social Science Computer Review, SAGE Journals, 2018
- <sup>4</sup> C. Schieband M.Preuss. “Governing hate speech by means of counterspeech on Facebook.”66th ICA annual conference, at Fukuoka, Japan, pp. 1-23, 2016
- <sup>5</sup> A. McGrath “Dealing with dissonance: A review of cognitive dissonance reduction.” Social and Personality Psychology Compass 11, no. 12, 2017
- <sup>6</sup> E. Kolbert “Why Facts Don’t Change Our Minds”, The New Yorker, February 2017
- <sup>7</sup> Lazarus and Clifford “Why Many People Stubbornly Refuse to Change Their Minds.” Psychology Today, 2018
- <sup>8</sup> M. Lanngren “Managing online harassment – public figure’s perception of online social support”, Master Thesis in Psychology, Department of Psychology, University of Gothenburg, 2020
- <sup>9</sup> M. Ziegele, P. Jost, D. Frieb and T. Naab “Collective Civic Moderation for Deliberation? Exploring the Links between Citizens’ Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussion”, Political Communication, 2019
- <sup>10</sup> C.Buerger “Dangerous Speech Project – the Anti-Hate Brigade” The Dangerous speech Project, In Publication, 2020

---

CASE STUDY

---

## 06 Galop and Community Partnership Working to Combat Hate Crime

---

EXTRACTS FROM:

Galop

---

---

## ABOUT GALOP

Galop is the UK's only specialist LGBT+ anti-violence charity. Our mission is to make life safe, just, and fair for LGBT+ people (lesbian, gay, bisexual, trans and other minoritised sexual and gender identities). For nearly 40 years we have provided advice, support and advocacy to LGBT+ victims and campaigned to end anti-LGBT+ violence and abuse. Galop works within three key areas; hate crime, domestic abuse, and sexual violence. We work to help LGBT+ people achieve positive changes to their current situation, through practical and emotional support, to develop resilience, and to build lives free from violence.

---

## SUMMARY

Galop's LGBT+ hate crime service gives independent advice to LGBT+ people facing homophobia, biphobia and transphobia. In London, we also offer a specialist casework service for LGBT+ people in London experiencing hate crime.

Galop's hate crime service works closely in collaboration with other specialist anti-hate community organisations, standing together against all kinds of hate crime. Our approach to combating hate speech and extremism is based on mutual support and cooperation. We coordinate with each other and build resilience among minoritised groups by training and supporting other community organisations.

### **This case study details our work in four key areas:**

- Victim-centred specialist hate crime advocacy
- Building an evidence base about LGBT+ hate crime
- The CATCH (Communities Against Hate) partnership
- The CATCH Together Against Hate training programme.



---

## VICTIM-CENTRED SPECIALIST HATE CRIME ADVOCACY

Our hate crime advice, advocacy and reporting services are considered an international model of best practice, with one of the first established third-party reporting sites, responsible for increased reporting of anti-LGBT+ hate crime. We remain responsive to evolving LGBT+ needs, improving support for victims of hate crime and building our communities' trust and confidence to report. We have extensive experience of working with people experiencing hate speech and hate crime both online and offline, and the interplay between these two spheres. We also have a specialist trans safety casework service and a young people's safety casework service.

### **Galop provides support, advice and advocacy, including the following:**

- Ongoing emotional support
- Information and advice
- Signposting and support
- Housing support
- Third party reporting
- One-off advice
- Assistance and advocacy around navigating the criminal justice system
- Representation at multi-agency panels
- Support at court

Below are a few case examples and how we help. Names and identifying details have been changed, apart from the first case which was reported widely in the media, including Galop's involvement in accordance with the wishes of the victims.

### **Homophobic attack in public- CJS support**

In May 2019, Melania Geymonat and Christine Hannigan were subjected to a homophobic attack on a bus. The perpetrators made sexual gestures towards them, demanding that they kiss and show how lesbians have sex. When they refused, they were surrounded, punched in the face several times, and robbed.

Galop worked with the police and CPS in ensuring that the hate element of the crime was recognised, and supported the victims through the criminal justice system, including during court. Three young men were convicted of the attack.

Commenting to Galop, Melania Geymonat said:

“Thank you to Galop for all the support they have kindly given to me. No one should be violated, especially just for being who they are. It is really important in these times of growing social conservatism and intolerance that we fight for our rights that took so long to be achieved.”

### **Homophobic harassment from neighbour- Non-CJS solutions**

Helen was referred to Galop after reporting multiple incidents of hate crime from her neighbour. Helen had been experiencing extreme amounts of verbal abuse and harassment from years from her neighbour but had not reported until the abuse became unbearable.

Galop worked with Helen to create a support and action plan, provide information about her rights and entitlements as a victim of hate crime, and assisted with making an incident log to record the abuse for statutory agencies.

We represented Helen at multiple community panels, and liaised with the council, police, social services and housing. We assisted Helen in applying for a priority move with her housing provider, by collecting supporting documents from the police and medical providers, and supporting Helen throughout the process. Helen was rehoused and we supported with the costs of moving via our emergency client fund.

Helen also faced some financial difficulties due to the COVID-19 pandemic. Galop assisted Helen with accessing grants, food bank vouchers and signposted to debt support organisations. This enabled Helen to get the support she needed, and she is now in a financially secure position, in her new home and has access to long term support.

The perpetrator has had restrictions placed upon them by their housing provider, and they are now getting the support needed via social services.

### **Transphobic harassment from neighbour- CJS support**

At the beginning of lockdown, Maja, a trans woman who lives alone, began to receive regular transphobic and homophobic abuse from her neighbour, including threats against her life. Maja was struggling to get local police to respond to the incidents and they were initially recorded as a “non-crime”. Maja requested that a Hate Crime Liaison Officer be involved in her case, but these requests were initially refused. She then got in touch with Galop.

We contacted her local Hate Crime Liaison Officer and established ongoing communication with them, Maja and the officer in charge of the case. We also contacted her housing officer, who initially framed the issue as one of clashing cultural differences between Maja and the perpetrator. We worked with them to see that Maja was being subjected to abuse and at risk of serious violence, which is unacceptable whatever the private views of the perpetrator. We ensured information sharing between the council and police to safeguard Maja effectively.

The perpetrator was charged and an injunction order was placed on the perpetrator upon bail. However, the perpetrator continued to harass Maja, repeatedly breaching his bail conditions, and when Maja called the police, the attending officers still failed to take it seriously. Due to consistent recording and reporting by Maja, we were able to evidence the on-going abuse to the Hate Crime Liaison Officer and officer in charge, who intervened and the perpetrator was eventually held on remand.

Awaiting court, Maja was anxious about appearing as a witness. We prepared her for what to expect on the day and what her rights were. On the day, the perpetrator changed his plea to guilty and was convicted. We ensured that the council housing team were informed of the conviction so they could take appropriate action to protect Maja from the perpetrator.

## Client comments illustrating key good practices for LGBT+ hate crime services:

“My wife and I were listened to, treated with respect and understood. The advocate was emphatic, knowledgeable and effective. First of all, our experience was understood and validated. We were reassured and treated seriously. After months of being ignored by the police and housing department, which left us at the mercy of the abuser, having someone who took us seriously meant a lot, gave us hope and allowed us to manage better. It was an absolute relief to know that there is somebody who gets it. Secondly, the advocate was the first person who fully explained to us our rights and the law, giving us chance to think clearly about options that we have. Finally, as a result of the advocacy the police got in touch with us, recorded the harassment that we were experiencing and took action.”

“The support I had was incredible – the advocate had the confidence, understanding and knowledge around the justice system which was vital for helping me feel better, as I didn’t know how it all worked. He was a very reassuring and kind and calming presence. He also went out of his way to help me out in this case – and offered to go with me to the court etc. It had a successful result as I didn’t feel the hate crime element was being heard before Galop got involved. What could have been really traumatic was significantly eased.”

“I received prompt, considerate and respectful help. Everyone was very kind and used the correct pronouns. I felt much more supported when dealing with the police, I wouldn’t have been brave enough to report the crime without Galop.”

“Sincere, supportive and felt really personal. I felt my feelings were absolutely validated.”

“Great, helpful, friendly, caring and getting things done. Emotionally it massively helped.”

---

## BUILDING AN EVIDENCE BASE ABOUT LGBT+ HATE CRIME

It is vital that policy makers in criminal justice, statutory and governmental agencies, as well as stakeholders like social media companies, understand the true extent of violence and abuse against minoritised groups and the impact it has on the lives of individual people, so that strategies are implemented to tackle this violence. We work towards this by facilitating reporting and ensuring that incidents are properly recorded as hate crime; providing training and resources to agencies to properly identify and recognise

hate crime; and providing information to LGBT+ people about their rights and options.

In 2016, we entered into a data sharing agreement with the National Police Chief’s Council, which enables us to receive anonymous detailed data on homophobic and transphobic incidents recorded by the police, as well as sharing information on the trends we are currently seeing. We have built a body of research on anti-LGBT+ hate crime, to show what our communities are facing, its impact, and what is needed. We also collaborate with university researchers, for example the Hate Lab at Cardiff University in tackling online hate speech.

---

### RESEARCH

---

- [Galop Online Hate Crime Report 2020](#)
- [Galop Online Hate Crime Report 2017](#)
- [All research publications](#)

---

### RESOURCES FOR COMMUNITY MEMBERS:

---

- [Hate Crime: A Guide for LGBT+ people](#)
- [What is online anti-LGBT+ hate speech and hate crime?](#)
- [A practical guide to tackling online anti LGBT+ hate crime](#)
- [All factsheets](#)

---

### RESOURCES FOR PROFESSIONALS:

---

- [Working with Victims of Anti-LGBT Hate Crimes: A Practical Handbook](#)
- [Online anti-LGBT+ hate crime: A guide for organisations](#)
- [All professional guides](#)

---

## THE CATCH (COMMUNITIES AGAINST HATE) PARTNERSHIP

Galop leads the CATCH partnership, which brings together the leading specialist community organisations that tackle hate crime, discrimination and abuse, across all hate crime strands: race, religion, disability, sexual orientation and gender identity. Between us, we have over 100 years of working to support victims of hate crime, and well-established partnerships with the relevant police, CPS, statutory and voluntary agencies to enable cross-service signposting and referrals. We are all organisations based within our own communities, run by the community, for the community. The partnership has an excellent track record of working together to challenge hate speech and hate crime, both online and offline.

This partnership has provided a unique opportunity to promote a better understanding of intersectional issues and learn from each other on how to ensure the cross-cutting needs of victims of hate crime can be fully supported. It has also enabled us to stand in solidarity with each other and show unity across minoritised groups in the face of adversity.

### The organisations in the CATCH partnership are:

- Galop - Anti-LGBT+ hate crime
- The Monitoring Group (TMG) - Race hate crime
- Community Security Trust (CST) - Antisemitic hate crime
- TellMAMA - Anti-Muslim hate crime
- Choice in Hackney- Anti-Disability hate crime
- Stay Safe East - Anti-Disability hate crime
- Real - Anti-Disability hate crime

# catch

Communities Against Hate

---





---

## TOGETHER AGAINST HATE

As well as the main CATCH service delivery partnership, in 2020 we had the opportunity to pilot a new CATCH venture. Together Against Hate is an intersectional training programme aiming to upskill small community organisations facing increased hostility in the current climate.

The programme provided training to recognise and challenge hate crime, support to build an online community campaign, and opportunities to network with other community organisations, policy makers and key figures in the anti-hate crime field. It comprised four key elements: a training course, production of resources, campaigning and an expert roundtable.

### The organisations taking part all met the following criteria:

- Based in communities that have not traditionally been the focus of anti-hate crime initiatives and were facing more hostility in the current climate.
- Have a good connection with their community through their work (social support, advice work, cultural activities etc.)
- Have a significant reach into marginalised communities (e.g. not just serving a specific community in a specific location).
- Need further knowledge and skills to provide advice and support around hate crime.
- Are based in London or work with people living in London.
- Are able to disseminate the training in community networks.

The organisations represented in the pilot programme in 2020 support a wide range of communities, including: Central & Eastern European; Chinese and South East Asian; Gypsy, Roma, Traveller; Latin American; LGBT+ asylum seeker & refugee; LGBT+ Muslim; LGBT+ young people; Sikh; and Somalian communities.



---

## TRAINING COURSE

The four 1-day training sessions took place via Zoom, and included a range of specialist speakers from Galop, CATCH organisations, police, the CPS and other external organisations.

### Hate crime and working with victims/ survivors

Session 1 was co-facilitated by trainers from Galop and the Monitoring Group, with an additional Galop speaker with expertise around trauma-informed practice. The topics covered were: the nature and impact of hate crime; barriers to accessing help; addressing barriers; hate crime and COVID-19; and trauma-informed practice. The purpose of the session was to give participants a basic knowledge in hate crime and how victims may present, to enable them to recognise hate crime, give initial advice and refer appropriately.

### Hate crime strands and CATCH partners

Session 2 was facilitated by a trainer from Galop, with speakers from The Monitoring Group, Tell Mama, CST, and Stay Safe East. The topics covered included: the CATCH partnership and intersectionality; anti-Muslim Hate Crime; anti-Semitic Hate Crime; race hate crime; disability hate crime; and anti-LGBT+ Hate Crime. The purpose of the session was to give participants an intersectional overview of hate crime, provide some insight into the nuances of different hate crime strands, and to build referral connections with CATCH organisations.

### The criminal justice system and non-CJS solutions

Session 3 was facilitated by a trainer from Galop, with speakers from the Met Police, the CPS, Why Me?, and a Galop hate crime advocate. The topics covered included: policing hate crime; prosecuting hate crime; client advocacy within the criminal justice system; and restorative justice. The purpose of the session was to give participants an overview of different solutions available to hate crime victims and how to navigate the criminal justice system, and to build connections with these professionals.

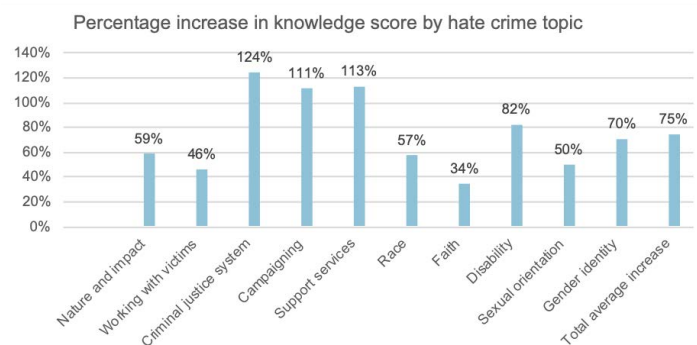
## Anti-hate crime campaigning

Session 4 was co-facilitated by trainers from Galop and the Antisemitism Policy Trust. The topics covered included: introduction to campaigns; the creative process; feedback on campaign plans; and technical online campaign tips. The purpose of the session was to give participants an introduction to the campaign planning process, show examples of successful hate crime campaigns and what could be improved, support them to begin to build their own campaign plan, and get feedback from each other on their initial ideas.

## Learning Objectives

The learning objectives for the program were for participants to gain increased knowledge on:

- The nature and impact of hate crime
- Working with victims/ survivors of hate crime
- Hate crime and the criminal justice system
- Anti-hate crime campaigning
- Support services for victim of hate crime in London
- Each hate crime area: race, faith, disability, sexual orientation and gender identity



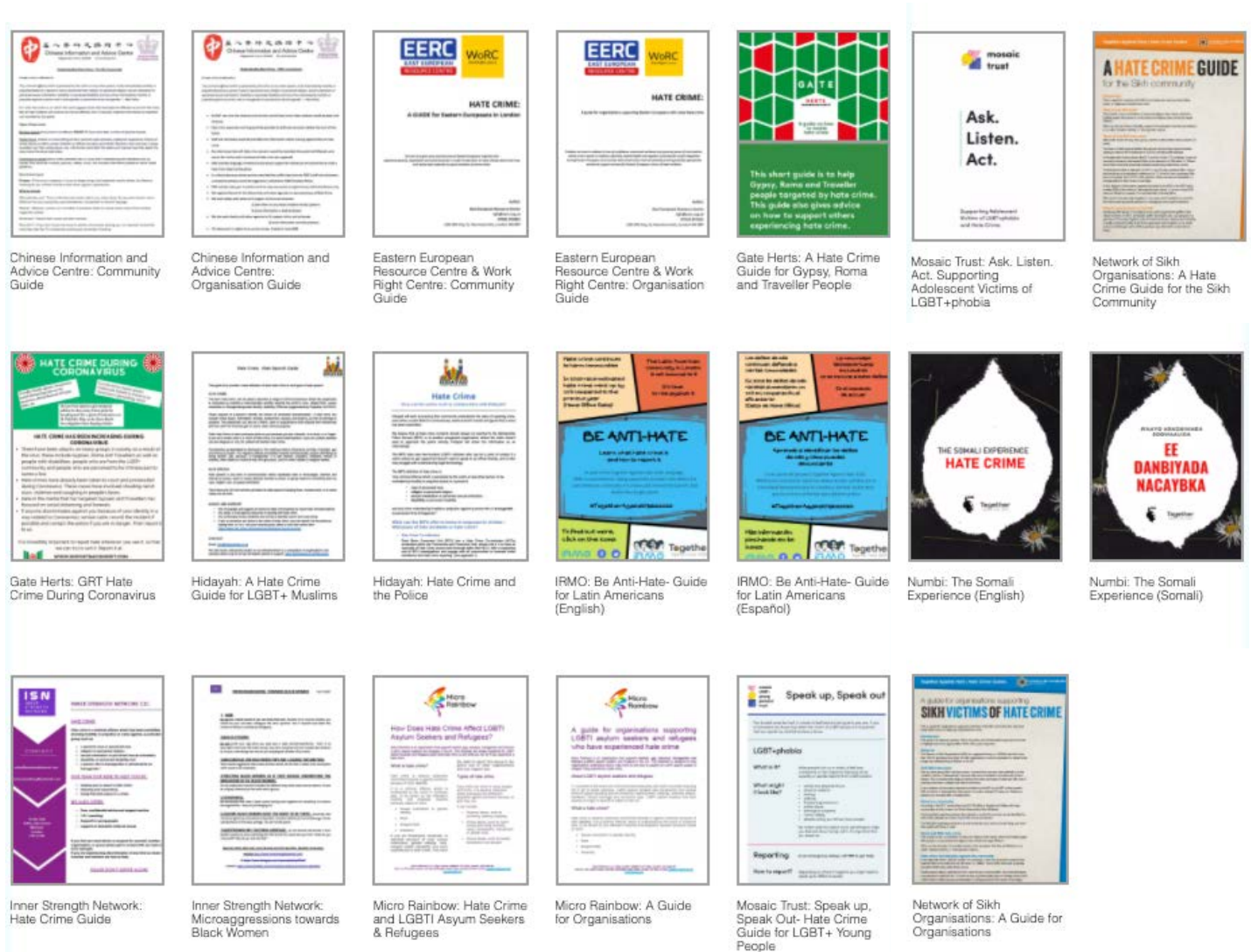
Success against each learning objective was evaluated by comparing scores in knowledge before and after the training. Participants were asked to score their knowledge in each learning objective area in a pre-training knowledge assessment survey before the beginning of the course, and then after in the post-session evaluation surveys. Participants knowledge increased in all subject areas, by an average of 75%.

---

## Community resources

Each organisation produced two factsheets on hate crime topics. The first factsheet was aimed at hate crime victims from their community, containing information about hate crime and where to get support. The focus of the second factsheet was flexible to best meet the needs of their community. Topics included: advice for professionals working with hate crime victims from their community, hate crime and COVID-19, and translations into alternate languages.

The factsheets can be viewed [here](#).



Chinese Information and Advice Centre: Community Guide

Chinese Information and Advice Centre: Organisation Guide

Eastern European Resource Centre & Work Right Centre: Community Guide

Eastern European Resource Centre & Work Right Centre: Organisation Guide

Gate Herts: A Hate Crime Guide for Gypsy, Roma and Traveller People

Mosaic Trust: Ask. Listen. Act. Supporting Adolescent Victims of LGBT+phobia

Network of Sikh Organisations: A Hate Crime Guide for the Sikh Community

Gate Herts: GRT Hate Crime During Coronavirus

Hidayah: A Hate Crime Guide for LGBT+ Muslims

Hidayah: Hate Crime and the Police

IRMO: Be Anti-Hate- Guide for Latin Americans (English)

IRMO: Be Anti-Hate- Guide for Latin Americans (Spanish)

Numbi: The Somali Experience (English)

Numbi: The Somali Experience (Somali)

Inner Strength Network: Hate Crime Guide

Inner Strength Network: Microaggressions towards Black Women

Micro Rainbow: Hate Crime and LGBTI Asylum Seekers & Refugees

Micro Rainbow: A Guide for Organisations

Mosaic Trust: Speak up, Speak Out- Hate Crime Guide for LGBT+ Young People

Network of Sikh Organisations: A Guide for Organisations

## CAMPAIGNS

Each organisation designed a campaign suited to the needs of their community, united under one hashtag, #TogetherAgainstHate2020. All the campaigns can be viewed via on Twitter, Facebook and Instagram via #TogetherAgainstHate2020.

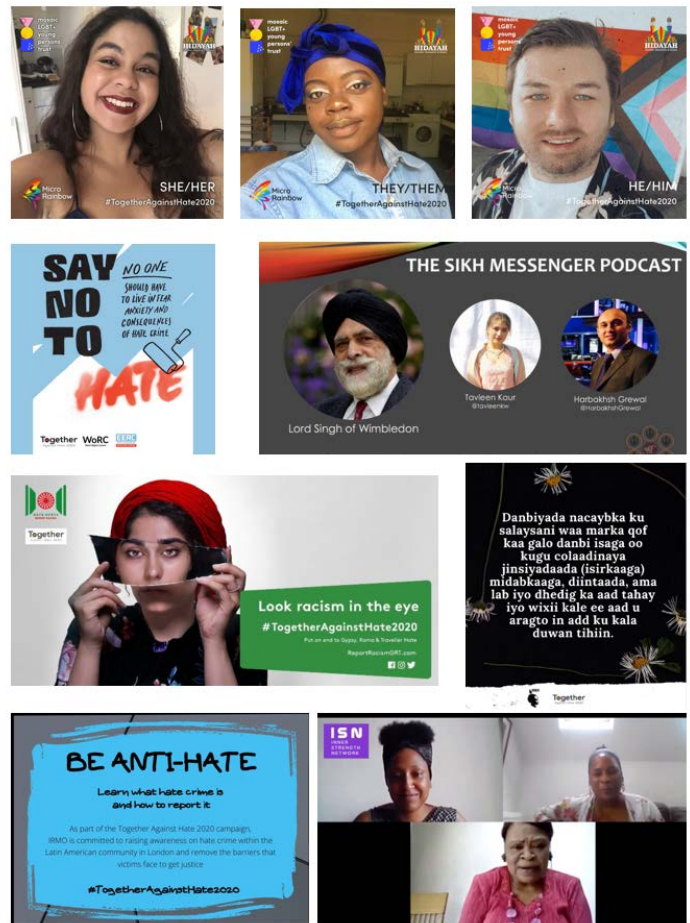
The campaigns launched on 17th August, aiming to:

- raise awareness of hate crime within communities
- launch and promote the resources created
- display unity among the diverse communities represented

The campaign methods were very varied, including podcasts, videos, photo filters, interviews and webinars. You can see some screenshot examples below. During the campaign period, the resources produced were disseminated to a total of 9,332 community members and 3,241 professionals, and collectively the social media posts made over 1 million impressions over 5 platforms: Facebook, Instagram, Twitter, Anchor and Spotify.

## ROUNDTABLE

The final roundtable took place at the end of September, in which the participants presented the needs of their communities and built links with key hate crime policy and research experts. It was a successful event, with a total of 37 attendees. Alongside the project and CATCH organisations, representatives attended from: the British Transport Police; Birbeck College/ Facing Facts; Cardiff University; Crown Prosecution Service; Government Equalities Office; Home Office; Mayor's Office for Policing and Crime; Ministry for Housing, Communities & Local Government; and the Metropolitan Police Service.



## FURTHER INFORMATION

### Make a referral

You can make a referral via our website:  
<http://www.galop.org.uk/report/>

If you have any questions about the services Galop can provide or would like to discuss a referral beforehand, you can contact Galop at: [referrals@galop.org.uk](mailto:referrals@galop.org.uk).

### General inquiries

For general inquiries, please contact: [info@galop.org.uk](mailto:info@galop.org.uk).

---

CASE STUDIES

---

# 07 Redirect North America: Challenging White Supremacist Extremism on Google Search

---

EXTRACTS FROM:

Moonshot

---

Moonshot is a social enterprise working to end online harms, applying evidence, ethics and human rights. We design new methodologies and technologies to enhance the capacity of our partners to respond effectively to violent extremism, disinformation, gender-based violence, and other online harms. Our work is rooted in the fundamental belief that people can change.

---

## THE PROBLEM

White supremacist extremism poses both a domestic and a global terror threat to countries around the world. It is an ideology based on the notion that the “white race” is threatened with extinction, the dehumanization of other races, and conspiracy theories that position particular ethnic and religious groups as “enemies.” Conspiracy theories underpinning this ideology include the belief that the white race is under attack from Jewish interests across industries and the government, which is referred to as the Zionist Occupied Government (ZOG). The virulent anti-Semitism that sits at the heart of this ideology is one of the many things it shares with jihadist organizations such as ISIS.

Instances of this form of terrorism are increasing across the globe. Norway saw the deadliest of these attacks in recent history, when a terrorist murdered 77 people in twin attacks on government buildings and the island of Utøya in 2011. And in March of 2019, attacks by a terrorist on two mosques left 51 people dead in Christchurch, New Zealand. In the past several years, we have seen these terrorists themselves become dangerous international ideologues and hate preachers. The Norway and New Zealand shooters published their own manifestos online, which serve to inspire others to act.

The internet did not create this global movement, but it supercharged its evolution. The current wave of white supremacist terrorism is intrinsically connected with the emergence of internet cultures. White supremacist extremists use technology to organize and recruit, and like jihadists, the rise of social media has provided a rich opportunity for these groups to support one another across borders.

---

## THE OPPORTUNITY

White supremacist extremist content remains highly accessible online. These groups are well aware of the legal boundaries and community standards that govern online spaces, and walk the line carefully to avoid moderation. Additionally, there are some spaces on digital platforms that remain unmoderated to a certain degree, such as search engines. Moonshot has developed a methodology, the Redirect Method, to reach users at risk of white supremacist extremism on search engines and offer them safer alternatives.

---

## THE REDIRECT METHOD

Moonshot’s Redirect Method was developed in 2015 in partnership with Jigsaw, Google’s in-house technology incubator. We worked together to design an approach that repurposes traditional marketing methods and advertising technology to reach people at risk of violent extremism.

The Redirect Method uses targeted advertising to connect people searching the internet for violent extremist content with safer alternatives. The Redirect Method aims to amplify pre-existing content made by communities across the globe, specifically to audiences that are in need of safer options to extremism.

Advertisements appear just above the organic search results and, as such, aim not to censor those results but merely to offer an alternative. When users click on the advertisement, they are taken to a range of content intended to undermine the messages that would otherwise be consumed. We have redirected at-risk users to a range of content types, including alternative and counter-narratives, de-escalation from violence, and direct access to online or offline services.

Moonshot has to date deployed the Redirect Method internationally in over 30 countries, in more than 20 languages, and in partnership with tech companies, governments and grassroots organizations. Since 2018, Moonshot has begun piloting deployments of the Redirect Method to reach people searching the internet for other kinds of dangerous content, including that related to gender-based violence, disinformation, human trafficking, and child sexual exploitation.

---

## CASE STUDIES

This document covers two case studies of deployments of the Redirect Method in North America to reach white supremacist extremist audiences.

### Redirect USA

In 2019, Moonshot launched a deployment of the Redirect Method across the United States in partnership with the Anti-Defamation League and the Gen Next Foundation. The campaigns were rolled out on Google Search across the entire country between May and November 2019. During this period, safer alternatives were offered over 179,000 times to people searching for white supremacist extremist content, resulting in over 4,000 clicks through to safer options.

### Redirect Canada

In 2019, Moonshot launched the Redirect Method in Canada with funding from the Community Resilience Fund and in collaboration with the Canada Centre for Community Engagement and Violence Prevention at Public Safety Canada. Redirect Canada was first deployed across all 13 Canadian provinces and territories, and in June 2019 Moonshot's campaigns were subdivided to incorporate 353 postcodes in Canada's six largest cities. These localized campaigns enabled Moonshot to collect granular data on extremist search appetite, test experimental messaging, and explore the viability of providing at-risk users with access to services in their communities.

---

## METHODOLOGY

The Redirect Method uses online advertising and curated content uploaded by people all around the world to confront online radicalization. It targets potential extremists most susceptible to messaging and redirects them towards curated content that undermines extremist themes.

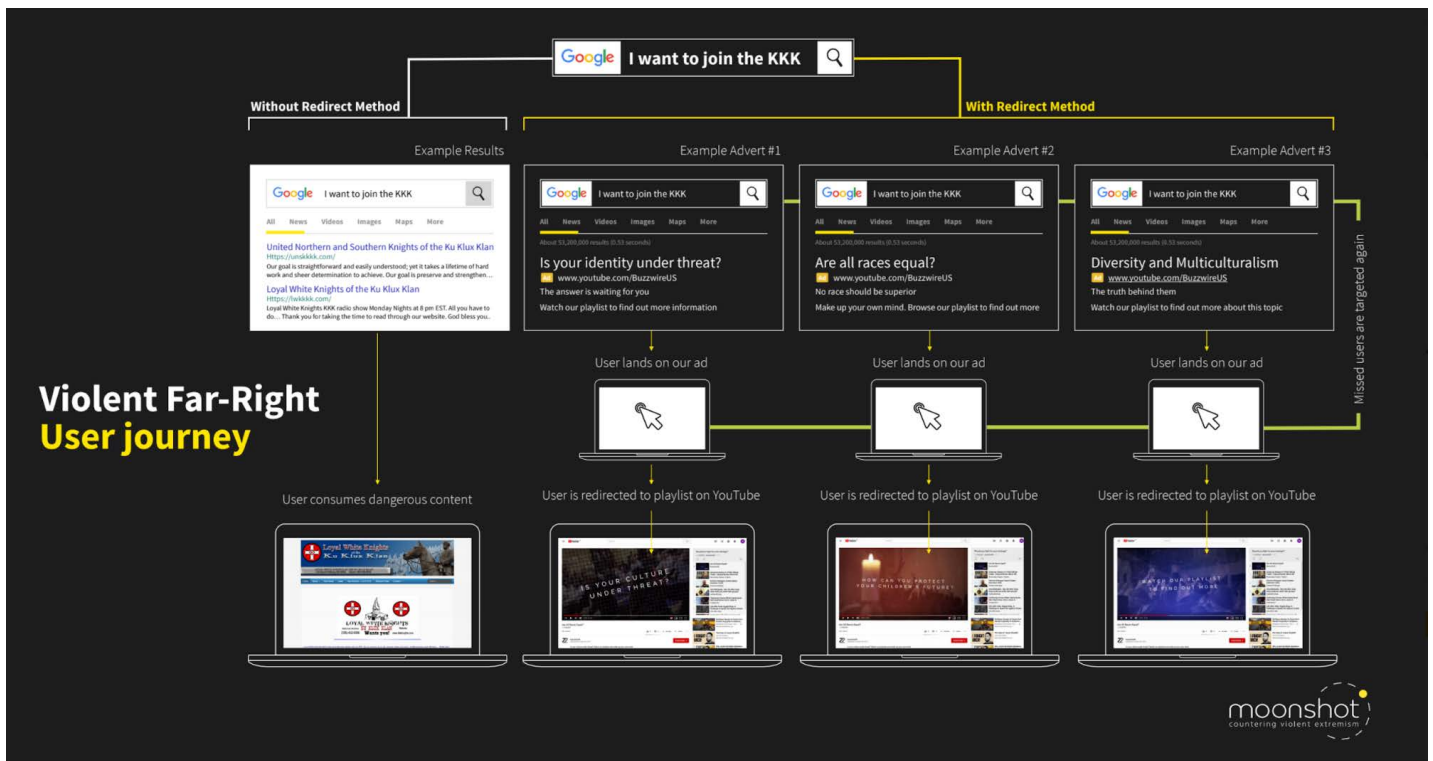
When someone performs a Google search using a keyword that demonstrates their curiosity or engagement with a violent ideology or movement, such as "Join Blood and Honor," they are taken to a Google results page which features one of our ads. These ads appear above organic search results, which oftentimes may include violent content. They hence act to safeguard a user's search by offering a contextual, credible, and safe alternative. In both Redirect USA and Redirect Canada, Moonshot's ads directed users to playlists of tailored counter and alternative content on YouTube.

This open-source methodology respects users' privacy and has been deployed to tackle a wide range of types of extremism. The methodology has been broken down into 44 detailed steps, which can be accessed and followed here. There are three primary phases to deployment of the Redirect Method: Research, Curation and Targeting.

## 1. Research

The first phase in any successful positive intervention is to research the group one is seeking to counter, and deeply understand their recruitment tactics, narratives, and appeal. Once this is complete, these insights need to be translated into indicator databases, to ensure the campaigns can effectively target and reach those most at risk. Depending on the platform involved, these indicators could take a wide range of forms. The two campaigns featured here were carried out primarily on Google Ads, so the primary indicators consisted of keywords.

These keywords indicate searchers' curiosity or engagement with violent extremist content, ideology, groups and personalities. For both projects, Moonshot consulted with external subject matter experts and researchers, including former far-right extremists, to add new indicators, which were also drawn from extremist groups' websites and forums, and counter-extremism resources.





## 2. Curation

The Redirect Method relies on curating and surfacing existing content rather than spending time and resources creating new material. During the curation phase of any Redirect Method project, the implementer should use the key narratives identified during the research phase to gather content that counters, undermines, or simply provides a safer alternative to these narratives. For the two case studies highlighted here, Moonshot used snowball sampling to identify suitable content.

We also recognized that content that was not created for the purpose of counter-messaging has the capacity to undermine extremist narratives when curated, organized and targeted effectively. Community activists and other experts were engaged to make recommendations around effective content. Our approach to selecting content aimed to mitigate the risk of lower retention rates resulting from bait-and-switch advertising, in which individuals are presented with content that differs significantly from what they were searching for.

Using this methodology, Redirect USA and Redirect Canada aimed to reach at-risk users with content that aligned as closely as possible with what they were searching for.

## 3. Targeting

During the targeting phase, the campaigns go live on the selected platform. The exact mechanisms depend on the platform, and for the two case studies the primary advertising platform was Google Ads. During this phase, the indicator database is uploaded to the platform to ensure that the campaign reaches only those at risk. The curated content is arranged and uploaded with accompanying ad text.

For both Redirect USA and Redirect Canada, Moonshot served content to search users using two unique variables: risk and content type. Every search keyword in our database was coded using a proprietary Risk Matrix. Risk factors included the use of specific words or phrases that imply a searcher's prior knowledge of extremist ideology or media (e.g. "NSBM Québec"; "download the Turner Diaries PDF"), or searches indicating an intent to join a group or commit acts of violence ("I want to kill blacks"; "how to join the KKK"). Moonshot then matched searches with playlists of video content that implicitly or explicitly challenge an extremist narrative or tenet: for example, the keyword "Brenton Tarrant video" was coded in a Videos Redirect category, while "the Kalergi Plan" and "genocide of whites" in the Conspiracy Theory category. Other content types included Slogans and Symbols, News, Influential Personalities, Music, and Literature.

## FINDINGS

### Redirect USA

Redirect USA ran in all 3,142 counties and county-equivalents in the United States between May and November 2019. Our primary objective was to increase the likelihood that those at-risk would consume safer options. Overall, our advertisements were shown to audiences searching for white supremacist content on Google over 179,000 times.

The deployment resulted in 179,684 impressions (the number of times the ad was shown to at-risk individuals), a significant number of clicks on the ads, and a high click-through rate (the percentage of clicks per impressions) and search impression share (the percentage of time an ad was shown every time an eligible user searched for an at-risk keyword).

In all, those searching for white supremacist extremist content consumed 5,509 minutes of videos undermining those messages — time that could have been spent consuming violent extremist content instead.

Metric	Redirect USA
Impressions	179,684
Clicks	4,295
Avg. Click-Through Rate	2.39%
Avg. Search Impression Share	95%

The program also yielded considerable insights into online propaganda more broadly. For example, those at risk of white supremacy often express a clear interest in consuming music by white supremacist bands such as Blue Eyed Devils and Vaginal Jesus, and in looking up influential personalities ranging from a neo-Nazi Canadian YouTuber to the perpetrator of the Christchurch massacre — and, of course, Adolf Hitler.

The aftermath of the tragic shooting in El Paso, Texas, during the time of the Redirect Method deployment illustrates the program's promise. After that tragedy, the campaign saw a 104% increase in impressions related to white supremacist extremism and a 59% increase in clicks. This effect was even more significant in El Paso itself, where a 192% increase in impressions was observed. The researchers measured a 224% increase in watch time for a playlist designed to undermine the white supremacist narrative of "Fighting for white heritage." This means that at-risk users looking for content based on searches such as "Prepare for race war" consumed alternative content at a higher rate in the aftermath of the attack. The increased demand for extremist content was surpassed by an increased willingness to engage with content that undermines such messages.

### Redirect Canada

Redirect Canada aimed to reach at-risk users with content that aligned as closely as possible with what they were searching for. We aimed to mitigate the risk of lower retention rates resulting from bait-and-switch advertising, in which individuals are presented with content that differs significantly from what they were searching for.

We aimed to reach audiences at risk of white supremacy across Canada who might be searching for a range of content types, such as music, gaming and literature, to deliver alternative messaging that matched the content type but offered safer alternatives. The campaign was deployed for over a year, from February 2019 to March 2020.

Across Canada, Moonshot's campaigns targeting searches related to the white supremacist extremist audiences redirected over 150,000 English language searches and over 3,000 French searches. Engagement with our ads resulted in over 2,000 views of playlists curated to challenge or provide an alternative to online extremist narratives.

Metric	Redirect Canada
Impressions	155,589
Clicks	2,234
Avg. Click-Through Rate	1.44%
Avg. Search Impression Share	98.89%

---

## EXPERIMENTATION

In both Redirect USA and Redirect Canada, we ran constant experimentation and A/B testing to continuously improve the campaigns, resulting in double and even triple-digit percent increases in search traffic captured and users engaged during the life of the campaigns. More than mere statistics, these achievements demonstrate a tighter focus on the at-risk population and a higher likelihood of redirecting these users away from harmful content online.

During our Redirect USA deployment, we worked to build a highly complex campaign infrastructure to run county-by-county campaigns for delivery of highly localized online prevention and intervention programming.

### Three main learnings emerged from our experimentation:

- 01** Localizing advertisements (including in the ad text the name of the city where the searcher is) increased engagement with Redirect messaging compared to ads that were not customized to a search user's location.
- 02** Including the keyword searched by a user in the ad text increased the likelihood that an at-risk user clicked on the ad. Interestingly, while this increased engagement by white supremacist extremists, it decreased engagement when we ran this experiment with audiences at risk of Jihadism.
- 03** We tested a hypothesis that customizing our alternative content to closely match the exact content that users were searching for (e.g. offering someone searching for "white genocide" safer information specifically challenging theories of white genocide) would keep redirected users more engaged. As it turns out, this did not increase average view duration. It also did not have an effect on users' average time spent in our playlists. This trend was statistically insignificant, however, so this method would benefit from further testing over a longer period of time.

---

## LEARNINGS

Over the course of many months, Moonshot safeguarded hundreds of thousands of searches across the US and Canada in two ideologies and multiple languages. We accomplished this work through productive collaboration with local partners, former extremists, translators and subject matter experts.

### Audience Lessons Learned

- In both the US and Canada, internet users aged 25-34 are the most interested in white supremacist content. Consequently, future positive intervention campaigns could place greater emphasis on messaging to this group, including for instance ad text that addresses the unique challenges individuals face at this age.
- Conspiracy theories that fuel violent white supremacist extremism ideologies are extraordinarily popular and must be challenged with new and creative solutions. Searchers sought information on white supremacist conspiracies such as the Kalergi Plan, The Great Replacement, and White Genocide at an alarming rate. These conspiracy theories can provide an entry point to the curious and cognitive reinforcement to the committed. While providing tailored, conspiracy-specific playlists did not generate a statistically significant increase in view time, other approaches should be tested and applied to deal with this challenge.
- Users seeking information specifically on white supremacist extremist groups (for example, seeking information about joining a group) were disproportionately likely to engage with alternative content offered by the Redirect Method. Over the course of the Redirect Canada campaigns, users looking for information on violent white supremacist groups clicked through to our content at more than twice the average of all categories (4.1% CTR, compared to an average of 1.4%). This suggests that individuals who are seeking to learn about, or engage with, these groups are also more willing to engage with counter and alternative content. This could be a valuable opportunity for future campaigns and interventions.
- Music provides a unique opportunity to keep the attention of users at risk of white supremacist extremism. Specifically in Canada, based on average watch-time, our Music playlist sustained the attention of viewers significantly more than others (an average of 104 seconds, compared to an average of 66 seconds). Future campaigns should build on this success, and continue to experiment with new methods to engage these users.

## Campaign Lessons Learned

- Risk assessment and risk-rating are critical to conducting ethical strategic communications, and their rigor can and should be evaluated. Moonshot verified the rigor of its risk-rating system during these Redirect Method deployments through structured inter-rater reliability testing managed by an external monitoring & evaluation consultant. This approach evaluated not just the risk-rating system's logic, but also how the ratings were applied by multiple raters. The evaluation confirmed that both the logic and process met customary social science standards. Online programming should not be exempt from such standards or scrutiny.
- More and better indicators for campaign targeting (for example, extremist keywords) are essential to success. While Moonshot steadily enhanced and optimized many aspects of the campaign, no single improvement had a greater impact than the continuous expansion of our indicator databases. This expansion, borne of intensive research and collaboration with experts and translators, fueled a large increase in impressions and clicks during the last six months of Redirect Canada. While more complex and sophisticated marketing strategies may show promise, there is no substitute for being able to be where your audience is. For the Redirect Method, that means using extensive databases of indicators for your targeting, to advertise as widely and deeply as possible.
- Customization and targeting do not guarantee success. Digital marketing strategies understandably emphasize the importance of a rich user experience, often achieved by providing the user with hyper-relevant content. We tested this theory during our experimentation phase by creating more customized ads and playlists, but found that viewership did not increase to a statistically significant degree, depending on the type of customization. This finding reminds us that ideas that make intuitive sense are not necessarily proven out empirically.

---

CASE STUDIES

---

## 08 The Online Civil Courage Initiative (OCCI)

---

EXTRACTS FROM:

The Institute for Strategic  
Dialogue (ISD)

---

**The Institute for Strategic Dialogue** (ISD) is a global think-and-do tank registered as an English Charity and a French Association. ISD designs innovative responses to polarisation, hatred and extremism in all its forms. ISD leverages its expertise in research, digital analysis and extremist movements in order to design resources, training programs and interventions for different audiences and in partnership with civil society organisations, academia, businesses, influencers, credible voices and public decision-makers. ISD has over a decade of experience in researching and challenging extremist groups, understanding their motivations and playbooks, monitoring disinformation and mapping hate online.

The OCCI is a strategic partnership between the ISD and Facebook that began in 2016. The programme launched in Germany in 2016 and expanded to France and the UK from 2017 onwards. It operates under four pillars of work: research, training, community and support, and aims to combine technology, communications, marketing and academic expertise to bolster the civic response to online hate and extremism. These tools and skills are often out of reach for grassroots or activist organisations, and the OCCI seeks to fill that gap.

The ultimate aim of the OCCI is to unite a diverse set of actors, provide them with the latest research into hate and extremism online – and resources on campaigning and digital citizenship education – and foster a collaborative environment via conferences and hackathons. This enables them to have a greater impact together than they would individually.

Activities of the OCCI in France are informed by consultation with our Steering Committee, which includes representatives from the following organisations:

- American Jewish Committee Paris
- Civic Fab
- Contre discours online
- Génération Numérique
- Institut français de Géopolitique
- Institut français de Relations Internationales
- Inter-LGBT (an umbrella group of 50 lesbian, gay, bisexual and trans organisations in France)
- Le Refuge
- Ligue Internationale Contre le Racisme et l'Antisémitisme (Licra)
- Ligue des droits de l'Homme
- No Hate Speech Movement in France
- Renaissance Numérique
- SOS Racisme

---

## HISTORICAL ACTIVITIES OF OCCI

Since 2016 and across all three countries, the OCCI has:

- released **26 research reports**, which provide in-depth and timely insights to practitioners, including analysis of extremist propaganda, hateful discourses and the co-option of current events by hateful groups. Feedback from network members indicate these reports have provided useful context for their campaigns and other activities, keeping them briefed on the latest online trends.
- organised **27 conferences and roundtables**, bringing together diverse actors from the three countries – and further afield in Europe. These events helped members to share knowledge and research on countering extremism and polarisation, and to foster innovative partnerships between like-minded or complementary groups. Over 1,000 practitioners have been trained, with participants consistently reporting increased knowledge and skills as a result.
- established vibrant communities of CSO actors who combat diverse types of hate and extremism. **Dozens of collaborations** have emerged from this network, including the French chapter of the initiative #IAmHere: #JeSuisLa.
- supported **45 counterspeech campaigns**, targeting a wide range of audiences – from vulnerable, at-risk groups to the general public – which have **reached almost 60 million people** with targeted, positive messaging to fight hate and extremism. The OCCI has also created two counterspeech campaigns as part of one-day hackathons – engaging over 2 million individuals to date.



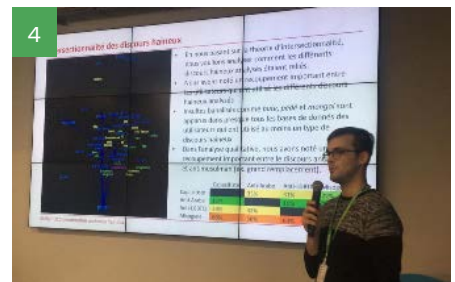
Picture 1 and 2: Pictures taken at the OCCI European Summit in Paris, in April 2019.

**CASE STUDY:  
SPOTLIGHT ON #JESUISLA (#IAMHERE)**

The online civic initiative #JeSuisLa (Picture 3) started at an OCCI workshop hosted by ISD and Facebook in 2018, which focused on tackling hate-fueled disinformation and addressing the limits of content moderation. At the event Shani Benoualid, Head of Communications at American Jewish Committee (AJC) Paris, and Xavier Brandao, founder of grassroots organisation Répondre aux Préjugés (Respond to Prejudices), met and brainstormed campaign ideas. They discussed their desire to encourage people who see hate online to challenge it. They then decided to create a French version of the [#IAmHere movement](#), which was initially launched by Mina Dennert in Sweden, and to adapt its strategy to a French context. Within a month, their Facebook group gathered thousands of members and a strong pool of hundreds of volunteers coordinating ‘love bombs’ to flood toxic comment threads with shows of support to victims of harassment, or to offer direct challenges to hateful, divisive and misinformed narratives with fact-based and empathy-driven arguments.

**i. Research**

In 2019, OCCI France released its first long-form research report, [Mapping Hate in France](#), the most comprehensive study analysing online hateful speech at the national level. The research used natural language processing systems, spanning 11 different categories of rhetoric, and analysing – among others – hateful misogynistic, homophobic, anti-Arab and ableist speech. The insights of ISD’s research were presented at OCCI events, providing participants with insights into the main trends in the online ecosystem as well as cross-sectoral policy recommendations to respond to online harms. The presentation of these insights contributed to increasing the OCCI network’s subject matter knowledge (see Figure 1 below) - an average of 7% increase.



Picture 3: The Home Page of #Jesuisla

Picture 4, 5 & 6: The launch event of the research report *Cartographie de la Haine en Ligne* ('Mapping Hate in France'), in November 2019



## ii. Trainings

In 2019, The OCCI organised three roundtables and conferences, **training over 100 participants**, on topics including disinformation and content moderation. All events included a series of presentations, from ISD, Facebook and members of civil society organisations, as well as collaborative group discussions which resulted in the development of cross-sectoral recommendations.

The first roundtable of the year, focusing on disinformation, took place in the aftermath of the European elections. ISD presented the main trends of disinformation from the European Parliamentary elections from our research, and participants from different backgrounds came together to discuss innovative solutions to counter disinformation online. Figure 2 demonstrates the impact of this roundtable, with participants increasing their knowledge with regards to disinformation online. The second round table on content moderation focused on hateful speech online, with the presentation of key findings from ISD’s Mapping Hate Online. Facebook and CSOs (including Licra and #JesuisLa) introduced their approach to tackling online hate. The founder of the app Bodyguard (an app that allows users to automatically set moderation standards for their social media profiles) also presented his tool as a new innovative solution to counter hateful content online.

The year ended with the **OCCI’s second European Summit**, held jointly with the European Commission’s Code of Conduct Dialogue. This event included a series of presentations by the European Commission, ISD, platforms and European CSOs, as well as a creative Hackathon which produced a European campaign to counter online hate.

As a result of these trainings, the OCCI helped CSOs establish links with each other and with key actors in the sector, including government agencies and tech companies. The figure below demonstrates participants’ level of satisfaction with the OCCI European Summit, which took place in Paris in December 2019.

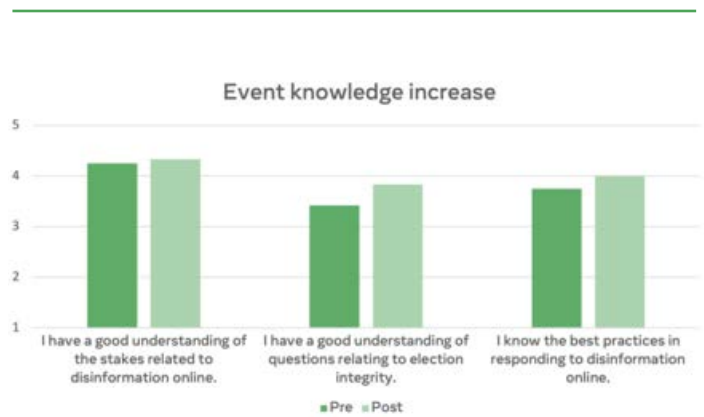


Figure 1: Example of the results of a survey demonstrating knowledge on disinformation before and after an OCCI event, in June 2019.

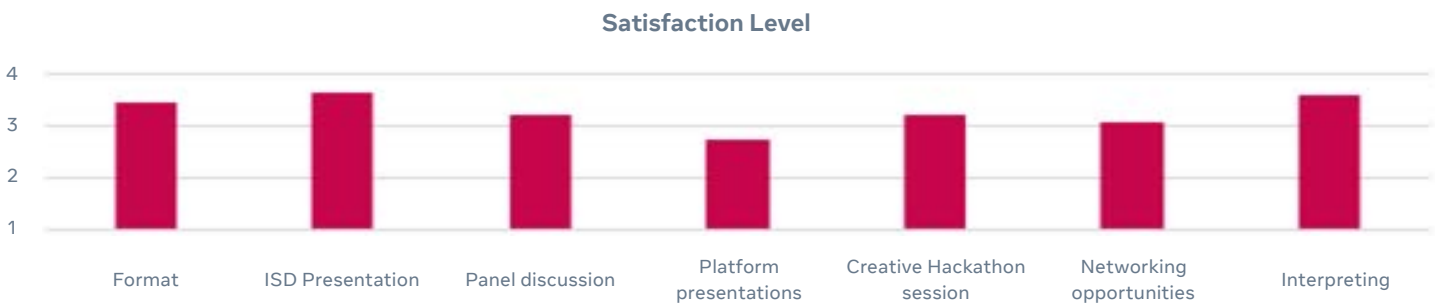
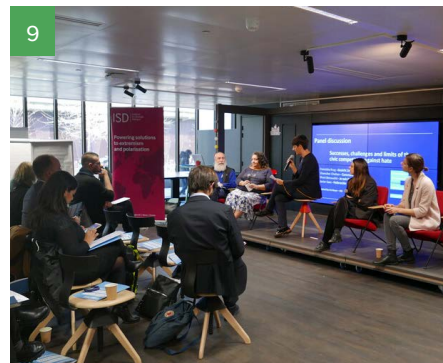


Figure 2: Satisfaction levels of participants with the OCCI European Summit in December 2019

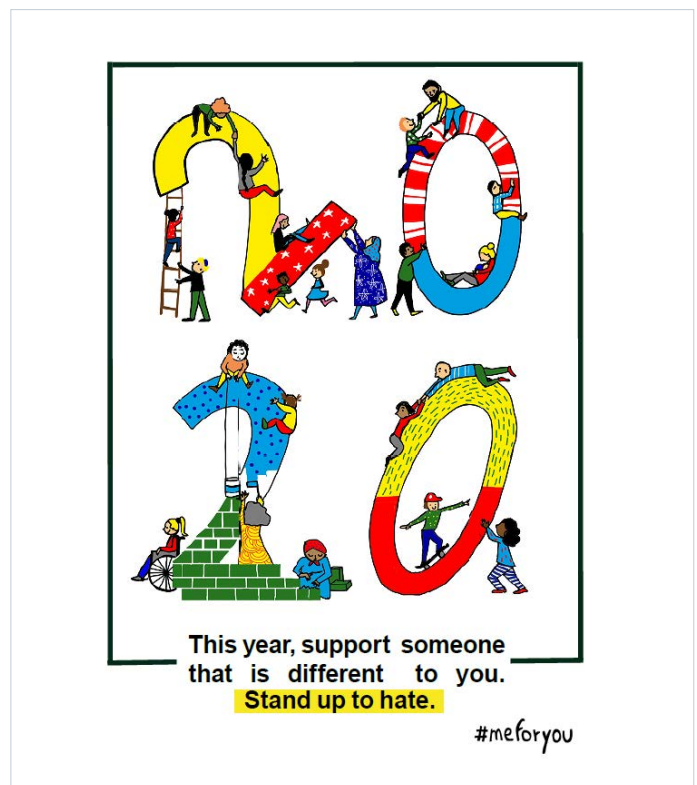


Photos 7, 8, 9, 10: Pictures taken at the second OCCI European Summit, in Paris 2019

### iii. Campaigns

The OCCI supported two coordinated campaigns to fight hate and disinformation online in 2019. One was in collaboration with #JeSuisLa to counter disinformation surrounding the European parliamentary elections and the other was with the European Commission Code of Conduct Monitoring Group to encourage positive and responsible engagement on social media. These campaigns have reached tens of thousands of users across social media platforms and across Europe.

The campaign developed in partnership with the European Commission Code of Conduct aimed to counter online hate. It resulted in seven campaign creatives which promoted standing up to hate online based on the theme of New Year's resolutions (as the campaign was shared on social media in early January 2020). This campaign was shared online by a collective of European CSOs which created the campaign during the hackathon session.



Campaign 1: Example of campaign creative from the OCCI European Summit campaign, 2019. This creative was shared by several European CSOs including the German Amadeu Antonio Stiftung which shared it both on Facebook and Instagram (leading to 16,000 impressions on FB and 7,500 on Instagram).

## OCCI CAMPAIGN SPOTLIGHTS:

### Entre SœursSoeurs (Between Sisters)

The OCCI worked with documentary filmmakers and activists countering Islamist extremism online. The filmmakers had interviewed French and Belgian female returnees from the Islamic State of Iraq and Syria (ISIS), who were highly skilled in engaging in online communities and discourse. Together they created an engaging Facebook page targeting young French women with snapshots of these interviews, attempting to dissuade them from joining Daesh by deconstructing the group's propaganda



(Campaign 2).

Campaign 2 : Graphic from the Entre Sœurs campaign

### Et toi, le Jihad?

Social enterprise [Civic Fab](#)'s project coordinator is a passionate cartoonist. She met the founder of [Et toi, le Jihad?](#) at an OCCI creative hackathon and started to volunteer and draw comic strips for this counterspeech



initiative, challenging Islamist extremism through satire.

Campaign 3: Illustration from Et toi, le Jihad?

### L'Association Zy'Va and France Fraternités

The founder of l'Association Zy'Va, a grassroots association, and a video producer at NGO France Fraternités met at an OCCI workshop and collaborated on a campaign promoting youth activism and local social engagement (Campaign 4).



Campaign 4: Screen capture of a post from the Zy'Va and France Fraternités collaboration

---

### 3 KEY LEARNINGS AND RECOMMENDATIONS:

The OCCI shows how the impact of individual organisations can be augmented through ongoing research and support, and by creating a diverse activist network. It has helped equip hundreds of practitioners with campaigning skills and improved understanding of trends in online extremism. More can still be done – for example, network members flagged the need for more technical and creative support to campaigns, combined with ongoing graduate training programmes and cutting edge research – but the OCCI has already demonstrated that it has had a durable impact on CSOs in France and across Europe.

### Several main recommendations were drawn from the OCCI programming:

- 01 For funders, ISD suggests funding bodies and other grant-making organisations should directly support CSOs that can tackle online harms creatively. CSOs have credibility that is based on their prior expertise and/or position in the relevant community, and are thus well placed to respond imaginatively to challenges as they emerge. Moreover, when analysing and communicating the impact of any CSO, funders must support in monitoring and evaluating them by providing frameworks and funding to do so. This will facilitate more strategic responses long term, including targeted investment and mobilisation around common aims. As the results of social good initiatives are often difficult to track or quantify, monitoring and evaluating in this sector can be piecemeal, and many organisations lack the expertise or resources to conduct rigorous, in-depth evaluations. CSOs need greater support – both financial and technical – to conduct thorough analysis of their efforts, including theories of change, viable qualitative and quantitative metrics, and appropriate data-gathering methods. These frameworks should be integrated into projects from the outset, rather than thrown together retrospectively.
- 02 Technology companies and social media platforms should provide in-kind marketing, analytic and technical support to the under-resourced civil society sector. Bolder and more sustained investment into multi-stakeholder frameworks is long overdue, which will help spearhead innovation in civic tech solutions.
- 03 Finally, CSOs need to seek out partnership opportunities, exploring how existing initiatives can complement each other and scale impact. Such a process should be coordinated by funding bodies, who often have a sector-wide perspective and are therefore well placed to broker links. It is important to share key learnings and collaborate in order to improve impact across the sector.

---

## APPENDICES

# The Institute for Strategic Dialogue (ISD)

---

### BIOS OF THE MAIN PRACTITIONERS

#### Iris Boyer - Deputy Head of Technology, Communications and Education

Iris oversees a number of programmes that support and amplify civil society's efforts against extremism through scaled partnerships with tech companies and grassroots organisations, including the OCCI France. Iris co-ordinates ISD networks that span government, academia, the media and the non-governmental organisation (NGO) sector, briefing them on ISD's latest insights into extremism and the most effective and innovative approaches to tackle related trends. Iris holds a five-year diploma in social sciences and humanities from Sciences Po, as well as an international Master's degree in Public Affairs from the Higher School of Economics in Moscow and the London Metropolitan University.

#### Cooper Gatewood - Senior Manager, Digital Research

Cooper is a senior manager within the Digital Research Unit of ISD, focusing on quantitative research into the spread of hateful and polarising narratives online, and how they are leveraged by extremist actors. Cooper is currently contributing to ISD's research on disinformation campaigns, particularly those that aim to influence and disrupt election processes. He also manages the OCCI in France, co-ordinating activities to support civil society's response to hate and extremism online. Cooper holds a Master of International Affairs from Columbia University and a Master in International Security from Sciences Po.

#### Cécile Guerin - Co-ordinator, Digital Research

Cécile Guerin is a coordinator at ISD, supporting the organisation's European development and analysis work. She works on the OCCI in France, as well as contributing to ISD's research and policy work, with a focus on social media analysis and network mapping related to hate speech, extremism and disinformation online. She has written for a range of publications, including the Guardian, Prospect and the Independent. Cécile holds an MSc in International History from the London School of Economics and an MA in English from the École Normale Supérieure in France.

#### Zoé Fourel - Associate

Zoé is an associate at ISD, working predominantly on the OCCI in France, for which she contributes research and co-ordinates on-the-ground activity. She also supports other ISD programmes that focus on empowering civil-society-led responses to hate and extremism. Zoé holds a five-year diploma from Sciences Po Lyon in International Affairs, which included studies at the School of Oriental and African Studies in London and Georgetown University in Washington, DC.

### LINKS TO RELEVANT PUBLICATIONS

#### Mapping hate in France

A panoramic view of online discourse, Cooper Gatewood, Cécile Guerin, Iris Boyer, Zoé Fourel <https://www.isdglobal.org/ISD-publications/mapping-hate-in-france-a-panoramic-view-of-online-discourse-2/>

#### Information Manipulations Around COVID-19

France Under Attack, Iris Boyer, Théophile Lenoir <https://www.isdglobal.org/ISD-publications/information-manipulations-around-COVID-19-france-under-attack/>

#### OCCI report

Building Digital Citizenship in France: Lessons from the Sens Critique project, Cooper Gatewood, Iris Boyer <https://www.isdglobal.org/isd-publications/fostering-civic-responses-to-online-harms/>

#### French language education resources

<https://www.isdglobal.org/ISD-publications/young-digital-leaders-2019-curriculum-all-languages/>

<https://www.isdglobal.org/isd-publications/young-digital-leaders-2020-ydl-parent-guide/>

<https://www.isdglobal.org/isd-publications/digital-citizenship-education-programming-toolkit/>

### CITATIONS

- <sup>1</sup> For more information on the OCCI see: <https://www.isdglobal.org/programmes/communications-technology/online-civil-courage-initiative-2-2/>
- <sup>2</sup> <https://www.facebook.com/groups/359820924602583/>

---

CASE STUDIES

---

## 09 Together #AgainstOnlineHate in Austria

---

EXTRACTS FROM:

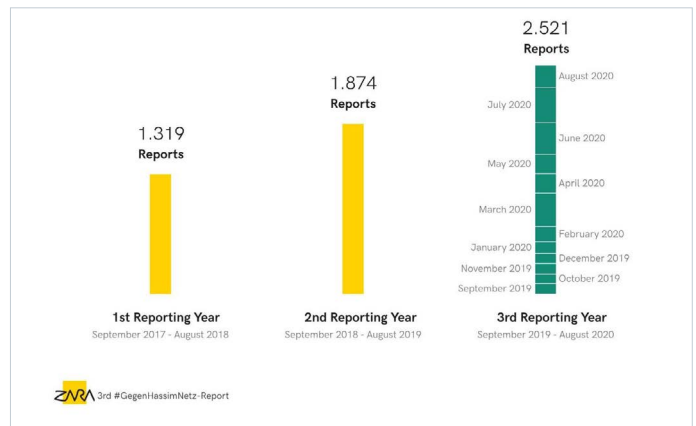
ZARA

---

The acronym ZARA<sup>1</sup> (Zivilcourage und Anti-Rassismus-Arbeit) stands for civil courage and antiracism work. We are a non-governmental organization (NGO) founded in 1999. Since 2000, we provide free legal and psychosocial support to those affected by and witnessing racism (and other intersecting forms of discrimination e.g. religious discrimination).

## IN SHORT, WHO IS ZARA AND WHAT DO WE DO?

ZARA is also at the forefront in the fight against online racism and online hate speech in Austria. Since 2017, we have broadened our mission and now cover further grounds of discrimination (for example gender, age, sexual orientation, disability, etc.) as well as cyber-bullying, cyber-stalking and other forms of online hate speech. Our team #AgainstOnlineHate (#GegenHassimNetz)<sup>2</sup> supports those affected by and witnessing online hate. According to ZARA's working definition, online hate includes inflammatory or hateful content, which is published on online platforms and social media, directed against a person or a (socially constructed) group, for example, as a result of their sexual orientation, their gender or disability.



Across the three years during which we have been active in this sphere, the number of cases of online hate reported to ZARA has increased from 1,319 counted over the period from September 2017 to August 2018 to now 2,521 from September 2019 to August 2020 (see Figure 1).<sup>3</sup> The current pandemic and the parallel transfer of public life online together with the strengthened #BlackLivesMatter movement led to a stark increase in the number of reported cases. In March, June and July 2020, the numbers have doubled or even tripled. The majority of cases of online hate speech, however, still remain unreported. The true incidence of online hate speech is therefore vastly underestimated by these numbers.

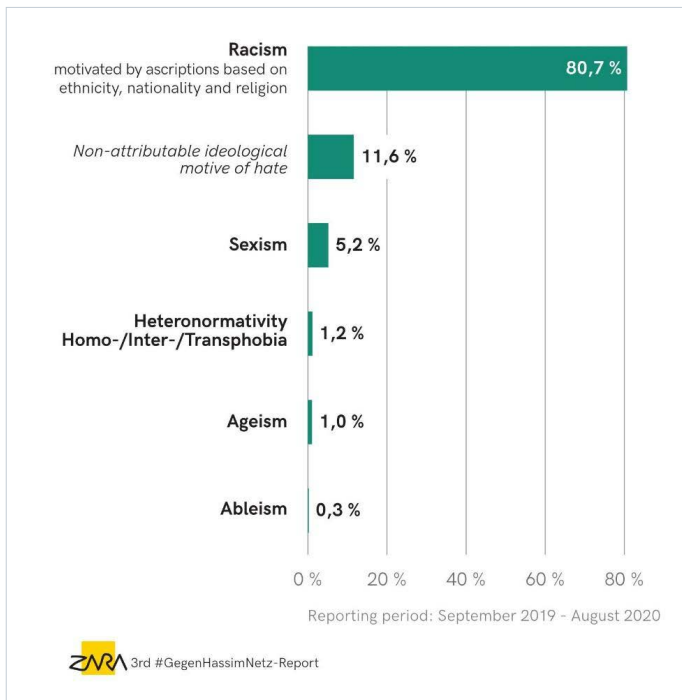
Our mission is to combat racism and online hatred and to promote, strengthen, and increase the value of civil courage within the Austrian society. We advocate for an inclusive society, in which individuals can live free from discrimination and exclusion and we promote a cyber space of freedom of speech which is guided by the principles of human dignity, respect and diversity.



## OUR TARGET GROUPS

Online hate is pervasive and threatening. The main ideological motives behind online hate speech reported to us (see Figure 2) are anti-Muslim racism, anti-black racism and racism directed at refugees. 80.7% of all submitted reports in the latest reporting period (09/2019-08/2020) concern these motives. Ranked second is a cluster of 11.6% of reports which cannot be attributed to an ideological motive and ranked third are 5.2% reports which can be attributed to sexist motives. Often these cases of online hate speech are directed against women. They receive threats of rape, are sexually harassed and stalked - this is extremely stressful for those affected. Further, 1.2% of reports document cases motivated by heteronormativity (these include homo-, inter- and transphobic cases), 1% document cases of ageism and 0.3% ableist online hate speech.

Apart from the people directly affected by online hate speech we seek to engage and address citizens, the civil society, the media, social media platforms, politicians, education and research bodies, administrative bodies, firms, influencers, people in the public eye, donors and international stakeholders, such as the International Network Against Cyber Hate (INACH) or the EU Commission.



## HOW DO WE COMBAT HATE SPEECH ONLINE?

We have a number of legally and psychosocially trained advisors in our team. Apart from documenting the reported cases of online hate speech, we provide confidential advice and support to those affected by online hate speech and map out individual empowerment strategies. These free services are available by encrypted email (via a secure platform called Aygonet), chat, telephone and face-to-face. If needed, an interpreter can be called in.

More specifically, the scope of our services covers: providing general information on online hate speech, documenting a case in our database, giving a general and especially a legal assessment of a case, informing on possible legal steps, briefing on and support in dealing with authorities, writing intervention letters and legal statements and monitoring the development of the case. Another important resource is our trusted flagger status on social media platforms. This allows us to seek rapid removal of hateful content and is a significant lever when the common flagging channels do not yield removal.

The most important task is, however, to provide psychosocial support. First, we listen and validate the situation of those affected, who very often feel isolated and scared. Then, we make suggestions on how the client can create an (online or offline) safe space. Together we devise individual counter strategies and ways to empower people against online hate. Further, we work in close cooperation with other CSOs and support providers to whom we refer our clients for further support if needed.

A few words on the Austrian legal context we operate in: In Austria, certain cases of online hate speech constitute a criminal offence. The Austrian criminal law foresees exceptions from the principle of freedom of opinion in those cases, where a statement is not simply an expression of personal opinion but rather incites hate or violates the Prohibition Statute (sec 283 Criminal Code). 35% of reported cases between 09/2019 and 08/2020 violated Austrian criminal law in this way or were instances of deliberate insult (a criminal offence). In these cases, we inform our clients about the possibility to file charges and assist them with the necessary steps.

ZARA is also actively campaigning for amendments to relevant laws and provides position papers on draft legislation. In September 2020, the Austrian government released several draft laws aiming at combating online hate. The draft laws fulfil some of ZARA's long-term recommendations targeting decision makers and make access to justice easier for people affected by online hate. The accompanying legal material repeatedly refers to the number of cases documented, and the work done by our team, including the legal actions taken to counter online hate.

The draft legislation proposes to include strong racial insults against individuals as an offense in the criminal code on incitement of hatred and violence. Two new procedures are foreseen, which enable individuals to identify people proliferating online hate against them. Furthermore, a new procedure is proposed, which allows people affected by online hate to ask a court to issue cease and desist orders against perpetrators who harm someone else's dignity. And lastly, the new provisions extend legal and psychosocial process support from the person directly affected to a wider group of people affected by criminal offences. ZARA submitted detailed statements pointing out positive aspects of the draft law and highlighting sections that need revision and/or clarification to allow us and others to better work with the law and support people affected by online hate.

Even if all the draft legislation is put into force, some cases can still only be pursued under civil law. The civil procedure is associated with tremendous personal and financial risk – only a few people can afford such lawsuits. Therefore, together with Sigrid Maurer, a politician who herself has been a target of online hate speech and sought support from us, we set up the legal aid fund #AgainstOnlineHate. Since October 2018, this fund can be used to finance civil lawsuits for cases of online hate. The legal aid fund allows those affected by online hate to hold perpetrators accountable for their actions.

Our team #AgainstOnlineHate will help review whether a personal characteristic was attacked. We then advise, support, and decide whether the legal aid fund can be tapped into for a case and arrange for a local attorney. We make use of the legal aid fund, if it is in the interest of the person affected to launch a lawsuit and it has a realistic chance of success.

Apart from all these activities, we also publish an annual report documenting the number of cases of online hate reported to ZARA. Thus, we shed light on the current situation in Austria. We spread it widely and engage the press, which raises awareness of the issue and the support we provide. Throughout the year, our outreach and social media team creates awareness raising and information campaigns to make our work visible and accessible even more. We run co-operations and fundraising campaigns in collaboration with bloggers, movie theatres, public figures, (social media) platforms and other NGOs.

We are very much aware that it takes much joint effort to change the situation of those affected by online hate for the better. We work in close contact with other CSOs, but also government bodies, transnational networks and institutions to highlight improvements needed and discuss possible solutions.

---

## A SELECTION OF OUR INITIATIVES

### ZARA Trainings

We have developed impactful trainings<sup>3</sup> on antidiscrimination, online hate and the promotion of (digital) civil courage which we provide for schools, companies and other institutions. We have been conducting trainings on civil courage, antidiscrimination and media literacy for more than 17 years and in 2014 founded a sister non-profit company, ZARA Training, focussing on training activities. Annually, ZARA Training conducts approximately 100 workshops on antidiscrimination, online hate and the promotion of (digital) civil courage - engaging with more than 1,500 participants. Our trainings are based on principles and practices of non-formal learning. ZARA Training works with a pool of currently 23 highly qualified trainers who are experts in the field of anti-discrimination and the promotion of civil courage. The training concepts have been tested for years and are based on practical experience from the anti-racism work in Austria as well as scientific theories and internationally recognized training methods.

Civil courage and solidarity make a big difference online (as well as offline), both for the directly affected and those witnessing online hate by reading along. During our trainings we empower users to stand up against online hate. Jointly, we develop counter strategies to prevent spreaders of online hate from silencing others.

### #calmdowninternet

In 2019, we launched the counter-narrative campaign #calmdowninternet. The award-winning<sup>4</sup> campaign was a joint project with TUNNEL23<sup>5</sup>, an advertising company specialised on digital campaigns. We combined AI, sentiment analysis, and antidiscrimination work in this undertaking. First of all, a crawler supported the identification of hateful content on Twitter. The content identified underwent a sentiment analysis, and when the tone of the content was considered to be hateful, automatic responses, including a text to calm down and watch an ASMR - Autonomous Sensory Meridian Response - video, were sent to the authors of the content. Sometimes, however, the ASMR videos were sent in response to tweets which were critical of hateful content and were only quoting these tweets. This was due to the imprecision of the algorithm identifying hateful content.

The intervention through these automatic responses led to mixed reactions. Some users, however, were indeed moved to reformulate or delete their tweets themselves. Others were motivated to get in touch with the ZARA Twitter account.

This campaign allowed us to try out a new creative approach in the fight against online hate. In contrast to deletion of content by external players, this strategy of raising awareness and shifting user behaviour is a more promising and sustainable approach in the fight against online hate speech. The potential we discovered has informed our plans for future campaigns, including the counter-bot project described further below.

### Schneller Konter ("Fast counter-speech")

Many users witnessing and receiving online hate report that fact-based and differentiated counter-speech is often ineffective. Also, it sometimes incites further hate-filled replies. These conventional counter-speech methods often seem slow, boring and 'lame'. All of these are attributes which lead to limited reach and effect. In September 2020, we therefore launched an online tool which enables users to counter online hate with easily and quickly compilable reactions. Visiting the website [www.schnellerkonter.at](http://www.schnellerkonter.at) users can select pictures, videos and pre-written responses from a database and create their own targeted memes or GIFs. Our tool provides a possibility to counter online hate in a fast and humorous way and, most importantly, without further putting oneself at risk. This increases users' scope of action and allows them to keep and strengthen their own voice online. Moreover, it moves users to stand up with courage and support others against online hate.

## Counter-Bot

At the beginning of 2020, we have designed a research project based on our experience with the dynamics and difficulties around online counter-speech. With this project we seek to contribute to the development of a system of artificial intelligence (AI) which identifies racist postings on social media and generates suitable counter-speech. The aim is to provide the scientific underpinnings for the eventual implementation of a bot – a counter-bot – which automatically detects online hate speech and generates counter-speech.

Real moderators lead to high personnel costs and cannot keep up with the speed of content produced on social media platforms. Also, moderators constantly confronted with online hate bear a high psychological brunt.

TUNNEL23 provides the AI, which identifies racist online content. A team of psychologists, linguists and human rights experts analyses to what extent the AI was capable of correctly identifying online content as racist. The content is first assessed linguistically and then statistically along linguistic variables. The results shall feed into a follow-up project, which should finally result in programming a counter-bot, which does not turn into a racist tool. Apart from this direct innovation, our aim is to contribute a human rights view and promote anti-discrimination standards within the AI discourse.

## Our primary practitioners

Our team #AgainstonlineHate is always there to support people directly targeted by or witnessing online hate. We have six legal and psychosocial advisors and would like to highlight three of them. Also, we regularly welcome volunteers who support us in our advisory work. Here is a brief overview of our primary practitioners, Dilber, Dunia and Lukas.

Dilber Dikme is the head of our advisory team. It is of utmost importance to her to ensure that people who seek us out are provided with the right tools so that they can effectively protect themselves against online hate speech. She supports clients who have been excluded, degraded or attacked in reclaiming a sense of safety in all areas of life.

Dilber studied law at the University of Vienna and the Sciences Po in Paris with a strong focus on human rights. Previously, she worked at the Vienna Intervention Centre against Domestic Violence as a legal and psychosocial advisor, leading the department for civil law.

After working as a trainee lawyer, she joined ZARA in March 2019 as a legal advisor and is now head of our legal advisory team. In addition, she is a member of the pool of trainers.

Dunia Khalil is a legal advisor at ZARA since June 2017 and our trusted flagger representative. Her priorities are to let people affected by any form of discrimination or online hate know that they are never the problem and to take away their feeling of loneliness.

Dunia is studying law at the University of Vienna. As our trusted flagger representative, she is in regular contact with the social media platforms Facebook, Instagram, YouTube and Twitter. She is an expert on monitoring, a ZARA trainer and involved in the “Working group on women and gender realities in the Civic Solidarity Platform of OSCE”.

Lukas Gottschamel is the head of the legal department and puts particular emphasis on empowerment through his advisory work. He supports those affected by online hate speech in their decision-making process in difficult situations by pointing out legal and non-legal possibilities for action. He studied law at the University of Vienna and is also a certified mediator. Before joining ZARA in August 2017, he held a university position and worked for the parliament’s general administration office. In his work with clients he weaves together his strong mediation skills and his detailed legal knowledge and is also a ZARA trainer.

---

## IMPACT OF ZARA'S WORK

The impact we aim at with our work is to promote an online space which is critical of hate speech, supports civil courage and guarantees freedom of opinion. At the same time, we fight for a change in the legal realm to make it easier for those affected by online hate to effectively hold perpetrators accountable without prohibitive costs.

Along the way, there are many steps we take to achieve this impact: through the availability of our services more and more people affected can get legal and psychosocial support to ease the emotional burden of being confronted with online hate. Also, our activities and reports promote online courage among users to stand up against online hate and increase the number of joint actions taken against it. We also fight for AI anti-discrimination standards by conducting campaigns and research in this crucial area. Further, we manage to increase the number of deletions of hateful content by either flagging online hate speech to platforms or by sharing instructions on how to proceed to get content deleted. Our work also contributes to legal sanctioning of online hate, which can be measured by the number of cases we report to the police, the prosecutor's office, or administrative authorities, of civil proceedings our clients launch and the resulting deletions of content and closing of accounts. As stated above, most recently, our work has led to draft legislation that would significantly improve the situation of people affected by online hate in Austria. When put into force, many gaps and shortcomings we previously highlighted and put forward in our (media) campaigns would be overcome. Increasing numbers of people affected by online hate turn to ZARA and get support, which is an indicator for increasing trust in our work. Overall, we have managed to increase awareness and knowledge of intervention possibilities regarding online hate speech.

---

## CITATIONS

# ZARA

---

- 1 <https://zara.or.at>
- 2 Our #AgainstOnlineHate work is funded by the Austrian federal government
- 3 <https://assets.zara.or.at/download/pdf/3-GegenHassimNetz-Bericht.pdf>
- 4 <https://zara.or.at/de/training>
- 5 [https://de.wikipedia.org/wiki/Deutscher\\_Preis\\_f%C3%BCr\\_Wirtschaftskommunikation](https://de.wikipedia.org/wiki/Deutscher_Preis_f%C3%BCr_Wirtschaftskommunikation) (search for ZARA)
- 6 <https://www.tunnel23.com/cases/calmdowninternet-ki-reagiert-auf-hass-postings/>

---

CASE STUDIES

# 10 Using AI and Advocacy-driven Counternarratives to Mitigate Online Hate

---

EXTRACTS FROM:

Textgain and Media Diversity Institute

---

[Media Diversity Institute](#) (MDI) works internationally to promote media literacy, combat disinformation and facilitate responsible coverage of diversity issues. In this case study, MDI's project [Get The Trolls Out](#) (GTTO) will be used as an example of counternarratives that challenge online hate. GTTO uses (social) media monitoring, social media campaigning, complaints, video production and memes to mitigate discrimination and intolerance based on religious grounds .

[Textgain](#) is a language technology spin-off company from the University of Antwerp that develops AI for addressing societal challenges, such as online hate speech, radicalization and extremism. In this case study, Textgain's project [Detect Then Act](#) (DTCT) will be used to discuss benefits and challenges of working with such tools in the wild.

---

## DETECT THEN ACT: AI-ASSISTED ACTIVISM AGAINST ONLINE HATE

In recent years, Machine Learning (ML) and more specifically Natural Language Processing (NLP) techniques have advanced to a point where they rival humans in tasks such as predicting state-of-mind, age or gender from (anonymous) text. Since a common feature of online hate speech is to use pejorative language (e.g., clown, thug, scum), it should in theory be possible to isolate it with automatic detection techniques. Theoretically, we could train a tireless AI to detect pejorative language on social media platforms, remove those messages and be done with it. However, the problem is more complex in practice. No doubt, AI is an integral part of the solution in managing hundreds of thousands of new messages per day, but careful consideration should be given to human rights and freedom of expression, as was also recognized by Facebook CEO Mar Zuckerberg.<sup>1</sup>

First, removing offensive messages does not remove the underlying drivers. If anything, those users that see their content blocked will likely only become more disgruntled. While most stakeholders might agree that nobody really minds if violent extremists are disgruntled when their inflammatory propaganda is removed, not all extremist content is violent, and not all offensive content is extremist. There is a large grey area of content in a metaphorical minefield of local government regulations, societal norms and tech company policy. To illustrate this, discriminatory online hate speech is illegal in many EU regions (cf. Germany's NetzDG) while freedom of expression is protected in the US by the First Amendment and *Brandenburg v. Ohio*, and tech companies have to navigate multiple regions worldwide.

Second, automatic techniques sometimes make mistakes, or even worse, perpetuate human prejudices, with the risk of overblocking (removing inconspicuous content) and underblocking (ignoring undesirable content). Contrary to human intervention, today's ML algorithms were not designed to account for their mistakes. This challenge is also highlighted in a recent publication in *Nature* (Rudin, 2019),<sup>2</sup> which advocates for simpler, more interpretable techniques for high-stakes decision making. If anything, one can argue that AI with societal impact should always have human supervisors in the loop. The aim of technology in a moderation setup is then not to replace humans, but to support them in their decision making by taking over the most repetitive tasks.

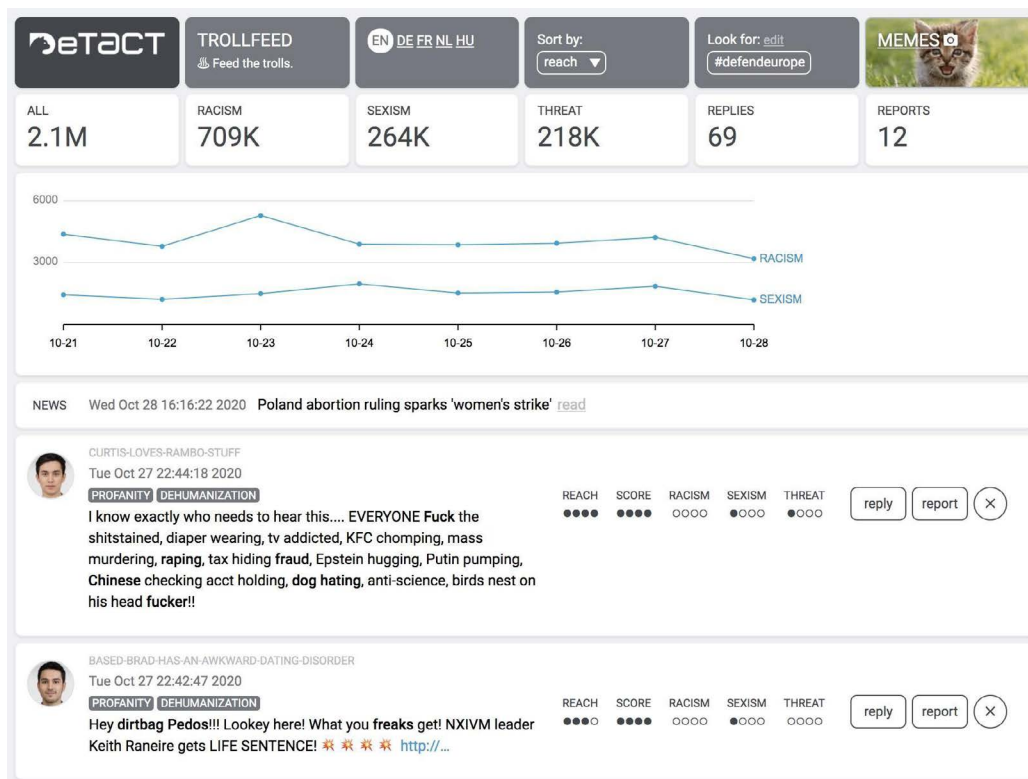


Social media platforms came with the vision of sharing knowledge globally, equal access to information, with equal voices. To uphold that vision is a shared responsibility of all members of society. Perhaps because we are still in our infancy as global citizens of a virtual community, some of our discussions still look like playground bullying, yet open debate is still democracy's best immune system. This kind of self-regulatory aspiration underpins our Detect Then Act project.

Detect Then Act is a collaboration between Textgain, the Media Diversity Institute, the University of Hildesheim (computer science, political science), the University of Antwerp (communication science) and the St Lucas School of Arts in Antwerp. The project is supported by the European Commission's Rights, Equality and Citizenship programme under the call REC-RRAC-ONLINE-AG-2018 (850511). Our aim is to counter online hate speech by encouraging bystanders to become upstanders. In effect, while online trolls test the boundaries and circumvent platform terms of service by framing their us-vs-them

narratives as 'funny memes,' the middle ground (Brandsma, 2017) largely stays silent in dismay.<sup>3</sup> The project encourages volunteers to stand up to online hate and bullying by training them in digital resilience, relevant regulations, by providing AI-powered dashboards and ready-made counternarratives, and by protecting their privacy when reacting.

The project is compliant with the EU's General Data Protection Regulation (GDPR). A private dashboard presents upstanders with a snapshot of today's and yesterday's most hateful messages on social media. These messages are collected by tapping into the platforms' official APIs, yet no content is stored in a database. After two days, any messages that the AI might have spotted are forever forgotten again. Also, the identity of the authors of such messages is never revealed, and when upstanders decide to react, neither is their identity. In academia this is also called a double-blind approach. Messages that attract a lot of buzz are shown with a computer-generated photo and pseudonym, to make them stand out.



Screenshot of the AI-powered dashboard for project upstanders.

---

## EXPLAINABLE AI FOR HIGH-STAKES DECISIONS

The AI that selects candidate messages for inspection is based on an Explainable AI principle, using approaches such as ontologies and decision trees instead of more complex Deep Neural Networks. By 2022 we aim to build an ontology for every European language, that is, a list of thousands of offensive expressions (e.g., stupid clown, stupid girl, stupid nazi), each with a score from 0 to 100 and several possible labels (e.g. is it an insult of intelligence, gender, sexuality, race or belief, is it aggressive, is it a conspiracy theory?). These lists can then be updated automatically by using machine learning to extract Word Embeddings and stay on top of the constantly evolving language use. This is explained further in a short technical report (Voué, De Smedt & De Pauw, 2020).<sup>4</sup>

An advantage is that our system can more easily account for its decision making, by highlighting known keywords. A drawback is that the construction of ontologies relies heavily on input from human experts, but the investment is well worth it: an independent review shows that the approach rivals recent Deep Learning systems (ours: 80.1% precision, BERT: 80.3% precision). The main disadvantage is low recall (ours: 64.8% recall, BERT: 74.9% recall), which means that our system will miss 3-4 out of 10 messages that might be relevant, but this will improve over time as the lists become more expansive.

---

## SIDE STORY: BUILDING AN ONTOLOGY FOR ANTISEMITISM

---

To offer a quick insight, here's how we built an ontology for identifying antisemitic messages. First, we compiled a list of all combinations of damn, dumb, dirty, ... + Jews, and all combinations of Jewish + scum, vermin, and so on. Hundreds of other offensive combinations are possible. We also added Wikipedia's Glossary of Nazi Germany and the like. Then, we searched for social media messages containing such expressions, discovering new ones such as lolocaust and holocough, which we also added to the list. In an online spreadsheet, annotators can then assign a toxicity score to each expression. This allows us to calculate a total score for any given message. By doing this for content that we know is unlikely to be offensive (e.g., Associated Press articles, Wikipedia articles), we can define a threshold score that represents neutral content. Any message with a score above this threshold is worth examining.

To illustrate this, in 1 million messages from 8chan/pol, we find that 20% of the content could be considered extremely offensive to Jewish people. About 10% is also threatening or violent to some degree, and 3% seems to propagate Jewish conspiracy theories (New World Order, Protocols of the Elders of Zion, etc.). The most frequently used slur is kike (65,000x).

## RESPONDING WITH WHOLESOME COUNTERNARRATIVES

Upstanders can react to an identified message in two ways: they can respond to it, or report it. To help decide, the project offers a set of training manuals with best practices, examples, and easy-to-use decision charts of local hate speech legislation. When an upstander decides to report a message, the project manager will pass it on the social media platform, and/or law enforcement in case of incitement to violence. When an upstander decides to respond to a message, they can write a response by themselves or rely on a non-offensive text generator (e.g., “While everyone has their own opinion, can we at least be civil?”) and a set of non-offensive memes – such as cute cat pics with a punchy slogan.

Such responses are also called counternarratives. The aim of these counternarratives is not to address the trolls and end up in a toxic discussion, but rather to try and defuse the situation, and demonstrate to the silent middle ground that hate is not necessarily the norm on social media. One of the reasons for trolls to resort to toxicity is because they may feel detached from society. In the field of Criminology this is called Social Control Theory, which states that an effective deterrent against undesirable behavior is to have strong involvement in society (family, friends, school, work, community, church, etc.).

In the Detect Then Act project we want to respond to hate with wholesome replies, balancing between involvement and sending a message. Getting that right is a challenging trial-and-error process, involving irony, creativity, tact, morality, and there is not a lot of prior academic evidence to base our work on. It is also not easy to find upstanders: many potential candidates have an understandable reserve of being in the center of the storm. In the next phase, we aim to quantitatively analyze the impact of upstanders’ responses, and in the meantime we can rely on MDI’s



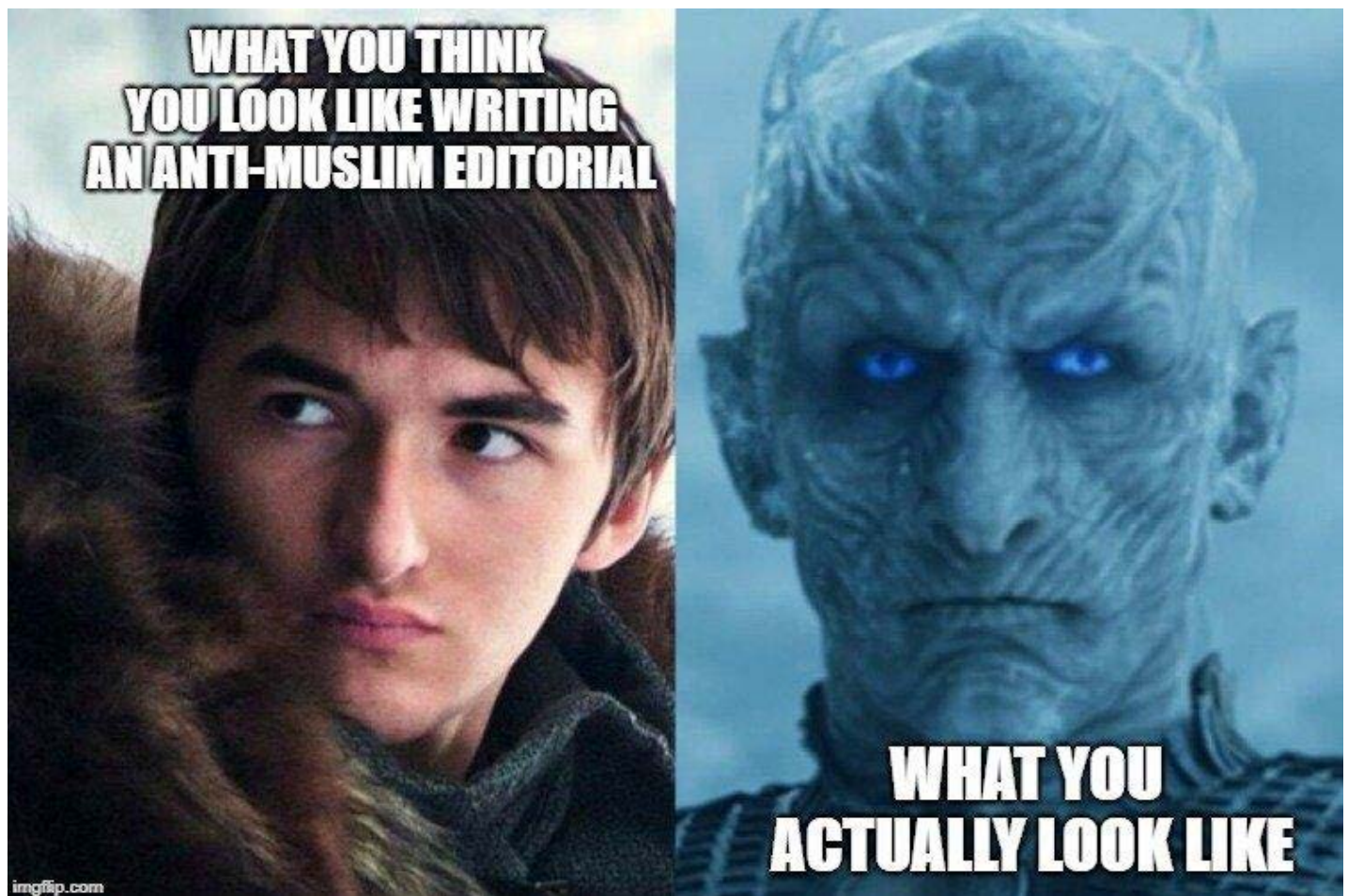
*Soothing cat meme that can be used to react to identified messages.*

first-line experience with the GTTO project.

---

### LESSONS LEARNED FROM GTTO: CREATING SUCCESSFUL COUNTERNARRATIVES

Since 2015, the aim of GTTO has been to leverage social media to engage in dialogue around diverse forms of hate, including antisemitism, Islamophobia, anti-Christian sentiment, and related attempts to turn public opinion against migrants and asylum seekers. GTTO's main audience is young people. Hence, the accompanying counternarratives have been specifically tailored to this group, which highlights an important aspect. The effectiveness of counternarratives depends heavily on demographics: who are we targeting, and what is the best way to target them?



Screenshot example meme from the [Game of Trolls campaign](#).

For example, millennials (25-35 year olds) are an ideal audience for educational narratives. They engage well with all types of informative media like explainer videos, podcasts, infographics, fact memes, and so on. This group constitutes the majority of GTTO followers and sharers. On the other hand, it has been more difficult to engage with younger audiences, 16-25 year olds in particular, due to a variety of factors. First, the choice of platform is key. Today, youngsters are more active on Instagram and TikTok than on Facebook and Twitter, where we communicate with video clips and cartoons. But the choice of content creators matters. Within GTTO, content is designed by millennials. This is something that we have become mindful of, and uptake will likely increase if we work directly with younger content creators.

One engagement technique used within the GTTO project is empowering young people to be producers instead of consumers. Two years ago, coinciding with the season finale of Game of Thrones, we launched a campaign called Game Of Trolls,<sup>5</sup> to help young people tackle online hate in an instructive, actionable way. Deciding to fight fire with fire, we recruited trolls to join our ranks and train them in ‘positive trolling’ through a series of hands-on tips on Facebook and Instagram. Then, our so-called Army of Good Trolls respond to calls for help submitted via the hashtag #TrollWithLove.<sup>6</sup> The campaign reached close to 1 million people on Facebook with the help of Facebook Ad Credits, which were critical to the success of the campaign. This shows how synergy between CSOs and social media platforms can result in powerful, broadly visible counternarratives. However, it also highlights one of the challenges faced by CSOs in mitigating online hate: without the support of the platform we could not have afforded the campaign. Creating effective counternarratives is not only about good ideas and demographics, but also about (financial) resources and tools to put them into practice.

There are many other reasons why counternarratives might succeed or fail. In GTTO, we constantly review and adapt our strategy, not only because we want to improve as first-line practitioners, but also because the effectiveness of techniques changes over time. As target audiences widen and expand their interests, so too must the content that they want to engage with and how. In the overview (right) are some practice-based insights that we have learned throughout this process:

Excerpts from the *Fantastic Trolls and How to Fight Them* guide. (below)



**Casual:** To make counternarratives appealing it is often important to strike the right tone. We avoid sounding like an NGO, which may represent part of an establishment that teens push against, and create distance. We try not to nag or preach and instead look to carry people with us.

**Fresh:** To keep people engaged, we work to keep our content fresh, by appropriating new trends and pop culture. For example, we used the movie release of *Fantastic Beasts and Where to Find Them* to launch a *Fantastic Trolls and How to Fight Them* guide,<sup>7</sup> a ‘bestiary’ with different types of online trolls, and how people can or shouldn’t engage with them.

**Stimulating:** It is one thing to share facts and figures, but to create effective engagement there needs to be a clear call to action for those who are consuming the content. How can they help, and why should they?

**Multimodal:** It is vital to use different forms of media throughout a campaign. Within GTTO, we use a mix of videos, images, infographics, memes, cartoons and articles, keeping in mind that walls of text are going to be scrolled faster than visual content.

**Persistent:** We constantly adapt with new tones, new trends, new modes, new content. In a way, campaigning is a case of attrition: the content needs to be kept flowing, and memes that don’t work once may work in the future if the global landscape changes to make it more relevant.

**Pragmatic:** While it is splendid to get support from sponsors and social media platforms, to keep day-to-day work going, it can be useful to rely on free apps for content generation. There is no formula or price setting for creating a viral meme, cheap & cheerful can also work.

**Practical:** In GTTO we continuously learn from others, for example Vox in the case of explainer videos. There is no need to reinvent the wheel. When time and resources are scarce, best practices from other initiatives can often be boiled down to basic yet effective output.

---

## YOUNG PEOPLE COMMUNICATE MULTIMODALLY

Some of our most successful content to date are videos that reappropriate pop culture references such as CinemaSins,<sup>8</sup> which we turned into JournalismSins to provide debunks in a visually pleasing and engaging way. Cartoons and carousels on Instagram have also been very successful, allowing youngsters to casually swipe through educational BLM content. It sometimes helps to lure users to our site, with a link accompanying short intro videos and troll graphics, even if it means the overall bounce rate is high.

Young people are now growing up with multimodal social media, and they are perfectly accustomed to communicating with short, fleeting video and audio clips instead of writing texts – dynamics not always fully understood by previous generations. In the overview below are some practice-based insights that we have learned in relation to video content:

**Tailored:** Understanding the target audience and using the right platform is key. Educational videos are not going to be watched on Instagram, while cartoons are not going to gain traction on Twitter.

**Concise:** The first two seconds (or the thumbnail) should be the most engaging or visually interesting, and the video should not be overloaded with text.

**Basic:** Videos should highlight a few core talking points, encouraging more in-depth reading on the accompanying website. Videos should be short and digestible, driving traffic to the campaign’s hub.

**Bold:** Videos should immediately debunk fake stories and conspiracies (ideally in the first few seconds), and not build up to an academic body of evidence. Conclusions come first, explanations later.

---

## WHAT ALWAYS WILL REMAIN CHALLENGING, AND WHY

Initiatives like Detect Then Act and Get The Trolls Out continually adapt to evolving audiences, behavior and technology. The greatest hurdle is the ever-changing tide of hate. Each day new hateful memes, new hashtags, images and videos emerge from the web's underbelly. Finding - let alone reacting to - everything is no longer possible, and ill-conceived reactions may also exacerbate the problem. Practitioners seeking to counter hate must now be selective in choosing what is having the most impact. AI can help, but it is not without pitfalls. Quantitative and qualitative approaches should work side-by-side as a key to success, but we need to close the gap between developers that might not fully grasp the problem and practitioners that might not fully grasp the technology. Algorithms can evaluate how influential keywords are, and whether they are going to 'explode' at some point in the future, but maybe only human experts should be able to operationalize this data. Academic groundwork can be an advantage too, and tech companies should perhaps be less afraid to adopt open sources, strengthening their accountability towards society.

Finally, mitigating hate is a responsibility shared by all members of society, demanding closer collaboration between law enforcement bodies, civil society actors, users, and tech companies.

---

## CITATIONS

# Media Diversity Institute

---

- <sup>1</sup> Zuckerberg, M. (2019). Standing For Voice and Free Expression. Facebook News. <https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression>
- <sup>2</sup> Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. <https://arxiv.org/pdf/1811.10154>
- <sup>3</sup> Brandsma, B. (2017). Polarisation management model. In Lenos & Keltjens RAN POL and EDU meeting on Polarisation Management, 2017, Stockholm. [https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/networks/radicalisation\\_awareness\\_network/ran-papers/docs/ran\\_edu\\_pol\\_meeting\\_polarisation\\_management\\_stockholm\\_10-11\\_05\\_2017\\_en.pdf](https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/networks/radicalisation_awareness_network/ran-papers/docs/ran_edu_pol_meeting_polarisation_management_stockholm_10-11_05_2017_en.pdf)
- <sup>4</sup> Voué, P., De Smedt, T., & De Pauw, G. (2020). 4chan & 8chan embeddings. arXiv preprint. <https://www.textgain.com/portfolio/4chan-8chan-embeddings-textgain-technical-report-1>
- <sup>5</sup> Game Of Trolls: <https://www.facebook.com/getthetrollsout/videos/1333644333438278>
- <sup>6</sup> #TrollWithLove: <https://www.facebook.com/hashtag/TrollWithLove>
- <sup>7</sup> Fantastic Trolls: <https://getthetrollsout.org/resources/guides>
- <sup>8</sup> <https://en.wikipedia.org/wiki/CinemaSins>

## REFERENCES

- <https://dtct.eu/upstanders>
- <https://getthetrollsout.org/what-you-can-do>

Eline Jeanné (MDI), Lydia El-Khoury (Textgain), Tom De Smedt (Textgain)



