Research report

# COVID-19, Disinformation and Hateful Extremism

## Literature review report

Kate Cox, Theodora Ogden, Victoria Jordan, Pauline Paille

March 2021

Prepared for the Commission for Countering Extremism (CCE)

# Preface

This report was produced by RAND Europe for the Commission for Countering Extremism (CCE) to examine hateful extremism within society during COVID-19. It presents the findings of a literature review that explored the links between hateful extremism and false information, and identifies associated online interventions and policy responses. **This report has been published by an independent body. The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the UK Government or the Commission for Countering Extremism.**

RAND Europe is a not-for-profit policy research organisation that aims to improve policy and decision-making in the public interest through research and analysis. RAND Europe's clients include European governments, institutions, NGOs and other organisations with a need for rigorous and independent interdisciplinary analysis.

For more information about RAND Europe or this study, please contact:

Ruth Harris

Research Group Director – Defence, Security & Infrastructure

Westbrook Centre, Milton Road

Cambridge CB4 1YG

United Kingdom

Tel: +44 (0)1223 353 329 x2624

ruthh@randeurope.org

# Executive Summary

## This study explores hateful extremism within society during COVID-19

The COVID-19 pandemic has provided a breeding ground for conspiracy theories, disinformation and hateful extremism. Pandemics are inherently fast-moving and information is constantly evolving, creating opportunities for hateful extremist groups to spread doubt, fear and suspicion among the public. Forums such as 4Chan and Reddit are hubs for real-time debate, conspiracy theories and disinformation. Similarly, social media platforms such as Facebook, Twitter and YouTube play a role in generating and amplifying false information. During lockdown and with rising unemployment, more people have been spending time at home and online, with greater exposure to false information and hateful extremist narratives.

Particularly in the COVID-19 context, it is important to ensure that today's digital generations are equipped to identify hateful extremism and false narratives in order to build societal resilience. As COVID-19 presents an unprecedented challenge and a catalyst for false information, this rapidly developing area requires research input. There is a need to consolidate existing research, better understand the evidence base and address gaps to inform primary research, policy planning and decision making.

In July 2020, the Commission for Countering Extremism (CCE) commissioned Ipsos MORI and RAND Europe to undertake a study to examine hateful extremism within society during COVID-19. This report presents the findings of a literature review conducted by RAND Europe on the links between hateful extremism and false information,[1] and on associated online interventions and policy responses. The study team addressed six research questions under the two review themes:

**Review theme 1: Links between hateful extremism and false information**

- 1.1: What impact can false information have on hateful extremist beliefs and behaviours?
- 1.2: In what ways do hateful extremist beliefs contribute to the spread of false information?
- 1.3: What trends and variations can be identified across different audience types, modes of false information, and extremist groups?

**Review theme 2: Associated online interventions and policy responses**

- 2.1: What insights can be identified from the literature on the effectiveness of existing interventions and policy responses?
- 2.2: What recommendations are put forward in the existing literature in relation to future interventions in this area?
- 2.3: What transferrable lessons/'good practices' from successful interventions in related policy areas can be identified?

[1] In this report, we use 'false information' as a catch-all term to refer collectively to online misinformation, disinformation and conspiracy theories. An overview of key definitions is provided in Section 1.2, and a more detailed glossary of terms is presented in Annex A.

The research questions were addressed through a Rapid Evidence Assessment (REA), which involved a review of 93 relevant papers across disciplines including psychology, political science, sociology and law.

## False information can shape hateful extremist beliefs and behaviours by leading to the growth of echo chambers and a rise in hate incidents (*Q1.1*)

The literature offers several hypotheses on the links between false information and hateful extremism, but the quality and quantity of the literature is not sufficient to provide strong, empirical evidence of these associations. The literature indicates that false information can lead to the growth of echo chambers, cementing hateful extremist attitudes by desensitising group participants to hateful language and narratives. As part of this process, echo chambers attract a concentration of extremists, without the presence of more moderate users to challenge their perspectives. False information has also been associated with increasing levels of hate crime and blame on minority groups for the current pandemic, although the direction of these relationships is not known. In the COVID-19 context – and with the increasing reach of false information – minority groups including Jewish, Chinese and Muslim communities have been blamed for the spread of the virus by hateful extremists.

## Hateful extremists are incentivised to spread disinformation and conspiracy theories by increased exposure and recruitment benefits (*Q1.2*)

While hate groups are not the only actors to disseminate false information, it is in the interests of hateful extremists to spread disinformation and conspiracy theories. The dissemination of this type of false information may give hateful extremist groups increased exposure, including through mainstream media reporting and public officials' statements. Hateful extremist narratives are also used as a recruitment tool and are tailored to appeal to people struggling from a lack of prospects in the current COVID-19 climate, in an effort to attract new supporters and sympathisers who might be susceptible to the influence of false information. Hateful extremist narratives typically incorporate disinformation by focusing on specific groups of people as 'out-groups', whether by expressing superiority over other groups, criticism of their opponents or victimisation by others. These types of narratives are increasingly shared by hateful extremists via automated social media accounts ('bots'), and dissemination techniques are shifting to more sophisticated online interactions through impersonation and amplification of organic posts. Using these techniques, hateful extremists are incentivised to spread false information by the allure of publicity and new supporters.

## Hateful extremist actors typically direct their narratives against 'out-groups', but these narratives frame the pandemic in different ways (*Q1.3*)

Common trends and variations can be identified in relation to how hateful extremist groups have framed the pandemic in their narratives. Many different types of hateful extremist actor operate online, ranging from far-right extremists to Islamist hate actors. While these various actors have exploited the current pandemic to advance their interests, they have taken different approaches in doing so. Far-right groups often blame migration, globalisation or the government for the virus, while

some Islamist extremist actors instead see the pandemic as a form of divine punishment against unbelievers. However, all groups tend to direct hostile narratives at 'out-groups', leveraging public fear and uncertainty surrounding the global pandemic. As more people – particularly young people – have been consuming online content during lockdown, this is likely to have exposed a greater cross-section of the population to hateful extremist recruitment.

## While empirical evidence on the effectiveness of online interventions is limited, the literature highlights that fact-checking, counterspeech, takedowns and education appear to work (*Q2.1*)

The review identified four categories of online initiatives for countering false information and hateful extremism: fact-checking, counterspeech, takedowns and education. While the reviewed literature does not offer rigorous empirical evaluations of such interventions, papers identify several promising practices promoted by civil society, government, media and social-media-company actors in terms of reducing the spread of false information and building societal resilience. The 'good practices' identified in this report are those highlighted in the reviewed literature as being effective, timely and relevant in curbing the online spread of false information and in building societal resilience to hateful extremism. For example, the UK-based organisation Stop Funding Fake News is working to prevent disinformation sites from earning advertising revenues. Policy interventions have also offered value in terms of holding social media platforms to account for removing false information, as evident in the 2018 German NetzDG Act. Media organisations counter hateful extremist disinformation by avoiding clickbait and maintaining transparency in factchecking, and social media companies contribute by modifying algorithms to avoid 'recommending' harmful content. It is worth noting that this report presents the findings and recommendations of the relevant literature, rather than offering an independent analysis of the effectiveness of counter measures.

## The reviewed literature offers a number of recommendations for the design and delivery of future interventions (*Q2.2*)

COVID-19 presents a unique challenge for policymakers and organisations seeking to tackle false information. Reviewed sources offer recommendations for decision makers tasked with designing and implementing future interventions and policy responses, though it should be noted that these recommendations are typically not based on robust empirical evaluations. For governments, sources highlight a need to dedicate more resources to combat false information in order to build societal resilience, as well as to conduct or commission further research into the impacts of hateful extremist narratives. Social media companies and media organisations are also urged to take more responsibility, respectively by managing the content on their platforms and ensuring that outlets adhere to good journalist practices (e.g. avoiding clickbait headlines).

## This report offers multidisciplinary findings on hateful extremism, but reviewed sources offered limited insight from other policy areas (*Q2.3*)

Successful online interventions in alternative policy areas were not identified in the reviewed papers and are accordingly not explored in this report. While one source referred to gang violence, the reference was made solely to distinguish this threat from extremism (Baldauf et al. 2019). The paper observed a link between criminal biker gangs and right-wing hate groups, but noted that the former lacks an ideology to legitimise their acts of violence, limiting the transferability of insights from this group to right-wing hate groups. Beyond this source, there was limited reference to other policy areas in the reviewed papers. It should nonetheless be noted that the interventions and policy responses presented in this report are identified from papers across a wide range of disciplines – including psychology, political science, sociology and law – offering a diverse range of perspectives and approaches.

## This report presents a set of policy considerations for CCE

The report sets out policy considerations for CCE based on the literature insights identified:

- **Investing in research could help address evidence gaps and strengthen responses to false information and hateful extremist narratives**. As outlined below, there is a need for further evaluations of existing interventions, as well as research on directional motivations and studies with a wider reach in terms of geography, languages and online content. Developing a more comprehensive understanding of the issue of false information and its use in hateful extremist messaging should help inform effective responses.
- **Holding tech companies to account could increase their responsiveness to false information and hateful extremism**. As the implementation of the 2018 NetzDG Act exemplifies, levying large fines on tech companies that do not remove false information and hateful extremist content in a timely way can increase companies' responsiveness in removing this content from their platforms.
- **Investing in education could help raise awareness of the dangers of false information and hateful extremism**. Particularly in light of review findings that younger people are more likely to encounter false information and to accept its presence online, investing in education and training is important in increasing public awareness of false information. Given increased exposure of online users to this content during COVID-19, there is a pressing need to educate the public about the threat of false information and its use by hateful extremists, and about public actions to support individual resilience.
- **Collecting and publishing information regarding indicators of hateful extremism could help improve policy responses**. The literature highlights a need for governments to collate and publish information on hateful extremism. There is a need to broaden the type of information collected (e.g. looking beyond text-based content to include images, audio, memes and other content), make greater use of computational advances (e.g. machine learning), and to ensure the quality of statistics (e.g. via independent peer review). A better understanding of the nature and scale of the threat could help enhance policy responses and improve public resilience.

- **Exploring the use of 'good' bots could support the spread of positive narratives online**. It is evident from online activity during major political events (e.g. the 2016 US presidential election) that the use of trolls and bots has manipulated voter behaviour and deepened societal divides. Noting the persuasiveness of these tactics, there could be scope to explore the adaptation of such techniques to instead promote democratic values of tolerance, acceptance and diversity on social media, as well as to constrain the reach and influence of online hate speech.
- **Collaborating across sectors could ensure that interventions are mutually reinforcing**. Coordination across UK policy officials, social media moderators, educators, journalists, civil society organisers, research experts, legislators and other national governments could help ensure that HMG policy and guidance reflects an understanding of the scale and nature of the challenge from hateful extremists' use of false information, and complements activities that are being delivered elsewhere.

## The study also identifies essential avenues for further research

Based on the evidence gaps identified, the study highlights areas that would benefit from further analysis and exploration:

- **Independent and robustly designed evaluations of existing interventions**. To inform a better understanding of the effectiveness of existing counter-measures, there would be merit in conducting independent evaluations of interventions dedicated to tackling hateful extremism and false information. Evaluations could focus on measuring the effectiveness of factchecking, counterspeech, takedowns or educational initiatives for countering false information and its use by hateful extremist actors. The results of these evaluations should be made publicly available to help inform effective future interventions.
- **Research on 'directional motivations' (an individual's propensity to hold onto existing attitudes)**. By focusing on 'directional motivations', future studies could improve understanding of the characteristics of individuals who are more prone to hateful extremist beliefs and behaviours. This could also help enhance awareness of the way these individuals respond to false information. This research could shape new interventions, informing an understanding of conditions under which particular responses to false information will be more (or less) effective.
- **Studies with broader coverage in terms of geography, languages and online content**. To develop a fuller evidence base for policy and decision makers, future studies should analyse a wider range of: (i) countries and regions, noting the Euro- and US-centric focus of the current literature;[2] (ii) languages, to understand non-English-language false information; and (iii) types of online content, moving beyond text-based content to images, audio files, memes, GIFs or videos. Researchers could rely more on automated processes and machine learning approaches to analyse large volumes of online content, allowing for larger and more varied datasets.

[2] It should be noted that this report has been written to inform CE policy in England and Wales. Our review of the literature nonetheless revealed that there is a broader shortage of empirical studies with a focus on geographical areas beyond Europe and the US, which could be of interest for future work.

# Table of contents

# Figures, Tables & Boxes

# Abbreviations

| | |
|---|---|
| BNP | British National Party |
| CCDH | Center for Countering Digital Hate |
| CCE | Commission for Countering Extremism |
| CoE | Council of Europe |
| CST | Community Security Trust |
| CVE | Countering Violent Extremism |
| EU | European Union |
| ICCT | International Centre for Counter-Terrorism |
| ISIS | Islamic State of Iraq and Syria |
| MENA | Middle East and North Africa |
| NHS | National Health Service |
| PVE | Preventing Violent Extremism |
| REA | Rapid Evidence Assessment |
| STREET | Strategy to Reach, Empower, and Educate Teenagers |
| UK | United Kingdom |
| UN | United Nations |
| UNESCO | The United Nations Educational, Scientific and Cultural Organisation |
| US | United States |

# 1. Introduction

In July 2020, the Commission for Countering Extremism (CCE) commissioned Ipsos MORI and RAND Europe to undertake a study to understand and respond to hateful extremism within society during the COVID-19 pandemic. This document presents the findings of a literature review conducted by RAND Europe on the links between extremism and false information, and on associated online interventions and policy responses.[3]

## 1.1. Background

The current COVID-19 crisis has seen the spread of conspiracy theories, disinformation and hateful extremist narratives.[4] Recent research indicates that a significant proportion of the UK population believes in conspiracy theories relating to COVID-19.[5] According to one survey, at the beginning of the pandemic some 3 in 10 respondents believed that COVID-19 was created in a lab, while 1 in 8 saw the pandemic as part of a global effort to force the public to be vaccinated.[6] Certain conspiracies surrounding COVID-19 are more sinister, deriving from false information around ethnic and religious minorities, particularly the Chinese, Jewish and Muslim communities. Harmful narratives can lead to hate crime and discriminatory behaviour, with such false narratives including disinformation that links COVID-19 to a fabricated Jewish plot to initiate a new world war,[7] blames Chinese communities for spreading the virus,[8] and accuses Muslims of flouting social distancing rules.[9]

During lockdown, more people are spending time online and on social media – a domain in which conspiracy theories, misinformation and disinformation thrive and multiply.[10] While false information

[3] The findings presented in this report derive from the reviewed literature analysed by RAND Europe researchers and do not necessarily reflect the views of the CCE.

[4] 'Hateful extremism' refers to behaviours that 'incite and amplify hate, or engage in persistent hatred, or equivocate about and make the moral case for violence', drawing on hateful, hostile or supremacist beliefs directed at an out-group, 'who are perceived as a threat to the wellbeing, survival or success of an in-group'; which can cause harm to individuals, communities or wider society as a whole. CCE (2019) (see Annex A).

[5] King's College London (KCL) & Ipsos MORI (2020).

[6] KCL & Ipsos MORI (2020).

[7] CST (2020).

[8] Pei and Mehta (2020).

[9] Ariza (2020).

[10] Ziems et al. (2020); EC (2018); Jones (2020); Polyakova & Fried (2019); Ayad (2020); 'Misinformation' refers to the distribution of incorrect information without the intention to mislead, while 'disinformation' refers to the deliberate dissemination of misleading information, Brennen et al. (2020) (see Annex A).

can be shared offline through family, social networks, books or newspapers, the online domain presents a considerable challenge. A lack of traditional media gatekeepers has made online platforms fertile ground for false information, where any internet user can broadcast their opinions, memes, doctored images or conspiracy films. As of June 2020, over 1 billion students worldwide were no longer at school and spent more time online,[11] providing hateful extremists with the opportunity to engage audiences with false-information-based propaganda on an unprecedented scale.[12]

Within this wider context, the UK faces particular challenges. A high proportion of far-right extremists with an extensive social media reach are UK citizens,[13] and there are gaps in the UK's response to hateful extremism. Unlike terrorism[14] – for which there is a more developed response under CONTEST[15] – hateful extremism is defined by its focus on inciting and amplifying hate, and making the moral case for violence.[16] Furthermore, hateful extremist narratives direct hostile or supremacist beliefs at other groups, with the potential to cause harm to individuals, communities and wider society. Hateful extremism falls under the 'umbrella' of extremism, but is distinct from violent extremism – which includes the use of terrorist tactics and/or violence.[17] While there are generally structured responses to terrorism and violent extremism, there is not such a developed framework to deal with hateful extremism.

The COVID-19 pandemic has provided a breeding ground for hateful extremism. Pandemics are inherently fast-moving contexts in which information – even from credible expert sources – is constantly evolving.[18] This can create opportunities for extremist groups to sow seeds of doubt and suspicion among the public. By leveraging heightened public fear and hijacking COVID-19 content, extremists can spread hateful views, particularly on race.[19] Forums such as 4Chan and Reddit have become hubs for real-time debate, conspiracy theories and mis/disinformation;[20] while social media platforms – such as Facebook, Twitter and YouTube – similarly play a role in hosting and amplifying false information.[21] An empirical study measuring the diffusion of falsehoods and truth on Twitter over a decade suggests that lies spread faster than truth online, due in part to such false information being novel, or invoking fear or disgust.[22] Furthermore, despite increasing moderation by social media companies, there have been sustained increases in traffic to far-right websites, and in the number of followers of far-right social media accounts.[23]

---

[11] UN (2020).
[12] UN (2020).
[13] Lowles and Levene (2019).
[14] 'Terrorism' can refer to the use or threat of action to intimidate the public and further political, religious, racial or ideological goals, CPS (2019) – though it should be noted that a range of differing definitions exist.
[15] HM Government (2018).
[16] CCE (2020).
[17] CCE (2020).
[18] Colliver and King (2020).
[19] Colliver and King (2020).
[20] Marwick and Lewis (2017).
[21] ISD (2020).
[22] See, for example, Vosoughi et al. (2018).
[23] Lowles and Levene (2019).

The ongoing pandemic has been conducive to the spread of hateful narratives, and lockdown conditions have seen an increase in online searches for extremist content.[24] In the US, for example, Moonshot CVE found a 21 per cent average increase in engagement with violent extremist content in states with lockdown measures in place for 10 or more days.[25] Similarly, in Canada, there has been a significant increase in extremist-related search traffic during COVID-19, with an average increase of 18.5 per cent across the country.[26] There has not yet been a comparable study published in the UK, though such a publication could provide important insights. In the UK, referrals to Prevent saw a 50 per cent decrease between 23 March 2020 (early lockdown) and 22 April 2020, because the usual in-person school or local authority referral channels had been disrupted by COVID-19. This prompted concerns that those who might otherwise need and receive a Prevent intervention were being targeted by extremist Islamist or right-wing radicalisation tactics, making them increasingly vulnerable to hateful extremist influences.[27]

Social media creates conditions for hateful extremist actors to mobilise by broadcasting false information, harassing opponents, and coordinating activity – including protests and publicity stunts.[28] Prominent online trolls, conspiracy theorists and ideologues are significant nodes within networks, holding disproportionate influence among other actors with the ability to amplify narratives and manipulate media.[29] Terrorist actors are similarly leveraging social media, particularly given that lockdown has reduced the effectiveness of common tactics – such as attacks on crowded spaces[30] – and disrupted their global and national supply chains.[31] In a shift away from physical attacks, many such groups have concentrated their efforts online.

As argued by Briant (2018), many of the online methods employed by hateful extremists reflect training and knowledge acquired in the military or intelligence.[32] While it is important to note that the vast majority of those posting hateful content do not have such backgrounds or training, it is possible that some of the methods are borrowed from military or intelligence playbooks. The evolution of sophisticated tactics – including deception techniques, demoralisation tactics and the exploitation of psychological weaknesses – requires governments, social media companies, news/media organisations and civil society to remain equipped to tackle these developments. Particularly in the COVID-19 context, the proliferation of hateful extremist propaganda and conspiracy theories surrounding the pandemic requires an urgent response, and there is a need to ensure that today's digital generations are equipped with the tools to identify hateful extremism and false narratives in order to build societal resilience. As the COVID-19 pandemic presents an unprecedented challenge and a catalyst for false information, this rapidly developing area requires research input. There is a

[24] Avis (2020).
[25] Avis (2020); Moonshot (2020a).
[26] Moonshot (2020b).
[27] Dodd (2020).
[28] Davey et al. (2020). For further analysis of the benefits that disinformation offers hateful extremist actors, see Section 3.1.
[29] Marwick and Lewis (2017).
[30] UN (2020).
[31] UN (2020).
[32] Briant (2018); Jones (2020).

need to consolidate existing research, better understand the evidence base and address evidence gaps to inform primary research, policy planning and decision making.

## 1.2. Purpose and scope

The purpose of this report is to present the findings of the REA to address research questions under each review theme, as presented in Box 1.

**Box 1: Review themes and supporting research questions**

**Review theme 1: Links between hateful extremism and false information**

- 1.1: What impact can false information have on hateful extremist beliefs and behaviours?
- 1.2: In what ways do hateful extremist beliefs contribute to the spread of false information?
- 1.3: What trends and variations can be identified across different audience types, modes of false information, and extremist groups?

**Review theme 2: Associated online interventions and policy responses**

- 2.1: What insights can be identified from the literature on the effectiveness of existing interventions and policy responses?
- 2.2: What recommendations are put forward in the existing literature in relation to future interventions in this area?
- 2.3: What transferrable lessons/'good practices' from successful interventions in related policy areas can be identified?

### Definitions and caveats

For the purposes of this review, we refer to '**hateful extremism**' as behaviours that 'incite and amplify hate, or engage in persistent hatred, or equivocate about and make the moral case for violence', drawing on hateful, hostile or supremacist beliefs directed at an out-group, 'who are perceived as a threat to the wellbeing, survival or success of an in-group'; which can cause harm to individuals, communities or wider society as a whole.[33] In this report, we use '**false information**' as a catch-all term to refer collectively to online misinformation, disinformation and conspiracy theories.[34] We draw a distinction between '**misinformation**' and '**disinformation**', based on the intent of an individual or group to spread that information. Misinformation refers to the distribution of incorrect information without the intention to mislead, whereas disinformation refers to the deliberate dissemination of misleading information.[35] In this report, '**conspiracy theories**' can be understood as narratives created to infer that

[33] CCE (2019).
[34] Kumar and Shah (2018); Brennen et al. (2020).
[35] Ball and Maxmen (2020).

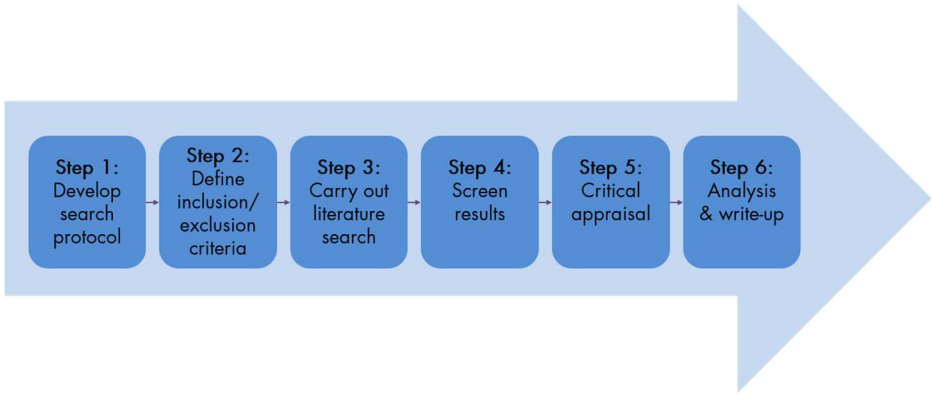
an event or situation is the result of a secret plan made by powerful individuals or groups.[36] An expanded set of definitions for the key terms used in this report is presented in Annex A.

There are several scoping considerations and caveats to note when considering the findings of this report. As outlined in Section 1.3 and Annex B, our literature review was based on a Rapid Evidence Assessment (REA)[37] approach. An REA was considered to be the most appropriate methodology for this study, offering a structured and robust approach within the constraints of the study. However, the review is not intended to be exhaustive or to include all papers on the topic of hateful extremism, false information and associated interventions. A further consideration to note is that when discussing existing and future interventions, this document reports on the findings and recommendations of reviewed sources rather than offering an independent assessment of their effectiveness. Reflecting the focus of the wider study, the review focused on interventions and policy responses with an online and social media dimension.

## 1.3. Research approach

The study team used an REA-based approach to deliver the literature review. The REA approach followed the six steps presented in Figure 1-1 and expanded on below.

**Figure 1-1: REA approach**




- **Step 1 – Develop search protocol**: RAND Europe developed a search protocol with a focus on the two review themes: (i) links between hateful extremism and false information; and (ii) associated online interventions and policy responses. The protocol involved the use of four search strings – combinations of terms to yield search results – covering the review themes from a general standpoint and with a focus on COVID-19. Sources were gathered via academic database searches (Academic Search Complete, Policy File Index, Scopus and Google Scholar), snowballing and targeted searches, as well as via sources provided by CCE and Ipsos MORI.

[36] Bolsen and Druckman (2018); Connolly et al. (2019); Douglas et al. (2019a); Vegetti & Levente (2020); ISD (2020); Fangen and Holter (2020); Holbrook (2020).

[37] A Rapid Evidence Assessment (REA) is a form of literature review that provides an overview of the quantity and quality of evidence in a particular field, but is not as exhaustive as a systematic review.

- **Step 2 – Define inclusion/exclusion criteria**: With a focus on the two review themes, the inclusion criteria were as follows:
    - **Geographic location**: UK (primary); rest of world (secondary);
    - **Source types**: academic, grey (research papers, evaluations, policy documentation), polling data;
    - **Language**: English-language sources;
    - **Publication date**: 2010–2020 (searches 2 and 4); 2020 (searches 1 and 3).
- **Step 3 – Carry out literature search**: An initial database search by RAND's Knowledge Services across Academic Search Complete, Policy File Index, Scopus and Google Scholar yielded 793 sources, which were loaded onto an EndNote database.
- **Step 4 – Screen results**: The 793 sources from the database search were screened against the inclusion criteria and narrowed down to 29 sources, which were added to the sources shared by CCE (29), provided by Ipsos MORI (5), identified through snowballing (14), and identified through targeted searches (16), resulting in a total of 93 sources.
- **Step 5 – Critical appraisal**: A full-text review of the 93 sources was then undertaken. For each source, data was extracted into a spreadsheet, with the content mapped against a set of categories including definitions and scope (e.g. country focus, type of hateful extremism); links between hateful extremism and false information; interventions and policy responses; and the research methods underpinning each reviewed source.
- **Step 6 – Analysis & write-up**: An Internal Synthesis Workshop was held on 11 August 2020, at which researchers from RAND Europe and Ipsos MORI discussed the emerging findings from the literature review and their implications for potential future research. Findings were then written up in a narrative synthesis and integrated into this report.

For a fuller description of the research approach, please refer to Annex B.

## 1.4. Structure of the report

In addition to this introduction, this report contains four substantive chapters:

- **Chapter 2** characterises the evidence base by providing an overview of the reviewed papers according to their geographic focus, underlying research methods and the type of hate/extremism examined.

- **Chapter 3** presents findings in relation to the first review theme, and describes how hateful extremism is linked to false information in the literature.

- **Chapter 4** outlines the findings of the second review theme, and presents a range of associated online interventions and policy responses.

- **Chapter 5** summarises the key findings of the review, identifies policy considerations for CCE and highlights areas for future research.

The report contains three annexes that complement and add further supporting detail to the core report chapters:

- **Annex A** provides expanded definitions of the key terms identified in the literature review, elaborating on those included in this chapter.

- **Annex B** describes the research methods used to undertake this study, expanding on the summary presented in Section 1.3.

- **Annex C** (separate) presents the underlying research methods that form the basis of the reviewed papers, expanding on Section 2.1.4 and offering clarity on the strength of the evidence base.

# 2. Characterising the evidence base

This chapter provides an overview of existing research on the links between hateful extremism and false information, and on associated online interventions and responses. The sections below are intended to provide the reader with a contextual understanding of the evidence base before considering the core findings of the review, which are presented in Chapters 3 and 4.

## 2.1. Overview of reviewed sources

The study team reviewed a total of 93 sources as part of the literature review, drawing on papers across disciplines including psychology, political science, sociology and law to offer a rich range of insights. The paragraphs below describe the geographic focus of the papers (Section 2.1.1); the type of hate or extremism described by sources (Section 2.1.2); the number of sources focusing specifically on COVID-19 (Section 2.1.3); and the distribution of underlying research methods used in the reviewed sources (Section 2.1.4).

### 2.1.1. The majority of papers focus on European countries, of which more than half examine the UK context

Of all the sources reviewed, 70 papers (75 per cent of all 93 papers) state an explicit geographic focus. Among these, the majority focus on the European context, with 26 papers (28 per cent) focusing on European countries or regions. Further to this, 23 papers (25 per cent) have a global focus spanning a wide range of countries. Remaining papers with an explicit geographic focus are concerned with the US (14; 15 per cent), Canada (2; 2 per cent), Australia (2; 2 per cent), the MENA[38] region (2; 2 per cent) and the Maldives (1; 1 per cent).

Of the papers focusing on the European context, a quarter cover the European Union (EU) or Europe as a whole; of these, four (4 per cent) focus on the EU only and three (3 per cent) consider Europe or the EU alongside other parts of the world. Of the EU-specific sources, 73 per cent focus on individual countries or comparisons between individual countries. Among these, 14 (15 per cent) focus on the UK, half of which cover the UK alone while the other half examine the UK in comparison with other countries. Six papers (6 per cent) focus on Germany, with five of these papers (5 per cent) covering Germany only and a single paper considering Germany alongside five other democracies: Sweden, Denmark, Austria, the UK and the US. The three remaining European papers focus respectively on Denmark, Finland and the Czech Republic.

[38] MENA: Middle East and North Africa.

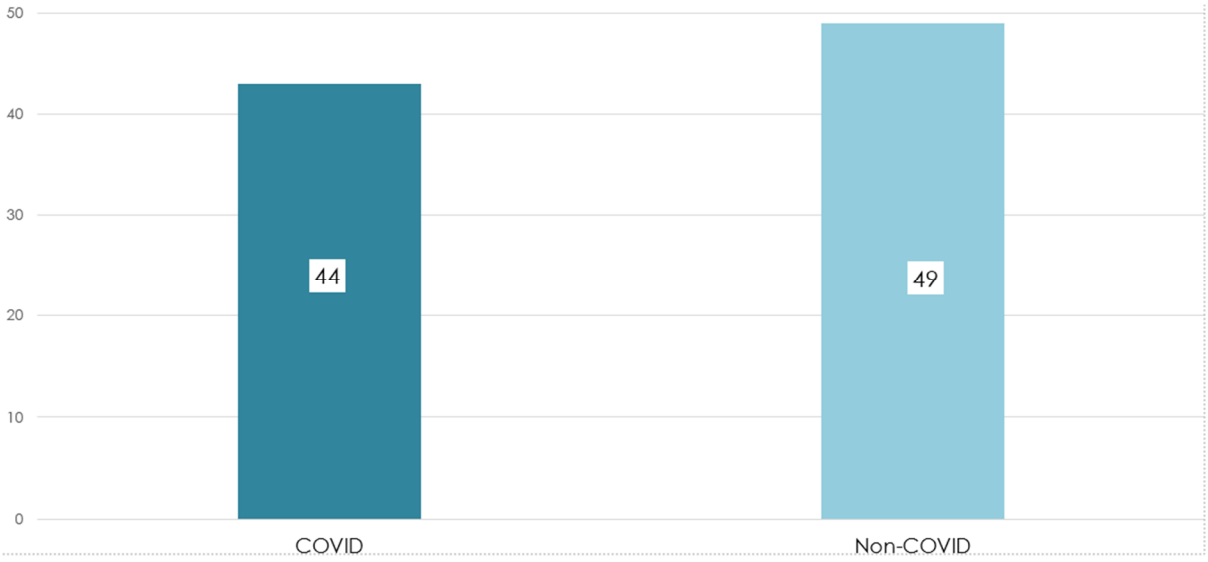
### 2.1.2. Most sources do not focus on specific types of hateful extremism but instead explore false information or hateful extremism more generally

Most of the reviewed sources do not focus on any specific type of hateful extremism, and instead explore broader trends relating to false information (35; 38 per cent). Other sources refer to hate speech in general (10; 11 per cent); (violent) extremism in general (9; 10 per cent); several types of extremism in combination (8; 9 per cent); xenophobia (4; 4 per cent); or conspiracy theories (2; 2 per cent). Where hateful extremist types *are* specified and focused on one area, the sources mostly focus on racism. With 24 papers (26 per cent) explicitly covering racism, this represents more than half of the sources focusing on a specific type of hateful extremism. Among the remaining papers, eight sources (9 per cent) exclusively cover right-wing extremism, five (5 per cent) exclusively cover Islamist extremism and four (4 per cent) exclusively examine antisemitism. While we have identified the above categories in the interests of clearly presenting the distribution of sources across hateful extremist 'types', it should nonetheless be noted that some reviewed papers conflate or identify links between different categories, particularly with regard to overlaps between racist narratives and right-wing extremism.

### 2.1.3. Of the reviewed sources, almost half focus on the COVID-19 context

As discussed in Chapter 2 and in Annex B of this document (see Section B.2), search strings were designed to identify papers that explicitly focused on the COVID-19 context, as well as those that explored false information, hateful extremism and associated responses more generally (i.e. outside of the COVID-19 context). Among the 93 reviewed sources, 44 papers (47 per cent) focus specifically on the COVID-19 context. Figure 2-1 illustrates the split between reviewed sources related to and not related to COVID-19.

**Figure 2-1: Distribution of sources related vs not related to COVID-19**




Source: RAND Europe (2020)

Papers that do not focus on the COVID-19 context represent a small majority of all reviewed sources (49 sources; 53 per cent). The focus of these sources, while diverse, can be grouped under three main themes:

- Around half of the non COVID-related sources describe policy and set out recommendations for governments and social media companies in relation to countering online hate, radicalisation and violent extremism.
- Around a third of non COVID-related papers present research on types of false information and how they further extremist agendas.
- Remaining sources provide mostly an analysis on the origins of different conspiracy theories and other types of false information.

### 2.1.4. The majority of reviewed papers are based on literature reviews

When examining the research methods used in the reviewed sources, the majority of papers rely on literature reviews. As shown in Table 2-1, 80 papers (86 per cent) are based on narrative literature reviews – which present a general, non-extensive assessment of relevant literature without disclosing selection methods – while three (3 per cent) are based on systematic literature reviews, offering a more thorough review approach with explicit selection criteria and extraction methods.[39]

Sources relying on other methods make up less than half of the reviewed sources. Three papers (3 per cent) are based on interviews, six on workshops/focus groups (6 per cent), five on analysis of secondary data (5 per cent) and 11 (12 per cent) on surveys. There were 19 papers (20 per cent) based on other methods, which included social media analysis, discourse analysis and case study analysis. As there are some overlaps (some sources employ more than one methodology), the figures exceed the 93 sources in the literature review. The distribution of reviewed papers according to underlying research methods is presented in Table 2-1.

[39] Ferrari (2015).

**Table 2-1: Distribution of reviewed sources according to underlying research method**

| Literature review (systematic) | Literature review (narrative) | Survey | Interviews | Workshops/ focus groups | Economic/ econometric analysis | Analysis of secondary data / management information | Other (please specify) |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

| Literature review (narrative) | Literature review (systematic) | Workshops/ focus groups | Interviews | Analysis of secondary data | Survey | Other |
|---|---|---|---|---|---|---|
| 80 | 3 | 6 | 3 | 5 | 11 | 19 |

Source: RAND Europe (2020)

Most of the reviewed sources are based on a single research method, particularly for those based on literature reviews and surveys. However, approximately 30 per cent of papers are based on a combination of methods:

- Some 27 per cent of sources based on narrative literature reviews and one source based on systematic literature reviews combine these methods with others, including workshops/focus groups, interviews, primary data analysis or surveys.
- The majority of papers relying on 'other' research methods (i.e. social media analysis, discourse analysis and case study analysis) are based on a combination of methods, including analysis of secondary data and, in a few cases, a literature review or interviews.

The literature review draws on a mixture of empirical analyses and expert opinions. As shown in the distribution of research methods in Table 2-1, a significant proportion of the research is based on secondary research. This highlights the importance of conducting primary research on this topic to generate new and timely evidence to inform policymaking, particularly as the COVID-19 pandemic continues to evolve and affect the spread of false information and hateful extremist behaviour. While the reviewed sources offer diversity in terms of geography, types of hateful extremism and COVID/non-COVID analytical framing, many papers are not robust in terms of providing strong, empirical evidence of the links between hateful extremism and false information, or offering policy recommendations based on tried and tested methods.[40] Nonetheless, as explored in Chapters 3 and 4,

[40] For a more detailed overview of the methods used in the sources reviewed, please refer to the separate Annex C.

the findings presented in this report provide a number of plausible hypotheses regarding associations and the effectiveness of online interventions and policies.

The 'good practices' identified in Chapter 4 are those identified in the reviewed literature as being effective, timely and relevant in curbing the online spread of false information and in building societal resilience to hateful extremism. It is worth noting, however, that the literature review examines material across several disciplines (including sociology, political science, psychology and law), and that some of the measures are implemented in different countries with distinct cultural, social, historical and legal contexts – meaning that what is reportedly effective in one context might not be in another. While measures to prevent hateful extremist speech might be accepted in one context, such efforts could be rejected elsewhere.

# 3. Links between hateful extremism and false information

This chapter presents findings in relation to the first review theme, which explores links between hateful extremism and false information. A summary of key findings in relation to the research questions set out in Chapter 1 is first presented in Box 2. The chapter then examines the incentives for hateful extremist groups and prominent individuals or influencers to share false information (Section 3.1), the content of hateful extremist narratives (Section 3.2), methods of circulating false information (Section 3.3), and how false information can shape hateful extremist beliefs and behaviours (Section 3.4). Finally, Section 3.5 outlines evidence gaps identified by the reviewed sources.

**Box 2: Key findings relating to review theme 1**

**1.1: What impact can false information have on hateful extremist beliefs and behaviours?**

- Overall, the literature sets out a number of hypotheses about the links between hateful extremism and false information, but there is limited empirical evidence that proves direct causality.
- The spread of conspiracy theories, misinformation and disinformation is associated with increasing levels of hate crime and blame on minority groups for COVID-19.
- The proliferation of false information has also led to the emergence of hateful extremist echo chambers, cementing hateful extremist attitudes by desensitising group participants to hateful content.

**1.2: In what ways do hateful extremist beliefs contribute to the spread of false information?**

- False information might serve extremist causes by:
  - Increasing exposure by infiltrating the mainstream, via the media or public officials who repeat hateful extremist messaging to a wider audience.
  - Supporting recruitment and fuelling violence by leveraging fear and blaming out-groups for crises (e.g. accusing Jews of orchestrating COVID-19).

**1.3: What trends and variations can be identified across different audience types, modes of false information, and extremist groups?**

- There is a shortage of research on the characteristics of individuals who are prone to hateful extremism, and a lack of evidence on the behavioural impacts of false information.
- During lockdown, a broader cross-section of the population – particularly young people – is consuming online content and potentially coming into contact with hateful extremist content.
- Far-right groups seek to blame migration, globalisation or the government for the virus, whereas Islamist extremists might see the pandemic as divine punishment against Westerners or infidels.

- Across different types of hateful extremists, groups tend to direct hostile narratives at 'out-groups', leveraging public fear and uncertainty surrounding COVID-19.

## 3.1. Benefits of false information for hateful extremist groups

As elaborated on below, hateful extremists are incentivised to spread false information by the heightened exposure and recruitment benefits afforded to them by doing so. With increasing social media engagement and time at home during lockdown, COVID-19 has also enabled hateful extremists to increase the reach of their false narratives and, in turn, increased opportunities for extremists to incite violence.

**False information serves extremist groups by offering them increased exposure**, particularly when it relates to COVID-19. The current pandemic provides an opportunity for hateful extremist actors to capitalise on public fears surrounding COVID-19, particularly when their narratives are featured in mainstream media.[41] Infiltrating the mainstream media provides hateful extremist actors with a platform to frame current issues, set agendas and spread their narratives to a wider audience.[42] The use of COVID-19 hashtags can also raise public visibility of hateful extremist messages, allowing extremist narratives to enter public debate.[43] The extremist narratives being reported are often deliberately shocking in their nature, generating clicks and drawing users to content. As individuals are increasingly relying on online channels for news content, this can have a damaging impact. Islamist extremist actors, such as ISIS, have been known to co-opt relevant hashtags and trending topics, creating so-called 'coronavirus pages' that funnel users to extremist content. Islamist extremist groups disguise content as 'health and wellbeing', where they celebrate the death toll in the West and link followers to extremist content, such as *The Punishment*, an Islamist extremist outlet.[44]

**Disinformation is a known tool for hateful extremist recruitment**.[45] As well as offering hateful extremists increased exposure, false information is also used to appeal to new recruits. Topics such as feminism and political correctness are examples of some of the 'gentle entries' to more extremist disinformation, and are used by hateful extremist recruiters to gauge how receptive potential supporters might be to more extreme narratives.[46] Disinformation on such topics is often touted as a 'red pill', revealing the underlying, unpleasant truths of the world, rather than the 'blue pill', which allows the majority of consumers to maintain 'blissful ignorance'.[47] Among far-right supporters, being 'red-pilled' refers to believing in narratives that go against the mainstream, such as Holocaust denial, the oppression of men by feminism, or white supremacy.[48] Becoming a believer in an issue such as

[41] Wilson (2020); Lewis & Marwick (2017).
[42] Lewis & Marwick (2017).
[43] Colliver and King (2020).
[44] Colliver and King (2020).
[45] Wallner (2020); Lewis & Marwick (2017).
[46] Baldauf et al. (2019).
[47] Baldauf et al. (2019); 'red pill', 'blue pill', was coined by The Matrix, a science-fiction action-film franchise.
[48] Lewis and Marwick (2017).

Men's Rights Activism reportedly makes an individual more likely to become 'red-pilled' on another more extreme belief. [49]

**Conditions for hateful extremist recruitment via hateful disinformation are better than ever before** in the COVID-19 context, with more people unemployed, housebound, online and consuming social media content.[50] The loss of social status and livelihoods has created an opportunity for hateful extremist actors to exploit individuals with a perceived or real lack of prospects. In this context, narratives of restored agency and purpose are used by hateful extremists to support recruitment and radicalisation.[51] Among children and young people, unsupervised screen time allows extremists to exploit their grievances during an uncertain period for families and communities.[52]

**Disinformation has also been used by hateful extremists to fuel hostility and strengthen resentment** in the COVID-19 context, which can lead to violence.[53] There has been a recent spike in online discussions around the 'boogaloo', a far-right term used to describe an impending 'second civil war' in the US, with calls for supporters to deliberately infect politicians, journalists and ethnic minorities.[54] A memo by the FBI notes how far-right groups have been urging followers to deliberately infect Jews with COVID-19, encouraging the use of spray bottles filled with infectious body fluids to be used in areas where Jews congregate, such as 'markets, political offices, businesses and places of worship'.[55] Targeted disinformation could accordingly lead followers to perceive minority groups as a threat, and subsequently encourage violence and hostility against such groups.

**The spread of disinformation during COVID-19 appears to have been beneficial for hateful extremists**. Pandemics could accelerate existing prejudices, contributing to the spread of hateful extremist beliefs and behaviours. This trend is evidenced by the historical association of infectious diseases with 'othering',[56] with late 19th-century US public officials blaming Asian immigrants for infectious diseases – including smallpox, leprosy and bubonic plague – referring to them as the 'disease ridden' carriers of sicknesses.[57] Today, narratives around COVID-19 have seen an increase in anti-Asian and antisemitic discourse, promoting harmful messaging about Chinese and Jewish communities. Allegations of 'networked complicity' have been present for centuries, asserting that the Jewish community holds excessive financial resources, political power and influence over media institutions.[58] During COVID-19 there has been more messaging of this type, in some cases depicting the pandemic as a Jewish hoax, a plot to depopulate the world or a scheme to start a new world war.[59]

[49] Lewis and Marwick (2017).
[50] Avis (2020).
[51] Avis (2020).
[52] Naseer (2020).
[53] Schwarz and Holnburger (2019).
[54] Colliver and King (2020).
[55] Malik (2020).
[56] Devakumar et al. (2020).
[57] Liz (2020).
[58] Holbrook (2020).
[59] CST (2020).

## 3.2. Hateful extremist narratives and the use of false information

As explored in more detail below, hateful extremist narratives tend to incorporate false information by focusing on 'othering' and victimhood; targeting minority groups, political opponents and other 'out-groups'; and highlighting perceived societal problems. COVID-19-related disinformation features in recent Islamist and far-right extremist narratives, and non-hateful or non-extremist narratives are also used to widen the support base for extremists.

**Hateful extremist narratives typically incorporate disinformation by focusing on 'the other'**. A 2019 report produced by the UK-based organisation Hope not Hate identifies several elements that are central to hateful extremist disinformation. First, hateful extremist narratives use disinformation to criticise other groups,[60] using these narratives to distinguish between the in-group ('us') and the undesirable out-group ('them'). Second, hateful extremist narratives use disinformation to highlight an unwillingness to mix or integrate with other groups, which is often coupled with a sense of victimisation by other groups.[61] Finally, these false narratives carry specific intentions, including the use or support of violence to achieve political goals.[62] Due to the generalist nature of hateful extremist narratives – as seen in their tendency to portray complex issues in simplistic terms – the disinformation shared by hateful extremists on issues such as COVID-19 is often focused on entire communities, interpreting the world via these simplistic understandings and stereotypes.[63]

**Far-right and Islamist extremists both use disinformation to promote narratives of victimhood** and 'othering'. For example, just as far-right extremists often spread disinformation referring to immigrants and Muslims as 'invaders' who seek to actively destroy European ethnocultural homogeneity, Islamist extremists portray 'nonbelievers' as their aggressors.[64] Some commentators highlight that the messaging of far-right extremists and Islamists is often interdependent, and that the activity of one group can ignite a reaction in the other.[65] Islamists have been shown to react strongly to far-right demonstrations and political activity, reinforcing narratives that the West is anti-Muslim. Similarly, there are reported spikes in the volume of anti-Muslim disinformation on social media following Islamist terror attacks.[66] Nonetheless, the narratives of the far right and Islamist extremist groups also overlap at times, with both groups focusing on common topics that are likely to draw media and public attention (e.g. transphobic narratives). Hateful extremist actors accordingly use disinformation in similar ways to generate a sense of 'otherness' and, in the case of far-right groups and Islamist extremists, the narratives often fuel one another.

**Extremist narratives and conspiracy theories often blame political institutions for societal problems**.[67] The 'out-groups' that feature in hateful extremist narratives can also be political, with such narratives

[60] Lowles and Levene (2019).
[61] Lowles and Levene (2019); Holbrook (2020); Guhl & Ebner (2018).
[62] Lowles and Levene (2019).
[63] Holbrook (2020); Ariza (2020); UNESCO (2020).
[64] Holbrook (2020)
[65] Guhl and Ebner (2018).
[66] Guhl and Ebner (2018).
[67] Vegetti and Levente (2020); Wilson (2020).

fuelling animosity towards institutions, procedures and key public actors.[68] For example, the statements developed by Uscinski (2016) to identify conspiracy thinking among survey respondents are designed to capture feelings of insecurity, political helplessness and a rejection of political institutions:

- 'Much of our lives are being controlled by plots hatched in secret places.'
- 'Even though we live in a democracy, a few people will always run things anyway.'
- 'The people who really "run" the country, are not known to the voters.'[69]

These sentiments can manifest in an unwillingness to engage with representative democracy or, at an extreme, attempts to overthrow democratic institutions.[70] In their empirical study, Jolley and Douglas (2014) found that exposure to false information affected focus group participants' intentions to engage in political processes, including voting.[71] This could be exploited by hateful extremist actors or adversary states in order to disrupt the democratic process or garner support from sceptics.[72] Indeed, hateful extremist actors often seek to alienate citizens from political institutions. For example, the German far-right group Der Dritte Weg (*the Third Way*), has stoked further distrust in politicians by publishing the narrative that German leaders have exploited the pandemic as a 'diversionary tactic' to distract from the oncoming 'flood' of refugees and migrants.[73]

**Far-right extremists have woven COVID-19 into their narratives**, hoping to exploit the global situation to create divisions and sow fear.[74] The far right has been highly responsive to COVID-19, prompting several studies on their narratives and activities during lockdown.[75] Far right conspiracy theories about COVID-19 often follow two trains of thought: (1) the belief that the virus is a hoax to justify imposing a totalitarian state; and (2) the view that the virus has been manufactured as a bioweapon.[76] The nuances of online far right narratives were examined by the International Centre for Counter-Terrorism in an analysis of the Telegram statements released by six far-right groups between 22 February 2020 to 22 April 2020.[77] The study categorises these narratives according to six frames:

1. **Migration** is a core driver of the spread of COVID-19;
2. **Globalisation** and multiculturalism have allowed COVID-19 to spread;
3. **Bad governance** is at the centre of the impact of COVID-19;
4. **Liberty** is at risk during COVID-19, as evidenced by the expansion of a 'security state';

[68] Vegetti and Levente (2020); Douglas et al. (2019a).
[69] Uscinski (2016).
[70] Lowles and Levene (2019); Jolley & Douglas (2014); Wilson (2020).
[71] Jolley & Douglas (2014).
[72] Lewis & Marwick (2017).
[73] Colborne (2020b); Der III. Weg (2020).
[74] UN (2020).
[75] McNeil-Willson (2020); Velasquez et al. (2020); Lu & Sheng (2020).
[76] Ariza (2020).
[77] McNeil-Willson (2020).

5. Far-right groups create **resilience** to COVID-19, by emphasising engagement in activities designed to build community resilience;
6. COVID-19 is a **conspiracy** or a deliberate distraction from more important issues.[78]

**COVID-19 similarly features in recent Islamist extremist narratives**, particularly in formal public statements issued by al-Qaeda and the so-called Islamic State providing guidelines to prevent the spread of COVID-19 within Islamist-controlled territory, with al-Qaeda highlighting that 'Islam is a hygiene-oriented religion'.[79] In this case, Islamist extremists are not discrediting COVID-19 as a hoax, but are instead treating the pandemic with caution. Islamist extremists also use the virus to target their enemies in disinformation campaigns, but they have taken a somewhat different tack to far-right extremists in weaponising the pandemic and attributing causation. Two potential narratives could emerge, depending on the trajectory of the pandemic: (1) if the virus spreads within Muslim-majority areas, there could be a rise in conspiracy theories blaming the West or Jews; and (2) if the spread of the virus does not reach Muslim-majority areas, especially regions under Islamist control, the pandemic could be framed as divine punishment against out-groups.[80]

By contrast to these narratives, extremists also use non-hateful narratives to attract support and to create favourable impressions of these groups. These tactics can help extremist groups generate sympathy for 'the cause',[81] potentially affording some legitimacy to the hateful narratives and actions of these actors. During the pandemic, for example, far-right groups have volunteered and delivered supplies to citizens,[82] with these actions forming part of a broader narrative that falsely positions hate actors as champions of the people.[83] In Germany, members of *Die Rechte* (the Right) have delivered supplies to low-income households with notes calling them the 'backbone' of the country in attempts to enlist recruits.[84] In the UK, the far-right group Britain First has similarly shared videos of members volunteering for the NHS.[85] In this way, McNeil-Willson (2020) states that extremist groups seek to win over the public by presenting themselves as public champions and service providers.[86]

These non-hateful narratives and approaches could win over those who are not convinced by hateful narratives alone, attracting sympathy from a wider selection of the population. While building support is a key goal for hateful extremist groups, there is little evidence to suggest that these hateful extremist efforts to attract public sympathies lead directly to a greater public following. Nonetheless, hateful extremist framing to build sympathy could prove problematic for policymakers; by emphasising the shortcomings of mainstream centrist parties and their responses to the pandemic, such narratives could potentially sow greater division within societies. Furthermore, it is difficult to

[78] McNeil-Willson (2020).
[79] Wilson Centre (2020).
[80] Campbell (2020); Avis (2020).
[81] Ariza (2020).
[82] Ariza (2020).
[83] Colborne (2020a).
[84] Ariza (2020).
[85] Ariza (2020).
[86] McNeil-Willson (2020).

frame a response to far-right groups volunteering in communities, as preventing them from helping communities could confirm the narrative that the government does not prioritise public interests.[87]

## 3.3. Hateful extremist methods of circulating false information

As this section will explore in more detail, hateful extremists are making increasing use of social media – particularly via 'bots' or automated accounts – to spread false information, relying both on explicit calls for violence and more ambiguous language to influence online users.

**Hateful extremist actors are increasingly using social media to spread false information**. Through analysis of over 600 million tweets and comparison against data from five years ago,[88] a 2020 Moonshot study found that there has been a clear spike in antisemitic and anti-Chinese narratives stemming from false information.[89] For example, some hate actors associate COVID-19 with a fictitious Jewish plot to initiate civil wars,[90] or blame Chinese communities for spreading the virus.[91] In addition, UK hate actors are using hashtags such as #GermJihad to target Muslim groups.[92] Hateful narratives circulated on social media are often expressed in subtle terms; in a 2019 analysis of 5.2 million tweets from British National Party (BNP) supporters, Vidgen et al. find that 10.8 per cent of content contains implicit Islamophobia, while 5.3 per cent of tweets are explicitly Islamophobic.[93] According to experts at the European Commission (2018), online hateful extremist content seeks to attract the support of 'out-groups'[94] to foster societal tensions, polarisation and suspicion, often in support of radical ideas and activities.[95]

**Hateful extremist disinformation is often disseminated via 'bots'** (automated social media accounts), and techniques are shifting to more sophisticated interaction with users through impersonation and amplification of organic posts.[96] During the May 2019 European elections, for example, foreign state disinformation campaigns amplified European extremist messaging, undermining centrists and establishment parties.[97] Coordinated inauthentic behaviour, particularly via bots and fake social media accounts, was used as an attempt to subvert the elections, posing a significant challenge to authorities and social media moderators.[98] Such techniques allow for fringe debates to enter the mainstream discussion, providing a platform for extremist messaging. 'Triggering' – the use of provocative language to cause an overreaction in mainstream reporting – is another method used by extremist actors to increase such engagement.[99] Alternatively, 'doxxing' is used to intimidate journalists by

[87] Colborne (2020a).
[88] Manavis (2020).
[89] Manavis (2020).
[90] CST (2020).
[91] Pei and Mehta (2020).
[92] CCDH (2020).
[93] Vidgen et al. (2019b).
[94] 'Out-group' is a group to which a person does not identify themselves as belonging (Abbink and Harris 2019).
[95] EC (2018b).
[96] Polyakova and Fried (2019); 'Organic posts' refer to those made by real people, as opposed to bot activity.
[97] Polyakova and Fried (2019).
[98] EC (2019).
[99] Baldauf et al. (2019).

disclosing their personal information online;[100] and 'source-hacking' entails sharing disinformation with credible media analysts so that this hateful extremist content is quoted by reputable sources and further pushed into the mainstream.[101]

**Hateful extremists rely both on explicit calls for violence and more ambiguous, coded language**.[102] 'Dog whistling' or cloaking language is often used online, whereby content uses coded or suggestive words or phrases, with a hidden meaning understood by some but not all groups. For example, the term 'Cultural Marxism' has antisemitic connotations in British politics,[103] similarly to how references to 'Zionist agents' is problematic.[104] It is difficult for governments, social media companies and civil society to navigate responses to such content, as these narratives are often not explicit in their hateful content, and hence may not violate social media terms of use. However, disguised hate could cause harm, as these can incorporate inflammatory content and a susceptible audience – two elements required for narratives to be dangerous.[105]

## 3.4. Implications of false information for hateful extremism

As this section explores in more detail below, false information is associated with increasing levels of hate crime, blame on minority groups for COVID-19, poor mental and physical health outcomes, and the emergence of hateful extremist echo chambers.

**A rise in hate incidents at a time when xenophobic language based on false information is used by certain elected politicians across the world** is noted in the reviewed literature, although not all sources agree there is a causal relationship.[106] Political rhetoric has been shown to influence public opinions and behaviour, with this rhetoric impacting on public perceptions of a foreign country, for example.[107] Some commentators argue that in addition to being misleading, the use of language such as 'China Virus' or 'Chinese Virus' by certain public officials can exacerbate the 'othering process' in the COVID-19 context.[108] A US-based analysis of the tweets of prominent political figures mentioning both China and COVID-19 has argued that there could be a spike in 'racial animus' on days when the tweets of such officials mention both China and COVID-19.[109]

**Higher exposure to false information in lockdown is associated with increased targeting of minorities**. In the US, there has been an increase in discriminatory behaviours during COVID-19, with the New York City Commission on Human Rights reporting a 92 per cent increase in anti-Asian discrimination incidents between March and May 2020, compared to the same three-month period in the previous year.[110]

---

[100] Baldauf et al. (2019).
[101] Baldauf et al. (2019).
[102] Benesch et al. (2020).
[103] Antisemitism Policy Trust (2020).
[104] BBC (2019).
[105] Benesch et al. (2020).
[106] Cabanatuan (2020); Jeung (2020) cited in Gover et al. (2020).
[107] Silver (2016); Lu and Sheng (2020).
[108] Gover et al. (2020).
[109] Lu and Sheng (2020).
[110] Liz (2020).

Similarly, in the UK, hate crime directed at South and East Asian communities increased by 21 per cent from March to May 2020 during the beginning of COVID-19, a surge believed to be driven by the increased use of social media and other online platforms during lockdown, exposing more people to false information.[111] Pei and Mehta (2020)'s analysis of 174,488 tweets with the hashtags #Chinesevirus and #Chinavirus finds that – amid the spread of false information – immigrant, ethnic and religious minority groups (e.g. Jewish, Chinese and Muslim communities) have been blamed for the spread of the virus by some members of the public.[112]

**False-information-driven discrimination has been linked to negative health outcomes in the literature**, particularly in the COVID-19 context. According to Priest et al. (2020), poor mental and physical health is linked to experiences of racial discrimination driven by false information. In this source, experiences of racial discrimination are associated with depression, behavioural difficulties, anxiety, sleep disruption, and a higher risk of suicide among children.[113] The same study also finds growing evidence that racial discrimination is associated with obesity, high blood pressure and inflammation, as well as epigenetic ageing among children and young people who are targeted by discrimination fuelled by false information.[114] A 2020 study from the University of Oxford identifies another way in which false information is linked to poor physical health, finding that people who hold conspiracy beliefs relating to COVID-19 are less likely to comply with social distancing guidelines or accept future vaccines.[115] These are two very different ways in which false information can have a negative effect on health.

**False information has also been associated with an increase in echo chambers**, in which extremist views reverberate with little opposition or exposure to alternate views. Exposure to hate has been linked to the normalisation and indoctrination of violent themes.[116] Hateful extremist groups often experience 'polarisation effects': as more moderate or sceptical users opt out of these chambers, the circle closes, leaving a concentration of like-minded believers without exposure to differing views.[117] In general, people with more extreme political views engage with a smaller number of individuals than those who hold more moderate views.[118] For example, on Facebook, content is posted in private groups, side-stepping social media moderators, which rely on users to flag false information or hateful extremist content.[119] These private groups act as echo chambers, in which misleading information is not reported as often due to inherent agreement and mutual support among members of the group. Furthermore, such chambers on mainstream platforms can act as a 'funnel', pulling individuals into less moderated platforms, such as 4Chan or Telegram.[120]

[111] Grierson (2020).
[112] Pei and Mehta (2020).
[113] Priest et al. (2020).
[114] Priest et al. (2020).
[115] University of Oxford (2020).
[116] Allington (2020).
[117] Marwick and Lewis (2017).
[118] CoE (2017a).
[119] Ball and Maxmen (2020).
[120] Ball and Maxmen (2020).

## 3.5. Evidence gaps

The reviewed sources highlighted two key gaps in relation to the first review theme: (1) a shortage of empirical evidence on the characteristics of individuals who are particularly vulnerable to the influence of conspiracy theories and hateful extremism; and (2) a lack of research on the behavioural impacts of online hateful extremist speech.

### Shortage of empirical research on characteristics of individuals who are susceptible to the influence of conspiracy theories and hateful extremism

Research on conspiracy theories tends to focus on online content rather than the motivations behind it. Fangen and Holter (2020) point out that research has, until now, 'largely focused on analysing digital content and not its producers', and note that the field could benefit from more studies on the self-perceptions and self-understanding of individuals who believe and spread conspiracy theories.[121] With regards to conspiracy theories in particular, Radnitz and Underwood (2015) suggest that current research neglects a large portion of believers entirely. According to those authors, the stigma attached to conspiracy theories prevents many individuals prone to conspiratorial thinking to admit it, leading to the exclusion in current research of so-called 'ordinary believers', in favour of louder voices. The lack of focus on these individuals could prevent a better understanding of conspiracy theories and their ubiquity.[122]

These evidence gaps are echoed by Connolly et al. (2019), who call for more research on how conspiracy theories spread through the information environment and on investigating which individuals are most at risk of adopting them.[123] In addition, several studies point to a lack of understanding regarding the dynamics behind directional motivations – that is, the individual's motivation to hold on to existing convictions and attitudes. Nyhan and Zeitzoff (2018) suggest that future studies should consider the role of individuals' media exposure in directional motivations. The same study recommends that future research should consider experimental manipulation of directional motivations in an ethical way to better understand the dynamics behind them.[124] Given the literature finding that susceptibility to conspiracy theories is linked to susceptibility to hateful extremist beliefs,[125] addressing this research gap could improve understanding of individuals who are vulnerable to the influence of both conspiracy theories and hateful extremism.

### Lack of evidence on the behavioural impacts of online misinformation and hateful extremist speech

Beyond links between misinformation and hateful extremist speech – which Ziems et al. (2020) see as an essential area for further study – some of the reviewed studies point to a lack of research on the direct behavioural impact of online false information. Albadi et al. (2019) see the impact of bot-

[121] Fangen and Holter (2020).
[122] Radnitz and Underwood (2015).
[123] Connolly et al. (2019).
[124] Nyhan and Zeitzoff (2018)
[125] Douglas et al. (2019); McNeil-Willson (2020).

disseminated content on human behaviour as an important field for future research. In particular, the study points to the need for investigation on whether hateful content disseminated by bots influences individuals' participation in hateful discourse.[126] In addition, Schild et al. (2020) highlight the need to develop new techniques to understand changes of online behaviour in the context of COVID-19, and to understand and prevent real-world violence resulting from these online behaviours.[127]

[126] Albadi et al. (2019).
[127] Schild et al. (2020).

# 4. Interventions and policy responses

The previous chapter set out the implications of false information and explored how hateful narratives can further the agendas of extremist groups. In this chapter, we present findings in relation to the second review theme, examining online interventions and policy responses for countering false information. As noted in Section 1.2, when discussing interventions, this report presents the findings and recommendations of reviewed sources rather than offering an independent assessment of their effectiveness.

A summary of key findings in relation to the research questions set out in Chapter 1 is presented in Box 3. An overview of existing and recommended future interventions and policy responses is then presented (Section 4.1), before the chapter describes 'good practices'[128] and challenge areas for actors involved in the delivery of initiatives to counter false information and hateful extremism as reported in the literature (Section 4.2). Finally, Section 4.3 outlines evidence gaps highlighted in the reviewed sources.

**Box 3: Key findings relating to review theme 2**

**2.1: What insights can be identified from the literature on the effectiveness of existing interventions and policy responses?**

- Four main types of online response to false information were identified in the literature review – factchecking, counterspeech, takedowns and education – and there is a shortage of existing empirical evidence on the impact of existing interventions.
- While the reviewed literature does not offer rigorous evaluations, papers identify several promising practices promoted by the following core actors with responsibility for tackling false information:
  - **Government** – reviewed sources observed for example that government actors have effectively held social media companies to account (e.g. 2018 German NetzDG Act), in addition to engaging in wider counter-extremism efforts.
  - **Social media companies** – papers highlighted the contribution of these actors to tackling false information by reconfiguring algorithms to ensure that hateful content is not promoted.

[128] 'Good practices' are those identified in the reviewed literature as being effective, timely and relevant in curbing the online spread of false information and in building societal resilience to hateful extremism. It is worth noting, however, that the literature review examines material across several disciplines (including sociology, political science, psychology and law), and that some of the measures are implemented in different countries with distinct cultural, social, historical and legal contexts – meaning that what is reportedly effective in one context might not be in another. Furthermore, it should be noted that the quality of source data varies across the reviewed papers: while some sources offer empirical evidence with a narrow focus on addressing a specific problem, others offer untested commentary, observations and broader recommendations from recognised experts (see Annex C).

- **Civil society** – papers noted the role of these actors in educating and raising awareness of false information in relation to hateful extremism.
- **Media organisations** – sources focused on good journalism practices, including the avoidance of 'clickbait' headlines and maintaining transparency in factchecking.

**2.2: What recommendations are put forward in the existing literature in relation to future interventions in this area?**

- The 'good practices' identified below are those highlighted in the reviewed literature as being effective, timely and relevant in curbing the online spread of false information and in building societal resilience to hateful extremism.
- Approaches that are recommended in the reviewed papers include:
  - Cooperation between social media companies;
  - National government dedication of resources to combating false information;
  - Increased education to build societal resilience;
  - Greater accountability for news organisations and use of responsible headlines;
  - Increased research into implications of online false information; and
  - Greater investment in education programmes to enhance critical thinking skills.

**2.3: What transferrable lessons/'good practices' from successful interventions in related policy areas can be identified?**

While this review set out to address this question, we did not identify successful interventions from other policy areas as they did not appear in the reviewed literature. Although there are no direct lessons or good practices to be transferred from other policy areas identified from the review, this report benefits from a multidisciplinary approach, drawing from literature across psychology, political science, sociology and law.

## 4.1. Categories of counter-measures

This section sets out current and potential counter-measures and interventions to mitigate the spread of hateful extremist false information, and to build societal resilience. Reflecting the focus of the reviewed literature – which spans the fields of psychology, sociology, communication studies and political science – these counter-measures are presented in relation to four areas of work: (1) **factchecking**; (2) **counterspeech**; (3) **takedowns**; and (4) **education**. The sections below describe insights from the reviewed literature on recommended approaches to designing and delivering each type of counter-measure, as well as associated limitations. However, it is worth noting that the impact of false information and of associated counter-measures differs from one society to another, depending on factors such as education levels, trust in institutions, inclusiveness of electoral systems and socio-economic inequalities[129] – so what is 'recommended' or 'effective' in one setting might not be in another. These factors have not been evaluated in this study, as it is beyond the scope of this report.

[129] EC (2018b).

### 4.1.1. Factchecking

Factchecking refers to activities to assess the veracity of content, particularly political claims.[130] Factchecking seeks to prevent the further dissemination of false information by a range of actors including hateful extremists, and to build resilience among consumers of online content. In many cases, factchecking can be understood as a type of persuasion that requires attitude change, to convince readers that the information they previously consumed is factually false.[131]

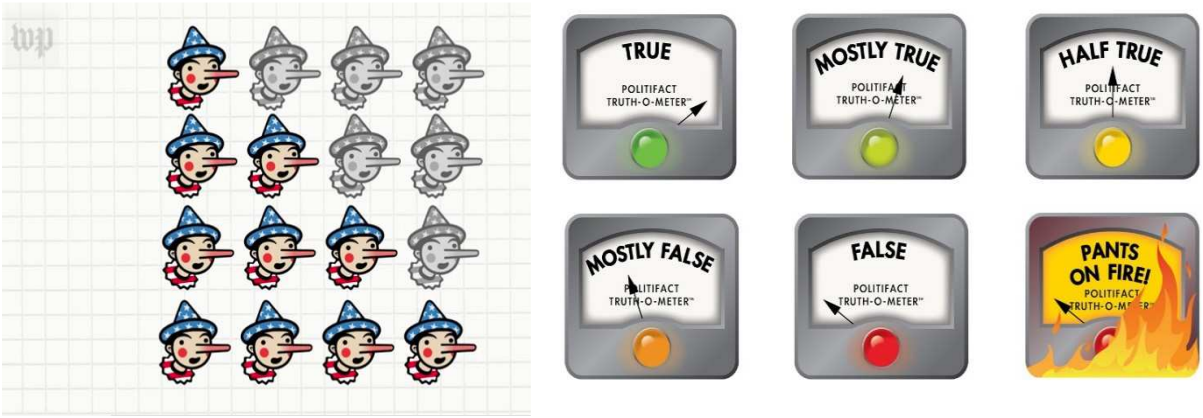
#### Insights from the literature on recommended approaches for factchecking

**Clarity and transparency are frequently referred to as important features for successful factchecking**, as this builds trust among readers.[132] At a minimum, Humprecht (2019) finds that organisations should aim to provide the background information of sources used for factchecking, to build credibility and reach those who might be susceptible to the influence of false information.[133] In addition, Thorson (2016) posits that it is important to explain the reason for false information being generated in the first place, although this might not always be understood by readers or indeed be possible to determine.[134] Furthermore, Thorson observes that affirmations are particularly likely to persuade readers.[135] For example, with reference to the Obama 'birther' conspiracy,[136] Thorson (2016) finds that the statement 'Obama is a Christian' is more likely to influence popular opinion than the negation 'Obama is not a Muslim'.[137]

**Independent factchecking rating scales could increase the effectiveness of correction among readers**. Empirical research undertaken by Amazeen et al. (2018) suggests that when independent factchecking organisations use rating scales, the effectiveness of correction among readers can be increased.[138] Rating scales illustrate the degree of veracity of content by colour-coding or allocating a numerical rating. According to this empirical research, these scales are demonstrated to be more effective for correcting non-political views than information that goes against the prior-held beliefs of readers.[139] Figure 4-1 presents two examples of rating scales to illustrate different ways of visualising factchecking.

[130] Amazeen et al. (2018).
[131] Amazeen et al. (2018).
[132] Humprecht (2019); Brandtzaeg & Folstad (2017); Poynter (2019).
[133] Humprecht (2019).
[134] Thorson (2016).
[135] Thorson (2016).
[136] The 'birther' conspiracy alleged that former President Barack Obama was born outside the US and ineligible to serve as president, with an ethnic and religious element injected into the conspiracy via the allegation that Obama was a Muslim.
[137] Thorson (2016).
[138] Amazeen et al. (2018).
[139] Amazeen et al. (2018).

**Figure 4-1: Examples of rating scales (Washington Post and PolitiFact)**[140]




**Big data and automation have the potential to enhance accuracy and reduce the human burden of factchecking**.[141] Human factcheckers face challenges in keeping up with the sheer volume and speed at which false information is proliferated online.[142] There have been many calls from communication specialists to automate factchecking.[143] Indeed, certain elements of factchecking are already being automated. For example, UK charity Full Fact has built a system to run factchecking tasks by identifying content that might be untrue, matching the language to facts within its database and subsequently publishing results online.[144] Similarly, the Duke Reporters' Lab and Chequeado have developed automated factchecking tools.[145] So far, these systems are only reported to be capable of identifying simple declarative statements, rather than implied claims or those embedded in more complex sentences.[146] For further progress to be possible, Graves (2018) note that governments will need to provide continued support for basic research and real world experiments, and cooperate with civil society organisations towards establishing open data standards. Furthermore, Graves notes that news organisations will have to become more active in this area, both by contributing their vast factchecking resources and expertise, and seeking to benefit from advances in this field.[147] In the future, entirely automated factchecking platforms could detect content in real time and rate its accuracy.

## Literature findings on the limitations of factchecking

Despite the reported benefits of factchecking, it is evident that such measures have inherent limitations and have to be designed carefully to avoid certain pitfalls.

**When readers already hold a particular belief, they are reportedly less likely to accept corrections**,[148] particularly when such views relate to political content.[149] This is often leveraged by extremist actors,

[140] Poynter (2019).
[141] Thorne & Vlachos (2018).
[142] Humptrecht (2019).
[143] Cohen et al. (2011); Graves (2018); Thorne & Vlachos (2018).
[144] Full Fact (2020).
[145] Poynter (2018).
[146] Graves (2018).
[147] Graves (2018).
[148] Flynn et al. (2017); Amazeen et al. (2018); Barrera et al. (2020).
[149] Amazeen et al. (2018).

who capitalise on political events to push forward their narratives. Here it is worth noting that populist undercurrents have politicised issues that have traditionally been apolitical or nonpartisan. COVID-19 has been politicised, with public officials and the media across the UK, Europe and the US highlighting funding cuts to healthcare, political leaders' handling of the virus, and the impact of lockdown measures on civil liberties.[150] 'Political content' is therefore a 'bigger basket' than ever, presenting a significant challenge to factcheckers as readers now hold political opinions on a larger range of topics. An empirical study conducted by Barrera et al. (2020) also finds that where factchecking is successful, it merely updates the factual knowledge of readers, but does not necessarily affect their policy conclusions or their support for the political candidate espousing false information.[151] The long-term effects of continuous factchecking and holding the statements of political candidates to account is an area in need of further academic input, to fully assess the impact and importance of factchecking.

**The evidence is mixed on whether acknowledging conspiracy beliefs tempers or contributes to such views**. Bolsen and Druckman (2018) emphasise the importance of engaging with readers, and report that acknowledging alternative beliefs is more likely to persuade readers than denying such beliefs. According to those authors, individual views can be changed when their beliefs in a conspiracy theory are recognised while also offering scientific consensus information.[152] While the acknowledgment of conspiracy beliefs can temper the impact of these beliefs, validation could also imply legitimacy and contribute to the spread of false information.[153] Further research is needed to assess how acknowledgment of false information can affect other beliefs and readers' openness to other information.

### 4.1.2. Counterspeech

Counterspeech entails the use of narratives to counter and offer alternative narratives to false or misleading information. While counterspeech can take various forms, online counterspeech is considered an important way to directly refute or challenge hateful extremist use of false information. Actors who engage in counterspeech include individual members of the public, civil society organisations or public sector officials.

#### Insights from the literature on recommended counterspeech approaches

**Sources note the need to design counterspeech in a nuanced way that taps into underlying public concerns**. To use counterspeech against hate actors, the response must be designed carefully.[154] If used successfully, counterspeech can undermine the authority of hateful extremist false information, as well as highlight to the public that there are organised movements working to actively counter hateful extremism and false information.[155] In addition to providing wide-ranging intervention programmes, for example, the UK's Prevent strategy and STREET (Strategy to Reach, Empower, and Educate Teenagers) focus on counternarrative messaging that deconstructs extremist narratives and promotes

[150] KHN (2020).
[151] Barrera et al. (2020).
[152] Bolsen and Druckman (2018).
[153] Bolsen and Druckman (2018).
[154] Allington (2020); CoE (2017b).
[155] CoE (2017b).

mainstream, moderate perspectives.[156] Similarly, Article 19 – a UK organisation focusing on human rights and freedom of expression and information – emphasises the importance of public officials condemning hateful extremist narratives online, establishing that the most effective responses are nuanced and provide persuasive counter-narratives that appeal to, and potentially challenge, underlying concerns and anxieties of the public.[157] To this end, the organisation encourages equality education and awareness raising for public officials, so that they are better equipped to respond to hateful narratives and false information online or offline (see Section 4.2). This could be facilitated by increased cooperation between government and civil society.[158]

**The Council of Europe (CoE) emphasises that framing counterspeech from a human rights perspective is important**, avoiding 'dichotomous and adversarial framing'.[159] For example, the slogan 'do not hate the migrants, hate the bankers' simply shifts the object of hate. Instead, the CoE proposes three core rules for counterspeech:

1. Counternarratives should not include hate, violence and discrimination;
2. Counternarratives should foster equality, respect and solidarity; and
3. Counternarratives should promote an understanding of the equal dignity of all human beings, and promote critical thinking, fair dialogue and correct information.[160]

**Counterspeech is also seen as more effective when refutations are concise**. While lengthier responses can allow for a greater degree of refutation, there are significant diminishing returns on the length of a refutation regarding its impact on observers, as many readers will not bother to digest the full response.[161] Allington (2020) argues that digital generations are moving towards a new type of learning and multitasking, which reduces their ability to write or read lengthy refutations online[162] – meaning that shorter responses are more likely to resonate. Similarly, participants in a 2014 survey agreed that YouTube videos offering counternarratives in response to Al-Qaeda-related content should remain short and to the point, as the inclusion of additional, potentially unnecessary information could 'lose' the viewer.[163] Excessive online engagement with hateful extremist actors – both in terms of time and word count – could prove counterproductive, attracting more attention to the messages of these actors.[164]

**The evidence is mixed on the effectiveness of using emotionally evocative counternarratives**. On the one hand, emotional responses could be viewed by hateful extremist actors and bystanders as 'taking the

[156] Bilazarian (2020); Barclay (2011).
[157] Article 19 (2018).
[158] Article 19 (2018).
[159] CoE (2017b).
[160] CoE (2017b).
[161] Allington (2020).
[162] Allington (2020).
[163] Hedayah & ICCT (2014).
[164] Allington (2020).

bait', and heightened emotional reactions to 'dog-whistling'[165] could also be viewed as disproportionate by bystanders unaware of the underlying meaning of the messaging,[166] highlighting the importance of using neutral, non-emotional language. On the other hand, some scholars posit that emotionally evocative counternarratives can surpass logical arguments when used as part of video-based counternarratives.[167] For example, ISIS defector videos featuring interviews with former extremists are often edited to highlight the most 'damaging, denouncing and derisive content'.[168] These counternarrative-based videos are subsequently uploaded with pro-ISIS titles, so that individuals seeking out ISIS recruiting material are instead directed to anti-ISIS content.[169]

Similarly to factchecking (Section 4.1.1), **automating counterspeech is presented as a promising avenue** for further investigation in the reviewed literature. Indeed, there has been considerable research into automated counterspeech, and how this can be done rapidly and effectively.[170] Natural language generation is presented by Tekiroglu et al. (2020) as a feasible automation strategy towards generating text responses to hateful extremist content, noting that there is still a need for verification systems to check responses.[171] In the coming years, automated counterspeech could become more effective, presenting a possible solution in tackling online false information and hateful extremist content. For now, this remains a promising area for further investment and research.

### Literature findings on the limitations of counterspeech

**There are nonetheless several challenges and limitations associated with counterspeech**, particularly given the sheer quantity of content being generated every day and noting the use of bots in spreading hateful extremist narratives and false information across platforms. As established above, counterspeech can be less effective when refutations are overly lengthy or excessively emotional. According to a recent study (Allington, 2020), in some cases counterspeech can cause backlash or appear defensive and reactive, playing into the intentions of hateful extremist actors who seek to capture wider attention and generate further debate in order to spread their narratives.[172]

[165] Content that uses coded or suggestive words or phrases, with a hidden meaning that is often understood by some groups but not others.
[166] Allington (2020).
[167] Tuck & Silverman (2016).
[168] McDowell-Smith et al. (2017).
[169] McDowell-Smith et al. (2017).
[170] Tekiroglu et al. (2020).
[171] Tekiroglu et al. (2020).
[172] Baldauf et al. (2019).

### 4.1.3. Takedowns

On major social media platforms, moderators often take down abusive posts – particularly content associated with automated accounts, pages and content working to actively manipulate discourse on social media (so-called 'inauthentic activity').[173] When it comes to hateful or misleading narratives espoused by prominent public figures, platforms take varying approaches. While some networking sites issue factchecking notices under inaccurate posts of public figures, others, such as Facebook, have argued that they should not be an 'arbiter of truth'.[174]

#### Insights from the literature on recommended approaches relating to takedowns

**Takedowns can prevent exposure to hateful extremist content**, making it harder for extremists to recruit followers or influence debates. Social media companies often rely on users to report comments, which are subsequently reviewed and potentially removed. However, implementing takedowns at the same rate as bots and human actors spreading hateful content is a significant task.

**Automated systems for reviewing and actively removing content can offer efficiencies** for social media companies when managing large volumes of data. The Global Internet Forum to Counter Terrorism, which brings together large US tech firms, has created a vast database of terrorist content, allowing companies to identify and remove content faster than otherwise possible.[175] More broadly, this forum offers an example of cooperation between social media and tech companies to share good practices, highlighting the value of collectively leveraging technology and collaborating to ensure that hateful extremist content is removed from online platforms in a timely manner.[176]

**National governments can lead the way in enforcing takedowns.** As explored in more detail in Section 4.1.4, Germany brought in its Network Enforcement Act, the 2018 NetzDG Act, which issues large fines to tech companies that do not remove hateful extremist content within 24 hours. Such measures can increase companies' responsiveness in removing hateful extremist content and false information from their platforms.

#### Literature findings on the limitations of takedowns

**Social media 'takedown rules' do not always ensure that false information is removed**. In relation to COVID-19, a recent study found that social media platforms had failed to remove 95 per cent of anti-vaccination misinformation reported to them.[177] A 2017 investigation by ProPublica, a US-based non-profit news organisation, analysed over 900 reported Facebook posts and reported that the social network giant applied its hate speech rules inconsistently, allowing content to slip through the net despite being reported by users.[178] Indeed, several social media companies have repeatedly failed to remove false information and hateful content, to the extent that consumer-goods giant Unilever recently removed its advertising from Facebook, Instagram and Twitter in the US, citing that 'in the areas of divisiveness and hate speech […] continuing to advertise on these platforms at this time would

[173] Polyakova and Fried (2019).
[174] Colliver and King (2020).
[175] Gibbs (2017).
[176] GIFCT (2020).
[177] CCDH & Restless Development (2020).
[178] Tobin et al. (2017).

not add value to people and society'.[179] Some 1,000 other companies joined the pledge led by Stop Hate for Profit, an initiative seeking to temporarily boycott advertising on Facebook due to hate and disinformation.[180]

**There has also been criticism that social media sites have historically applied takedown measures inconsistently**. Activists and journalists in several countries and disputed territories – such as the Palestinian territories, Kashmir and Crimea – have reported having their posts deleted, prompting criticism.[181] A ProPublica report analysed the rules used to train Facebook content reviewers, finding that there was a distinction between 'protected categories' and 'subsets'. Using an algorithm that viewed all ethnicities and genders equally, 'protected categories' – based on race, sex, gender identity, religious affiliation, nationality, ethnicity, sexual orientation and disability – include white men, as both characteristics are protected, but not female drivers or black children, who fall into the 'subset' group, as one of their traits is not protected.[182] Such criticism highlights the difficulty of taking down harmful content without curbing free speech. While human moderators alone cannot filter the sheer multitude of online posts, algorithms are likely to exhibit flaws as they are developed and refined. This is therefore an area in need of significant further research and investment.

### 4.1.4. Education

Educational approaches are designed to help build societal resilience to false information. These approaches constitute an important long-term measure and are targeted towards young people – a cohort who are particularly vulnerable to the influence of false information given the time they spend online consuming social media.[183]

#### Insights from the literature on recommended educational approaches

**Education is generally recognised as important in building societal resilience to false information** and in preventing radicalisation,[184] particularly among young people in schools.[185] In general, younger people are reportedly more likely to encounter false information and to accept its presence online.[186] Noting the heightened vulnerability of younger generations, education can prevent the spread of false information in several ways[187]:

- Education can promote historical awareness, values of citizenship and civic participation;
- Education can play a part in developing the critical thinking skills of young people, creating a 'safe' environment for discussing controversial topics and allowing for a range of views; and

[179] Hern (2020).
[180] Stop Hate for Profit (2020).
[181] Angwin et al. (2017).
[182] Angwin et al. (2017).
[183] Ofcom (2020).
[184] 'Radicalisation' refers to the process an individual goes through towards becoming involved in extremism. Department for Education (2017).
[185] Flynn et al. (2017).
[186] Ofcom (2020).
[187] Wallner (2020).

- The 'contact hypothesis' suggests that by facilitating contact between different population groups, education programmes can reduce prejudice between majority and minority groups.

'Critical literacy' measures the ability of individuals to assess the soundness and credibility of arguments, respond to arguments and come to conclusions by analysing information.[188] A 2018 report into fake news and critical literacy in British schools by the National Literacy Trust found that only 2 per cent of children and young people in the UK possessed the critical literacy skills to assess whether a news story is true or false.[189] Education in itself is not guaranteed to build critical thinking skills, but teachers can help by facilitating discussion on challenging or controversial topics, allowing for a plurality of opinions and revealing different perspectives.[190]

Finland offers an example of how countering false information has been incorporated into educational approaches. In Finland, educators integrate teaching on false information into all subject areas from a young age.[191] In art, history, maths and language lessons, students learn how messaging can be manipulated, by analysing how the meaning of images can be misconstrued, studying historical propaganda campaigns, learning about how statistics can be used to deceive, and being educated on how language is used to confuse and mislead.[192] Such initiatives are likely to build more robust societies, in which fake news does not find as ready a foothold.[193] Some researchers suggest that using a theatre format[194] to address sensitive and controversial issues can prove particularly effective, pointing to several studies highlighting the effectiveness of 'entertainment-education'[195] in increasing knowledge, generating favourable attitudes and influencing positive behaviours.[196]

Overall, education is a method for countering false information that can foster critical thinking and societal resilience, and combat the perceived legitimacy of political violence.[197] Furthermore, educational programmes can help shape how students think without policing their thoughts and beliefs.[198] As an intervention method for preventing the proliferation of false information, Wallner (2020) recommends further investment in education, as well as a move away from traditional learning styles, where the teacher imparts knowledge to students and the former's authority is not to be questioned, towards a more interactive model of learning.[199]

---

[188] Machete & Turpin (2020).
[189] National Literacy Trust (2018).
[190] Wallner (2020).
[191] Henley (2020).
[192] Henley (2020).
[193] In 2019, Finland topped the Media Literacy Index, a ranking of 35 countries where each country was allocated scores based on indicators relating to media freedom, levels of education, trust in others and use of the Internet. In contrast, the UK was ranked in 12th place (Lessenski 2019).
[194] A theatre format could see a speaker deliver a monologue that has been prepared with the help of a professional theatre company (Parker & Lindekilde 2020).
[195] 'Entertainment education' includes addressing sensitive topics in a play or musical (Parker & Lindekilde 2020).
[196] Parker & Lindekilde (2020).
[197] Parker & Lindekilde (2020).
[198] Wallner (2020).
[199] Wallner (2020).

### Literature findings on the limitations of educational approaches

The reviewed literature appears to be largely supportive of educational approaches, with relatively few limitations and challenges highlighted. It should nonetheless be noted that while education can promote historical awareness, values of citizenship and civic participation, such education programmes can be counterproductive where they deliver restrictive narratives relating to national identity and history, which could alienate parts of the population.[200] A further consideration to note when developing educational measures is that extremist narratives tend to target individuals with different levels of education in various ways – using monetary incentives and intellectual narratives to target (respectively) those from areas with lower quality and higher quality education.[201] By extension, it could be beneficial to tailor education programmes to these different cohorts in recognition that individuals are targeted by hateful extremists in different ways.[202]

## 4.2. Actors involved in the design and delivery of interventions

The literature sets out a number of reported 'good practices', challenges and considerations in relation to online interventions for countering false information and its use by hateful extremists. When exploring the 'good practices' and challenges identified in the reviewed papers, this section distinguishes between several actors involved in the design and delivery of online interventions:

- **National governments**
- **Social media companies**
- **Civil society organisations**
- **News/media organisations**

The counter-measure 'types' explored in Section 4.1 are implemented by different actors, often with overlaps. While roles are unlikely to be as clear-cut in practice, Table 4-1 provides an indicative overview of actors' roles in delivering the various measures. The measures outlined below are already being taken by governments, social media companies, civil society actors and media organisations, though the literature also contains recommendations to build on and improve such measures.

[200] Wallner (2020).
[201] Wallner (2020); Ghosh et al. (2016).
[202] It is not clear from the reviewed literature whether this is an approach that has been taken by educational institutions to date.

**Table 4-1: Actors and interventions**

| Actors | Factchecking | Counterspeech | Takedowns | Education |
|---|---|---|---|---|
| Governments | x | x | x | x |
| Social media companies | x | | x | |
| Civil society | x | x | | x |
| News/media organisations | x | | | x |

The sections below explore some of the reported 'good practices', challenges and considerations across different actor types as regards the implementation of interventions to counter false information and its use by hateful extremists, according to the reviewed source material.

### 4.2.1. National governments

Table 4-2 summarises some of the reported 'good practices', challenges and considerations for national governments in tackling the spread of hateful extremist false information, as set out in the reviewed papers.

**Table 4-2: National government: reported 'good practices', challenges and considerations**

| Reported 'good practices' | Challenges and considerations |
|---|---|
| • Taking action to hold social media companies to account.<br>• Investing in education and prevention programmes to reduce the spread of false information and to build societal resilience (see also Section 4.1.4).<br>• Speaking out against hateful narratives, and refraining from engaging in hate speech.<br>• Monitoring disinformation dynamics and their impacts on society. | • Relying on tech/social media companies to prioritise the wellbeing of their platform users over advertising revenues.<br>• A need to increase the quality and public availability of statistics on hateful extremism and hate crime.<br>• Budgetary challenges – particularly in the COVID-19 context – and a need for greater investment in education and awareness-raising programmes. |

National governments have promoted policy initiatives to hold social media companies to account, as the pursuit of advertising revenues could compromise the integrity of online content. In Germany, the 2018 NetzDG Act forces tech companies to remove hate speech from their platforms within 24 hours

of it being reported, to avoid a €20 million fine.[203] With this law in place, one in six Facebook moderators now operates from Germany to ensure that hate speech is removed in a timely way, which can be viewed as an indicator of success for the NetzDG Act.[204] Many large tech/social media platforms only provide headline statistics about the targets and perpetrators of online hate speech, showing only the breakdowns by country, while other websites such as Google and BBC News provide no such information.[205] In addition, each platform varies in its frameworks, guidelines, moderation processes and reported-content takedowns, meaning that there is a lack of consistency and comparability across platforms. For example, Facebook currently reports on the number of abusive posts while, Twitter logs the number of abusive users.[206] Vidgen et al. (2019a) encourage standardised reporting to allow for comparison across platforms, recommending that governments impose reporting requirements on large tech companies. However, such an initiative could be difficult to implement practically across all companies, as tech companies vary greatly in size.

UK organisation Article 19 calls for public officials to speak out against hateful narratives, and to refrain from engaging in hateful speech themselves.[207] Members of the 'No Hate Parliamentary Alliance' at the Council of Europe also pledge to raise awareness and take action against hatred and intolerance by exchanging information on best practices, engaging in campaigning activities and speaking out. In terms of educating public officials, further equality training to educate officials on countering discrimination and hate speech could be beneficial, particularly when such training is clearly communicated to the public in order to build trust and ensure transparency.[208]

The European Commission (2018) emphasises the need for governments to monitor disinformation dynamics and their impacts on society.[209] To this end, Vidgen et al. (2019a) recommend that governments consider collating statistics on the different types of illegal online abuse – including hate speech and online harassment – and publish these in a single bulletin.[210] The Equality and Human Rights Commission takes an active role in reporting hateful extremist narratives, collecting data on hate speech online and offline.[211] Every October, the Home Office issues a response to the annual Hate Crime statistics. These resources could provide a good starting point for a regular bulletin to provide an overview of online harm. Vidgen et al. (2019a) recommend further efforts to improve the coverage, comparability (across years as well as with other countries) and quality of government statistics, which could include the reinstatement of the UK Home Office's reporting of online hate crime.[212] In the Home Office's 2018/19 hate crime report, concerns about the quality of statistics meant that no figures were provided for online hate.[213] In general, there is a need for continued

[203] Digital, Culture, media and Sports Committee (2019).
[204] Digital, Culture, media and Sports Committee (2019).
[205] Vidgen et al. (2019a).
[206] Vidgen et al. (2019a).
[207] Article 19 (2018).
[208] Article 19 (2018).
[209] European Commission (2018).
[210] Vidgen et al. (2019a).
[211] Article 19 (2018).
[212] Vidgen et al. (2019a).
[213] Vidgen et al. (2019a).

support for research and real world experiments, in addition to cooperation with civil society organisations to establish open data standards.[214]

Furthermore, according to Vidgen et al. (2019a) a publicly accessible monitoring platform could be established to provide real-time insight into online abuse; and researchers should leverage recent computational advances including machine learning models, deep neural networks,[215] and contextual word embeddings.[216] Hatemeter, an EU-funded initiative, uses natural language processing, machine learning and big data analytics to monitor and analyse anti-Muslim content from social media, flagging hate speech in real time.[217] This system was established to understand the patterns in online Islamophobia, develop effective strategic and tactical response plans, and produce a counterspeech framework for tackling anti-Muslim hatred.[218] Government funding for such initiatives could provide further insights for researchers and practitioners alike.

Bilazarian (2020) highlights a need for greater public sector investment in education and training to raise awareness and develop societal resilience.[219] Article 19 (2018) recommends increased investment in digital literacy skills to enable the public to understand the benefits of digital engagement, particularly the opportunities to foster pluralism, which could encourage people to engage in counterspeech.[220] As noted in Section 4.1.4, education on false information is integrated into subject areas including art, history, maths and language lessons in Finnish schools.[221] Such initiatives can help increase critical literacy and raise awareness of the impact of false information.

Beyond training and education, reviewed papers suggest that governments can cooperate with civil society to develop counterspeech tactics to educate bystanders while combating online false information (see also Section 4.1.2). For example, the US government response to al-Qaeda-inspired online propaganda – referred to as 'Think Again' – saw State Department officials use social media to spread counternarratives, directly challenging the portrayal of ISIS published by its supporters.[222] Some of the reviewed literature explores the question of whether such narratives hold greater sway if they are shared by official government accounts rather than by independent organisations, but the review did not identify a strong, empirical body of literature to provide conclusive evidence on this topic. The strengths and shortcomings of counterspeech are discussed in Section 4.1.2.

[214] Graves (2018).
[215] 'Deep neural networks' are artificial deep learning systems that allow the recognition of patterns and the sophisticated interpretation of data.
[216] Vidgen et al. (2019a); 'contextual word embeddings' allow systems to independently encode words from a document.
[217] Hatemeter (2020).
[218] Hatemeter (2020).
[219] Bilazarian (2020).
[220] Article 19 (2018).
[221] Henley (2020).
[222] Bilazarian (2020).

### 4.2.2. Social media companies

Table 4-3 summarises the reported 'good practices', challenges and considerations for social media companies in relation to countering hateful extremist false information as articulated in the reviewed literature.

**Table 4-3: Social media companies: reported 'good practices', challenges and considerations**

| Reported 'good practices' | Challenges and considerations |
|---|---|
| • Adopting the use of upvote/downvote systems that potentially reduce the visibility of misleading or hateful content.<br>• Using 'good' bots to share positive narratives on social media and online platforms.<br>• Taking down abusive posts and comments (see Section 4.1.3).<br>• Factchecking content, including that published by public officials (see Section 4.1.1). | • Oversights in 'recommendation' algorithms can direct users to harmful online content.<br>• A need to provide transparency regarding any algorithm changes.<br>• A need for faster and stronger action against malicious bots. |

False information poses a significant challenge for social media companies.[223] Nonetheless, there is some consensus within the reviewed literature that companies should take greater responsibility for the content hosted on their platforms.[224] To this end, Carter (2020) suggests that social media companies should monitor content and change algorithms to avoid the 'recommendation' of extreme or harmful content on their platforms.[225] Notably, Ribeiro et al. (2019)'s study suggested the existence of a 'radicalisation pipeline' on YouTube, positing that the recommendation algorithm was pointing users to increasingly extreme content.[226] The Council of Europe (2017) suggests that any changes to algorithms that could down-rank content should be accompanied by transparent criteria to avoid claims of bias and censorship from content producers.[227] Jones (2020) also advocates for the use of algorithmic 'throttling', or a system of down-ranking of misleading or hateful content.[228] Reddit, for example, uses an upvote/downvote system, which encourages users to moderate forums by raising (and lowering) the profile of certain content. However, it remains to be seen whether an upvote/downvote system would reduce the visibility of false information on a larger platform such as Facebook. In addition, upvoting/downvoting approaches evidently require careful policing – these systems have been manipulated on forum boards such as 4Chan and 8Chan by right-wing extremists to upvote extreme content, while downvoting 'tamer' content that is subsequently removed after 24

[223] Ariza (2020).
[224] CoE (2017a); Caplan (2017); Ockenden (2020); Digital, Culture, media and Sports Committee (2019).
[225] Carter (2020).
[226] Ribeiro et al. (2019).
[227] CoE (2017a).
[228] Jones (2020).

hours. A further issue to consider is the potential risk for automated manipulation if the visibility of content is determined by consensus voting.

According to the Council of Europe (2017), social media companies need to take stronger and faster action against automated accounts (bots) that amplify content.[229] During the 2016 US election, for example, Twitter reported that nearly 1.4 million human accounts interacted with content created by bots or trolls.[230] Bots are used by hate actors to rapidly amplify and multiply false information, allowing harmful narratives to have a wider reach. Given that bots have such an effect on hate speech, Albadi et al. (2019) ask whether 'good' bots could be used to promote tolerance, acceptance and diversity values, in order to decrease online hate speech.[231] To this end, in 2018 the Fundamental Rights Agency of the European Union unveiled its pilot project 'FRAbot', an automated Twitter account that responds to hateful extremist content online.[232] However, it remains to be seen whether such an initiative could be rolled out on a large scale, and whether this could be implemented by social media platforms themselves. Overall, Carter (2020) recommends that social media companies take a more active role in promoting positive counternarratives and ensuring that these are actively shared by online influencers.[233]

### 4.2.3. Civil society

Table 4-4 summarises reported 'good practices', challenges and considerations for civil society organisations with regard to countering false information and its use by hateful extremists, according to the reviewed literature.

[229] CoE (2017a).
[230] Albadi et al. (2019).
[231] Albadi et al. (2019).
[232] Fundamental Rights Agency (2018).
[233] Carter (2020).

**Table 4-4: Civil society: reported 'good practices', challenges and considerations**

| Reported 'good practices' | Challenges and considerations |
|---|---|
| • Acting as an 'honest broker' to bring together different actors to generate discussion and foster collaborative solutions against false information.<br>• Providing direct support to victims of hateful extremism (e.g. via hotlines).<br>• Raising awareness and petitioning for the removal of false information.<br>• Providing educational material and educating the public. | • Lack of awareness among companies regarding the placement of advertisements (e.g. on disinformation sites) and a need for greater transparency.<br>• A need for more resources and funding to tackle false information on a greater scale. |

**Civil society organisations can act as 'honest brokers', bringing together different actors in the fight against false information**, and creating fora for engagement between social media companies, news organisations, research institutes and governments.[234] A number of civil society organisations are also playing an active role in supporting victims of hateful extremism; for example, Stop Hate UK offers a hotline for victims who wish to seek advice and guidance. Furthermore, these organisations actively report online abuse, highlight areas for the attention of governments and social media companies, and raise public awareness of disinformation. For example, the UK-based organisation Stop Funding Fake News has pointed out that companies advertising via Google are often unaware of the placement of their advertisements, and has called for greater transparency from Google. Stop Funding Fake News aims to raise awareness of this issue and petition companies to remove their advertising from websites that it deems as hosting disinformation or stories that are inaccurate and sensationalist.[235] These petitions have resulted in action by large brands such as Sky, eBay and WWF to remove their advertising from certain sites.[236]

**There is a need to educate the public on an ongoing basis about the threat of disinformation and the persuasive techniques used by hateful extremist actors**.[237] Commentators have identified a need to raise awareness about the risks of disinformation to society, including growing public distrust in official sources and the deepening of societal divisions.[238] Disinformation can cast doubt on government communications, scientific consensus and historical facts.[239] This can be particularly detrimental at a time when there is a global health pandemic, with the spread of disinformation sparking concerns

[234] CoE (2017a).
[235] Stop Funding Fake News (2020).
[236] Stop Funding Fake News (2020).
[237] CoE (2017a).
[238] CoE (2017a).
[239] Connolly et al. (2019).

regarding societal divisions in countries across the globe.[240] Ultimately, it is the goal of some hateful extremist actors to bring about the collapse of society so that they can build a 'new world' based on their belief system.[241] In the context of this threat, it is particularly important to educate the public about the dangers of disinformation.

According to the European Commission, enhancing education and public awareness is a role that could be taken on by civil society.[242] Already, civil society fills in the gaps left by government educational initiatives (Section 4.2.1), providing educational material to schools and educators to build societal resilience, improve detection of racism and antisemitism, and enable students to critically assess how they consume online media.[243] As civil society actors are independent, they could also be seen as relatively non-political or 'more trustworthy' by some members of the public.[244] To inform educational programmes and provide further insight, some commentators suggest that civil society and/or government could administer a dedicated annual representative survey to understand the online abuse experienced by people in the UK.[245]

### 4.2.4. News/media organisations

Table 4-5 summarises reported 'good practices', challenges and considerations for news/media organisations, according to the reviewed literature.

**Table 4-5: News/media organisations: reported 'good practices', challenges and considerations**

| Reported 'good practices' | Challenges and considerations |
|---|---|
| • Good journalism practices (e.g. avoiding clickbait) to reduce the spread of false information.<br>• Demonstrating integrity in calling out misleading statements from prominent public figures and other sources.<br>• Maintaining transparency in factchecking. | • Reporting on news issues when increased exposure feeds into extremist agendas.<br>• Avoiding clickbait and misleading headlines.<br>• Ensuring that quality control is sufficient and avoiding reliance on post-publication correction. |

It is important for news/media organisations to take responsibility for headlines and to avoid 'clickbait', particularly amid a global pandemic.[246] Outrage attracts readers and generates revenue; and news publications accordingly post clickbait headlines containing sensationalist and misleading information in order to increase views.[247] Furthermore, many news sources similarly report on cases

[240] Banjo & Lung (2019).
[241] Wilson (2020).
[242] EC (2018b).
[243] Vidgen et al. (2019a).
[244] CoE (2017a).
[245] Vidgen et al. (2019a). OxIS and Ofcom's 'Adult Media Use and Attitudes' survey offers a good starting point.
[246] CoE (2017a).
[247] Marwick and Lewis (2017).

of false information they consider newsworthy, unintentionally giving such content more exposure.[248] For this reason, the Council of Europe (2017) recommends that news and media organisations should consider cooperating to agree on when to hold 'strategic silence'.[249] In addition, the Council recommends that news organisations should ensure that they pursue good journalism practices and ensure internal quality control.[250] As many organisations rely on outrage, sensationalism or clickbait to generate revenue, there is a need to carefully establish good journalism practices and determine how these can be adhered to without incurring losses.

In light of the difficulties outlined in Section 4.1.1 regarding persuading readers of corrections, it is essential that external post-publication corrections do not substitute internal processes of quality control.[251] Furthermore, despite running the risk of alienating supporters, Brennen et al. (2020) suggest that news/media organisations should take a more active role in calling out false information from prominent politicians.[252] Finally, Amazeen et al. (2018), Thorson (2016) and Douglas et al. (2019a) suggest that news organisations should maintain transparency in how they factcheck, using credible sources and data, and making them accessible to audiences.[253]

## 4.3. Evidence gaps

The reviewed papers identified two key gaps in relation to the second review theme: (1) a lack of research on the impact of existing interventions and empirical evidence on 'what works'; and (2) a lack of varied datasets used in primary data analysis regarding hateful extremist activity online and associated responses.

### Lack of evidence on the impact of existing interventions and 'what works'

With regards to research on the impact of interventions, Amazeen et al. (2018) and Thorson (2016) emphasise the need for a better understanding of the effectiveness of factchecking, counterspeech, takedowns and other such techniques used to counter false information. This is echoed by Douglas et al. (2019a), who call for researchers to evaluate these techniques to address false information and their potential effects in relation to the spread of conspiracy theories.[254] Similarly, research focusing on hate speech highlights current gaps in understanding effective ways to tackle this challenge. With a focus on the backlash caused by some counterspeech campaigns against hate speech, Baldauf et al. (2019) call for research into innovative alternatives, echoed by Muller and Schwarz (2020), who highlight the need to demonstrate effective techniques against hate speech.[255] Finally, studies by Parker and Lindekilde (2020) and by Wallner (2020) highlight the need for robust

[248] Marwick and Lewis (2017).
[249] CoE (2017a). Further research is needed to establish what the threshold for a 'strategic silence' could be.
[250] CoE (2017a).
[251] Cherilyn and Posetti (2018).
[252] Brennen et al. (2020).
[253] Graves (2013); Poynter (2019); Humprecht (2019); Douglas et al. (2019a).
[254] Amazeen et al. (2018); Thorson (2016); Douglas et al. (2019a).
[255] Baldauf et al. (2019); Mueller and Schwarz (2020).

impact assessments and public evaluations of existing PVE and CE/CVE[256] interventions to inform future policy and interventions.[257]

### Limited variation in datasets used in primary data analysis

Several studies point to the need for more varied datasets to analyse false information, hateful extremist content online and associated interventions. This lack of diversity concerns both the format of content as well as its origins. Regarding format, several studies highlight the need for further analysis of multimodal content: Alan Turing Institute (2019) points to the 'severe restriction' in terms of the lack of research into non-text based abusive content online, including the risk of neglecting other forms of content like images, audio files, memes, GIFs or videos in a multimedia environment.[258] This is echoed in a study by Ziems et al. (2020) on racism online, and in Schwarz and Holnburger (2019)'s study on the dissemination of disinformation online.[259] Studies also highlight the need for an extended geographical and linguistic scope of research. Regarding conspiracy theories, Nyhan and Zeitsoff (2018) point out that most research focuses on Europe and the US, with little focus on the developing world. This reflects our study finding regarding the Eurocentric focus of the reviewed literature on false information and hateful extremism online (see Section 2.1.1). On online abuse, Vidgen et al. (2019) also note the lack of focus on non-English language abuse.[260] Research in this area could be important in England and Wales, particularly among communities where English is not the first language. Finally, Albadi et al. (2019) comment on the lack of research on bot behaviour on Arabic social media.[261] Research into this area could potentially shed light on immigrant communities based in the UK and their experiences of exposure to hateful extremist disinformation. This narrow understanding of hateful content online – which does not appear to cover multimodal hateful online content across countries and languages – could limit the ability of governments, social media companies and other actors to develop effective interventions and policy responses.

[256] PVE: Preventing Violent Extremism; CVE: Countering Violent Extremism.
[257] Parker and Lindekilde (2020); Wallner (2020).
[258] Alan Turing Institute (2019).
[259] Ziems et al. (2020); Schwarz and Holnburger (2019).
[260] Alan Turing Institute (2019).
[261] Albadi et al. (2019).

# 5. Key findings and next steps

This chapter presents an overview of the main study findings in order to address the research questions set out in Chapter 1 and presented in the box below (Section 5.1). Based on these findings, Section 5.2 then sets out policy insights for the consideration of CCE. Finally, Section 5.3 outlines areas for recommended future research to help address the evidence gaps identified in Chapters 3 and 4.

**Review theme 1: Links between hateful extremism and false information**

- 1.1: What impact can false information have on hateful extremist beliefs and behaviours?
- 1.2: In what ways do hateful extremist beliefs contribute to the spread of false information?
- 1.3: What trends and variations can be identified across different audience types, modes of false information and extremist groups?

**Review theme 2: Associated online interventions and policy responses**

- 2.1: What insights can be identified from the literature on the effectiveness of existing interventions and policy responses?
- 2.2: What recommendations are put forward in the existing literature in relation to future interventions in this area?
- 2.3: What transferrable lessons/'good practices' from successful interventions in related policy areas can be identified?

## 5.1. Summary of findings

**Review theme 1: Links between hateful extremism and false information**

The literature review draws on a combination of expert opinion and empirical analysis by the authors of selected sources. The findings presented here provide a number of plausible hypotheses identified in the literature, but the quality and quantity of the literature is not sufficient to provide strong, empirical evidence of the links.

False information can enable the spread of hateful extremist attitudes and beliefs, as extremist narratives circulate online and pick up traction among those who might not usually consume such content. This report has also identified how false information can lead to the emergence of echo chambers, which can strengthen existing hateful attitudes by further desensitising group participants

to hateful language and narratives. In this process, moderate users drop off, leaving a concentration of extremists and no opposing views to challenge their perspectives.

Hateful extremist beliefs can contribute to false information by leveraging such narratives to serve their causes. Hateful extremist narratives are used as a recruitment tool and claim to offer a 'red pill'[262] to people struggling from a lack of prospects in light of the current COVID-19 crisis, in an effort to make the extremist narratives seem more appealing. Furthermore, false information gives hateful groups increased exposure – including through the statements of public officials and mainstream media reporting – and can attract sympathy for their causes. In combination, recruitment benefits and increased exposure present considerable incentives for such groups to spread their narratives.

Many different types of hateful extremist actor operate within the online domain. While these actors have made use of the current pandemic to further their interests, they have taken different approaches in doing so. While far-right groups seek to blame migration, globalisation or the government for the virus, Islamist actors might see the pandemic as a form of divine punishment against infidels. However, all groups seek to direct hostile narratives at 'out-groups', leveraging public fear and uncertainty amid the global pandemic. As more people – particularly young people – have consumed online content during lockdown, this has exposed a greater cross-section of the population to recruitment by hateful extremist groups.

**Review theme 2: Associated online interventions and policy responses**

According to the reviewed source material, several existing measures offer promise in terms of reducing the spread of false information and building societal resilience. Civil society is playing an important part in raising awareness and campaigning to prevent false information. However, the review did not identify any interventions that have been subject to rigorous empirical evaluation. The UK-based organisation Stop Funding Fake News, for example, has taken on the task of preventing disinformation sites from earning advertising revenues. In terms of policy interventions, the 2018 German NetzDG Act has shown promise in holding social media platforms to account and offering a radical solution to stem the spread of false information. News/media organisations support the fight against hateful extremism by adhering to good journalistic practices (e.g. avoiding clickbait), while social media companies can contribute by modifying their algorithms to prevent the 'recommendation' of harmful content.

COVID-19 presents a unique challenge for policymakers and organisations seeking to tackle false information. In terms of the role of governments, the reviewed sources highlight a need to dedicate more resources to combat false information in order to build societal resilience, and to support further research into the impacts of hateful extremist narratives. Social media companies and news/media organisations are also urged to take more responsibility by the authors of reviewed papers,

[262] The 'red pill' is a concept coined by The Matrix, a science-fiction action film franchise, wherein the underlying, unpleasant truths of the world are revealed to the consumer, rather than the 'blue pill', which allows the majority of consumers to maintain 'blissful ignorance'. Baldauf et al. (2019).

respectively by managing the content on their platforms and ensuring that outlets adhere to good journalism practices and avoid clickbait headlines.

This report set out to address all research questions presented in Section 1.2, including Q2.3: '*What transferrable lessons/good practices from successful interventions in related policy areas can be identified?*'. While this report initially set out to answer this question, the study team did not identify successful interventions from alternative policy areas, as this area was not revealed in the reviewed papers. While one source referred to gang violence, the reference was made solely to distinguish this threat from extremism. The paper observed a link between criminal biker gangs and right-wing hate groups, but noted that the former lacks an ideology to legitimise their acts of violence, limiting the transferability of insights from this group to right-wing hate groups.[263] Beyond this source, there was limited reference to other policy areas in the reviewed papers. It should nonetheless be noted that the range of approaches presented in this report are drawn from a wide range of disciplines, including psychology, political science, sociology and law, which offers a range of views and approaches. Although the review did not identify direct lessons or transferrable practices from other policy areas, insights on interventions in this report benefit from a multidisciplinary approach across a broad scope of literature.

## 5.2. UK policy considerations

Based on the review findings of this report, we present some policy insights for the consideration of CCE:

- **Investing in some of the notable research gaps in the field of false information**. As identified in the previous section, there are significant research gaps requiring further input. There is a need for further evaluation of existing interventions, as well as research on directional motivations and a wider range of studies in terms of geography, languages and online content. Regarding appropriate responses to false information, there is a need to develop a better understanding of the issue and its wider context. Furthermore, there is a need for additional research into 'what works', to understand intervention effects.

- **Holding tech companies to account can increase their responsiveness to false information**. As the implementation of the 2018 NetzDG Act shows (see Section 4.2.1), large fines may incentivise tech companies to remove hateful extremist content in a timely manner, increasing companies' responsiveness in removing false information from their platforms.

- **Investing in education programmes can help raise awareness of the dangers of hateful extremism**. As noted in Section 4.2, there is a role for government and civil society to play in developing education and training to increase public awareness of hateful extremist uses of false information. Given the large volume of false information online and the increased exposure of online users to this content during COVID-19, there is a pressing need to educate the public

[263] Baldauf et al. (2019).

about the threat of false information, persuasive techniques used by hateful extremist actors, and actions to support individual resilience.

- **Collecting and publishing information regarding indicators of hateful extremism could help improve policy responses**. The literature highlights a need for governments to collate and publish information on hateful extremism. As outlined in Section 5.2, there is a need to broaden the type of information collected (e.g. looking beyond text-based content to include images, audio, memes and other content), make greater use of computational advances (e.g. machine learning), and to ensure the quality of statistics (e.g. via independent peer review). A better understanding of the nature and scale of the threat could help enhance policy measures and improve public resilience.

- **Exploring the use of 'good' bots to support the spread of positive narratives online**. It is evident from online activity during major political events (e.g. the 2016 US presidential election) that trolls and bots have manipulated voter behaviour and deepened societal divides. Noting the persuasiveness of these tactics, there could be scope to explore the adaptation of such techniques to instead promote democratic values of tolerance, acceptance and diversity on social media, as well as to constrain the reach and influence of online hate speech.

- **Collaborating across sectors can ensure that interventions are mutually reinforcing**. Engagement between UK policy officials, social media moderators, educators, journalists, civil society organisers, research experts, legislators and other national governments could help ensure that HMG policy and guidance reflects an understanding of the scale and nature of the challenge from hateful extremists' use of false information, and complements activities that are being undertaken elsewhere.

## 5.3. Avenues for further research

In light of the evidence gaps identified in the reviewed sources (see Sections 3.3 and 4.3), we propose several areas for future research:

- **Independent and robustly designed evaluations of existing interventions and 'what works'**. To inform a better understanding of the effectiveness of existing counter-measures, there would be merit in conducting independent evaluations of interventions dedicated to tackling hateful extremism and false information. Public authorities should consider this need when designing or commissioning interventions and collaborating with private sector platforms. Evaluations could focus on measuring the effectiveness of factchecking and other techniques used to counter false information – such as counterspeech and takedowns – and assess interventions that are delivered online and those implemented in physical environments, such as schools or universities. The results of these evaluations should be made publicly available to help inform the development of innovative and effective future interventions.

- **Future research on 'directional motivations'**. By focusing on 'directional motivations' – i.e. the individual's motivation to hold on to existing convictions and attitudes – future studies could

help improve understanding of the characteristics of individuals who are more prone to hateful extremist beliefs and behaviours. This could also help enhance awareness of the way these individuals respond to online content and false information. Research on this topic could support the development of new interventions, informing an understanding of the circumstances under which false information corrections will be more (or less) effective.

- **Studies with a wider reach in relation to geography, languages and online content**. To equip policy and decision makers with a fuller evidence base, future studies should analyse a wider range of: (i) countries and regions, shifting away from the Euro- and US-centric focus of the literature to incorporate a broader cross-section of countries and regions across the globe;[264] (ii) languages, in order to understand non-English language hate speech and misinformation, particularly among immigrant communities in England and Wales; and (iii) types of online content, moving beyond text-based content to other types of content including images, audio files, memes, GIFs or videos. Researchers could rely to a greater extent on automated processes and machine learning approaches to analyse large volumes of online content, allowing for larger and more varied datasets. Vidgen et al. (2019a) also observe that researchers have made little use of freely available Google Trends data, with scope for this data to inform future research.[265]

[264] While this report has been written to inform CVE policy in England and Wales, our review of the literature revealed that there is a shortage of empirical studies with a focus on geographical areas beyond Europe and the US.
[265] Vidgen et al. (2019a).

# References


Abbink, Klaus & Harris, Donna. 2019. 'In-group favouritism and out-group discrimination in naturally occurring groups'. 4 September 2019. As of 19 August 2020: https://doi.org/10.1371/journal.pone.0221616

Alan Turing Institute. 2019. 'Hate speech: measures and counter-measures'. As of 8 October 2020: https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures

Albadi, N., Kurdi, M. & Mishra, S. 2019. 'Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media'. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 61. As of 9 October 2020: https://doi.org/10.1145/3359163

Allington, W. 2020. 'Antisemitic Memes and Naïve Teens: Qualitative and Quantitative Impacts of the Internet on Antisemitism, the Evolution of Antisemitism 2.0, and Developing Adaptable Research Methodologies into Online Hate, Abuse, and Misinformation'. As of 8 October 2020: https://hdl.handle.net/2123/22444

Amazeen, Michelle A., Emily Thorson, Ashley Muddiman & Lucas Graves. 2018. 'Correcting political and consumer misperceptions: the effectiveness and effects of rating scale versus contextual correction formats'. *Journalism & Mass Communication Quarterly* (95:1).

Angwin, J., ProPublica & H. Grassegger. 2017. 'Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children'. ProPublica. 28 June 2017. As of 14 September 2020: https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

Antisemitism Policy Trust. 2020. 'Antisemitism and the Conservative Party'. As of 20 August 2020: https://antisemitism.org.uk/wp-content/uploads/2020/06/Antisemitism-and-the-Conservative-Party-4-11-2019-V5.pdf

Archetti, C. 2015. 'Terrorism, Communication and New Media: Explaining Radicalization in the Digital Age'. *Perspectives on Terrorism* 9(1). February 2015. 49–59. As of 14 September 2020: https://www.jstor.org/stable/pdf/26297326.pdf

Ariza, C. 2020. 'From the Fringes to the Forefront: How Far-right Movements across the globe have reacted to COVID-19'. Tony Blair Institute. As of 8 October 2020: https://institute.global/policy/fringes-forefront-how-far-right-movements-across-globe-have-reacted-covid-19

Article 19. 2018. 'Responding to "hate speech" with positive measures: A case study from six EU countries'. As of 15 September 2020: https://www.article19.org/wp-content/uploads/2018/06/Responding-to-%E2%80%98hate-speech%E2%80%99-with-positive-measures-A-case-study-from-six-EU-countries-.pdf

Avis, W. 2020. 'The COVID-19 Pandemic and Response on Violent Extremist Recruitment and Radicalisation'. As of 8 October 2020: https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/15322/808_COVID19%20_and_Violent_Extremism.pdf?sequence=1&isAllowed=y

Ayad, M. 2020. 'The Propaganda Pipeline: The ISIS Fuouaris Upload Network on Facebook'. ISD. As of 14 August 2020: https://www.isdglobal.org/wp-content/uploads/2020/07/The-Propaganda-Pipeline-1.pdf

Baldauf, J., Ebner, J. & Guhl, J. 2019. 'Hate Speech and Radicalisation Online: The OCCI Research Report'. ISD.

Ball, P. & Maxmen, A. 2020. 'The epic battle against coronavirus misinformation and conspiracy theories'. *Nature*, 581(7809), 371–374. As of 8 October 2020: https://doi.org/10.1038/d41586-020-01452-z

Banjo, Shelly & Lung, Natalie. 2019. 'Fake news and rumours deepen distrust, divide as protests roil Hong Kong'. *The Print*. 12 November 2019. As of 29 September 2020: https://theprint.in/world/fake-news-and-rumours-deepen-distrust-divide-as-protests-roil-hong-kong/319643/

Barclay, J. 2011. 'Strategy to Reach, Empower, and Educate Teenagers (STREET): A Case Study in Government Community Partnership and Direct Intervention to Counter Violent Extremism'. London, UK: Center on Global Counterterrorism Cooperation.

Barderi, D. 2018. 'Antirumours Handbook, Council of Europe'. As of 8 October 2020: https://rm.coe.int/anti-rumours-handbook-a-standardised-methodology-for-cities-2018-/168077351c

Barrera, O., Guriev, S., Henry, E. & Zhuravskaya, E. 2020. 'Facts, alternative facts, and fact checking in times of post-truth politics'. *Journal of Public Economics* 182. February 2020. As of 14 September 2020: https://www.sciencedirect.com/science/article/pii/S0047272719301859

Bartoš, V., Bauer, M., Cahlíková, J. & Chytilová, J. 2020. 'Covid-19 Crisis Fuels Hostility against Foreigners'. As of 8 October 2020: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3618833

British Broadcasting Corporation (BBC). 2019. 'Jeremy Corbyn and Labour's anti-Semitism row explained'. BBC News. 27 November 2019. As of 20 August 2020: https://www.bbc.com/news/newsbeat-43893791

Benesch, S., Buerger, C., Glavinic, T. & Manion, S. 2020. 'Dangerous Speech: A Practical Guide'. Dangerous Speech Project. As of 8 October 2020: https://dangerousspeech.org/guide/

Bilazarian, T. 2020. 'Countering Violent Extremist Narratives Online: Lessons From Offline Countering Violent Extremism'. *Policy & Internet* 12 (1): 46–65.

Bolsen, T. & Druckman, J. 2018. 'Validating conspiracy beliefs and effectively communicating scientific consensus'. *Weather, Climate and Society* (10).

Brandtzaeg, P. B. & A. Følstad. 2017. 'Trust and Distrust in Online Fact-Checking Services'. Communications of the ACM 60 (9). 65–71.

Brennen, J.S., F. Simon, P.N. Howard & R.K. Nielsen. 2020. 'Types, sources, and claims of COVID-19 misinformation'. Reuters & University of Oxford. As of 8 October 2020: https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation

Briant, E. 2018. 'Building a stronger and more secure democracy in a digital age'. Written Evidence FKN0071 on Fake News submitted to the Digital, Culture, Media and Sport Committee.

Cabanatuan, M. 2020. 'Coronavirus: Asian American groups compile hate crime reports as Trump persists in "Chinese virus" attacks'. San Francisco Chronicle. As of 12 August 2020: https://www.sfchronicle.com/bayarea/article/Coronavirus-Asian-American-groups-compile-hate-15144295.php

Campbell, J. 2020. 'How Jihadi Groups in Africa Will Exploit COVID-19'. Council on Foreign Relations. As of 19 October 2020: https://www.cfr.org/blog/how-jihadi-groups-africa-will-exploit-covid-19

Carlo, A. 2019. 'The far right paint Muslims as the enemy of the LGBT+ community – but they are the real danger'. *The Independent*. 30 March 2019. As of 14 September 2020: https://www.independent.co.uk/voices/far-right-lgbt-muslims-christchurch-shooter-salvini-le-pen-a8846031.html

Carter, R. 2020. 'Young People in the Time of COVID-19: A Fear and Hope Study of 16–24 Year Olds'. Hope Not Hate. As of 8 October 2020: https://www.hopenothate.org.uk/wp-content/uploads/2020/08/youth-fear-and-hope-2020-07-v2final.pdf

Center for Countering Digital Hate (CCDH). 2020. 'Dealing with Hate & Misinformation around COVID-19'. CCDH. As of 8 October 2020: https://www.counterhate.co.uk/our-response

Center for Countering Digital Hate (CCDH) & Restless Development. 2020. 'Failure to Act: How Tech Giants Continue to Defy Calls to Rein in Vaccine Misinformation'. As of 30 September 2020: https://252f2edd-1c8b-49f5-9bb2-cb57bb47e4ba.filesusr.com/ugd/f4d9b9_8d23c70f0a014b3c9e2cfc334d4472dc.pdf

Commission for Countering Extremism (CCE). 2019. 'Challenging Hateful Extremism'. October 2019. As of 13 August 2020: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/874101/200320_Challenging_Hateful_Extremism.pdf

———. 2020. 'COVID: How hateful extremists are exploiting the Pandemic'. As of 8 October 2020: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/898925/CCE_Briefing_Note_001.pdf

Chaudhry, I., and A. Gruzd. 2020. 'Expressing and Challenging Racist Discourse on Facebook: How Social Media Weaken the "Spiral of Silence" Theory'. *Policy & Internet* 12 (1): 88–108.

Cherilyn, I., & Posetti, J. 2018. 'Journalism, "Fake News" & Disinformation: Handbook for Journalism Education and Training'. UNESCO.

Cobain, I., Parveen, N. & Taylor, M. 2016. 'The slow-burning hatred that led Thomas Mair to murder Jo Cox'. *The Guardian*. 23 November 2016. As of 15 September 2020: https://www.theguardian.com/uk-news/2016/nov/23/thomas-mair-slow-burning-hatred-led-to-jo-cox-murder

Council of Europe (CoE). 2017a. 'Information Disorder: Toward and interdisciplinary framework for research and policy making'. As of 8 October 2020: https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77

———. 2017b. 'WE CAN! Taking Action against Hate Speech through Counter and Alternative Narratives'. As of 30 September 2020: https://rm.coe.int/wecan-eng-final-23052017-web/168071ba08

Cohen, Sarah, Li, C., Yang, J. & Yu, C. 2011. 'Computational Journalism: a call to arms to database researchers'. Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR 2011) California. 148–151.

Colborne, M. 2020a. 'For the Far Right, the COVID-19 crisis is a PR opportunity'. *Fair Observer*. 13 April 2020. As of 16 August 2020: https://www.fairobserver.com/region/europe/michael-colborne-far-right-coronavirus-pandemic-assistance-covid-19-crisis-pr-news-10109/

———. 2020b. ‘As world struggles to stop deaths, far right celebrates COVID-19’. Al Jazeera. 26 March 2020. As of 15 September 2020: https://www.aljazeera.com/indepth/features/world-struggles-stop-deaths-celebrates-covid-19-200326165545387.html

Colliver, C. & King, Jennie. 2020. ‘The first 100 days: Coronavirus and Crisis Management on Social Media Platforms’. ISD. As of 8 October 2020: https://www.isdglobal.org/wp-content/uploads/2020/06/20200515-ISDG-100-days-Briefing-V5.pdf

Connolly, J.M., Uscinski, J.E., Klofstad, C.A. & West, J.P. 2019. ‘Communicating to the Public in the Era of Conspiracy Theory’. *Public Integrity* 21(5): 469–476. doi: 10.1080/10999922.2019.1603045

Community Security Trust (CST). 2020. ‘Coronavirus and the plague of Antisemitism’. Research Briefing, CST.

Crown Prosecution Service (CPS). 2019. ‘Terrorism: Guidance in relation to the prosecution of individuals involved in terrorism overseas’. As of 19 October 2020: https://www.cps.gov.uk/legal-guidance/terrorism-guidance-relation-prosecution-individuals-involved-terrorism-overseas

Davey, J., Hart, M. & Guerin, C. 2020. ‘An Online Environmental Scan of Right-wing Extremism in Canada: Interim Report’.

Dearden, L. 2018. ‘Far-right “yellow vest” Brexiteers chase Anna Soubry shouting “traitor” and “Hitler”’. *The Independent*. 20 December 2018. As of 15 September 2020: https://www.independent.co.uk/news/uk/politics/anna-soubry-brexit-leave-traitor-hitler-peoples-vote-threat-a8692261.html

Department for Education. 2017. ‘Safeguarding and radicalisation’. As of 19 October 2020: https://www.gov.uk/government/publications/safeguarding-and-radicalisation

DePaula, N., Fietkiewicz, K. J., Froehlich, T. J., Million, A. J., Dorsch, I. & Ilhan, A. 2018. ‘Challenges for social media: Misinformation, free speech, civic engagement, and data regulations’. *Proceedings of the Association for Information Science and Technology* 55(1): 665–668. doi:https://doi.org/10.1002/pra2.2018.14505501076

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A. & Larson, H. 2020. ‘The pandemic of social media panic travels faster than the COVID-19 outbreak’. *Journal of Travel Medicine* 27(3). doi: https://doi.org/10.1093/jtm/taaa031

Der III. Weg. 2020. ‘Diskussionsbeitrag zur Corona-Krise: Notwendiges Vorgehen oder Ausbau der Diktatur?’ 23 March 2020. As of 3 December 2020: https://der-dritte-weg.info/2020/03/diskussionsbeitrag-zur-corona-krise-notwendiges-vorgehen-oder-ausbau-der-diktatur/

Devakumar, D., Shannon, G., Bhopal, S. S. & Abubakar, I. 2020. 'Racism and discrimination in COVID-19 responses'. *The Lancet* 395(10231): 1194.

Digital, Culture, Media and Sports Committee. 2019. 'Disinformation and "fake news"'. Report. As of 12 August 2020: https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

Dodd, V. 2020. 'Fears of rise in UK terrorist recruits as anti-radicalisation referrals collapse'. *The Guardian*. 22 April 2020. As of 14 September 2020: https://www.theguardian.com/uk-news/2020/apr/22/fears-of-rise-in-uk-terrorism-recruits-after-anti-radicalisation-referrals-collapse-coronavirus

Douglas, K., Usinski, Sutton, R. Cichocka, A. Nefes, T Ang, C. & Deravi, F. 2019a. 'Understanding conspiracy theories'. *Political Psychology* 40 (1).

———. 2019b. 'Why do people adopt conspiracy theories, how are they communicated, and what are their risks?'. Centre for Research and Evidence on Security Threats.

Duffy, B. 2020. 'Coronavirus uncertainties: vaccines, symptoms and contested claims'. King's College London.

European Commission (EC). 2018a. 'Action Plan Against Disinformation'. Joint Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions.

———. 2018b. 'Tackling online disinformation: A European Approach'. Joint Communication from the Commission to the European Parliament, the Council, the European. Economic and Social Committee and the Committee of the Regions. As of 12 August 2020: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236

———. 2019. 'A Europe that protects: EU reports on progress in fighting disinformation ahead of European Council'. As of 15 September 2020: https://ec.europa.eu/commission/presscorner/detail/en/IP_19_2914

Fangen, K. & Holter, C.R. 2020. 'The battle for truth: How online newspaper commenters defend their censored expressions'. *Poetics* 80.

Ferrari, R. 2015. 'Writing narrative style literature reviews'. *Medical Writing* 24(4): 230–235.

Flynn, D., Nyhan, B. & Reifler, J. 2017. 'The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics'. *Political Psychology* 38(1).

Full Fact. 2020. 'Automated Fact Checking'. As of 14 September 2020: https://fullfact.org/automated

Fundamental Rights Agency. 2018. 'FRAbot to be presented at GAME CHANGER Conference'. As of 15 August 2020: https://fra.europa.eu/en/event/2018/frabot-be-presented-game-changer-conference

Funke, D. & Flamini, D. 2020. 'A guide to Anti-misinformation actions around the world'. Poynter. As of 14 August 2020: https://www.poynter.org/ifcn/anti-misinformation-actions/

Ganesh, B. & Bright, J. 2020. 'Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation'. *Policy and Internet* 12(1): 6–19. As of 8 October 2020: https://doi.org/10.1002/poi3.236

Ghosh, R., Chan, A., Manuel, A., Dilimulati, M. & Babaei, M. 2016. 'Education and Security: A Global Literature Review on the Role of Education in Countering Violent Religious Extremism'. Tony Blair Institute for Global Change. As of 19 October 2020: https://institute.global/sites/default/files/inline-files/IGC_Education%20and%20Security.pdf

Gibbs, S. 2017. 'What can be done about abuse on social media?'. *The Guardian*. 12 December 2017. As of 18 September 2020: https://www.theguardian.com/media/2017/dec/13/what-can-be-done-about-abuse-on-social-media

Global Internet Forum to Counter Terrorism (GIFCT). 2020. 'Global Internet Forum to Counter Terrorism: Evolving an Institution'. As of 18 September 2020: https://www.gifct.org/about/

Gover, A.R., Harper, S.B. & Langton, L. 2020. 'Anti-Asian Hate Crime During the COVID-19 Pandemic: Exploring the Reproduction of Inequality'. *American Journal of Criminal Justice*. 1–21.

Graves, L. 2013. 'Deciding What's True: Fact-Checking Journalism and the New Ecology of News'. Doctoral Dissertation, Columbia University.

———. 2018. 'Understanding the Promise and Limits of Automated Fact-Checking'. Technical report, Reuters Institute, University of Oxford.

Grierson, J. 2020. 'Anti-Asian hate crimes up 21% in UK during coronavirus crisis'. *The Guardian*. As of 3 August 2020: https://www.theguardian.com/world/2020/may/13/anti-asian-hate-crimes-up-21-in-uk-during-coronavirus-crisis

Guhl, J. & Ebner, J. 2018. 'Islamist and Far-Right Extremists: Rhetorical and Strategic Allies in the Digital Age'. Radicalisation Research. As of 8 October 2020: https://www.radicalisationresearch.org/debate/ebner-islamist-far-right-extremists-rhetorical-digital-age/

Han, N. 2020. 'I Don't Scare Easily, But COVID-19 Virus of Hate Has Me Terrified: Reporter's Notebook'. ABC News, May 23. As of 14 August 2020: https://abcnews.go.com/US/asian-americans-covid-19-racism-virus-hate-reporters/story?id=70810109

Hardage, D. & Najafirad, P. 2020. 'Hate and Toxic Speech Detection in the Context of Covid-19 Pandemic using XAI: Ongoing Applied Research'. As of 3 August 2020: https://openreview.net/forum?id=7HP_0BgVX7v

Hatakka, N. 2019. 'Expose, debunk, ridicule, resist! Networked civic monitoring of populist radical right online action in Finland'. Information Communication and Society. As of 8 October 2020: https://doi.org/10.1080/1369118X.2019.1566392

Hatemeter. 2020. 'The Project'. As of 15 September 2020: http://hatemeter.eu/?page_id=197

Hedayah & ICCT. 2014. 'Developing Effective Counter-Narrative Frameworks for Countering Violent Extremism: Meeting Note September 2014'. As of 30 September 2020: https://www.dhs.gov/sites/default/files/publications/Developing%20Effective%20Frameworks%20for%20CVE-Hedayah_ICCT%20Report.pdf

Heft, A., E. Mayerhöffer, S. Reinhardt & C. Knüpfer. 2020. 'Beyond Breitbart: Comparing Right-Wing Digital News Infrastructures in Six Western Democracies'. *Policy & Internet* 12 (1): 20–45.

Henley, J. 2020. 'How Finland starts its fight against fake news in primary schools'. *The Guardian*. 29 January 2020. As of 14 September 2020: https://www.theguardian.com/world/2020/jan/28/fact-from-fiction-finlands-new-lessons-in-combating-fake-news

Hern, A. 2020. 'How hate speech campaigners found Facebook's weak spot'. The Guardian. 29 June 2020. As of 19 October 2020: https://www.theguardian.com/technology/2020/jun/29/how-hate-speech-campaigners-found-facebooks-weak-spot

HM Government. 2018. 'Counter-terrorism strategy (CONTEST) 2018'. As of 19 October 2020: https://www.gov.uk/government/publications/counter-terrorism-strategy-contest-2018

Holbrook, D. 2020. 'The Challenge of Conspiracy Theories for Strategic Communications'. *The RUSI Journal*, 165:1, 26–36. doi:10.1080/03071847.2020.1734384

Hrčková, A., Srba, I., Móro, R., Blaho, R., Šimko, J., Návrat, P. & Bieliková, M. 2019. 'Unravelling the basic concepts and intents of misbehavior in post-truth society'. *Bibliotecas, Anales de Investigacion* 15(3): 421–428.

Humprecht, E. 2019. 'How Do They Debunk "Fake News"? A Cross-National Comparison of Transparency in Fact Checks'. *Digital Journalism* 8 (3): 310–327. As of 14 September 2020: https://www.tandfonline.com/doi/full/10.1080/21670811.2019.1691031

Imhoff, R. & Bruder, M. 2014. 'Speaking (Un-)Truth to Power: Conspiracy Mentality as a Generalised Political Attitude'. *European Journal of Personality* 28: 25–43. doi:10.1002/per.1930

Institute for Strategic Dialogue (ISD). 2020. 'Anatomy of a Disinformation Empire: Investigating NaturalNews'. ISD. As of 14 August 2020: https://www.isdglobal.org/isd-publications/investigating-natural-news/

Jasko, K., LaFree, G. & Kruglanski, A. 2017. 'Quest for Significance and Violent Extremism: The Case of Domestic Radicalization'. *Political Psychology* 38, 815–831.

Jeung, R. 2020. 'Incidents of coronavirus discrimination March 26–April 1, 2020: A report for A3PCON and CAA'. Asian Pacific Policy and Planning Council.

Jolley, D. & Douglas, K. 2014. 'The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint'. *British Journal of Psychology* 15(1). 35–56. As of 15 September 2020: https://onlinelibrary.wiley.com/doi/full/10.1111/bjop.12018

Jones, K. 2020. 'Online Disinformation and Political Discourse: Applying a Human Rights Framework'. Chatham house. As of 12 August 2020: https://www.chathamhouse.org/publication/online-disinformation-and-political-discourse-applying-human-rights-framework

Kaiser Health News (KHN). 2020. 'Listen: How the Pandemic Further Politicized Public Health'. As of 30 September 2020: https://khn.org/news/listen-how-the-pandemic-further-politicized-public-health/

Kim, Y. 2019. 'How conspiracy theories can stimulate political engagement'. *Journal of Elections, Public Opinion and Parties* 0(0): 1–21. As of 14 August 2020: https://www.tandfonline.com/doi/full/10.1080/17457289.2019.1651321

King's College London & Ipsos MORI. 2020. 'Covid conspiracies and confusions: the impact on compliance with the UK's lockdown rules and the link with social media use'. The Policy Institute, King's College London.

Kruglanski, A., Jasko, K., Webber, D., Chernikova, M. & Molinario, E. 2018. 'The Making of Violent Extremists'. *Review of General Psychology* 22 (1): 107–120.

Kumar, S. & Shah, N. 2018. False Information on Web and Social Media: A Survey. As of 19 October 2020: https://arxiv.org/pdf/1804.08559.pdf

L1ght. 2020. 'Rising Levels of Hate Speech & online toxicity during this time of crisis'. As of 16 August 2020: https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf

Larsson, P. 2020. 'Anti-Asian racism during coronavirus: How the language of disease produces hate and violence'. The Conversation. As of 4 August 2020: https://theconversation.com/anti-asian-racism-during-coronavirus-how-the-language-of-disease-produces-hate-and-violence-134496

Laub, Z. 2019. 'Hate Speech on Social Media: Global Comparisons'. CFR. 7 June 2019. As of 14 September 2020: https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

Lee, B. 2020. 'Countering Violent Extremism Online: The Experiences of Informal Counter Messaging Actors'. *Policy & Internet* 12 (1): 66–87.

Lessenski, M. 2019. 'Just think about it: Findings of the Media Literacy Index 2019'. Open Society Institute. As of 14 September 2020: https://osis.bg/wp-content/uploads/2019/11/MediaLiteracyIndex2019_-ENG.pdf

Lewis, A. & Marwick, A. 2017. 'Taking the Red Pill: Ideological Motivations for Spreading Online Disinformation'. Understanding and Addressing the Disinformation Ecosystem, University of Pennsylvania Annenberg School for Communication, Philadelphia, PA. As of 15 September 2020: http://www.tiara.org/wp-content/uploads/2018/05/lewis_marwick_redpill_ideological_motivations.pdf

Li, Y. & Galea, S. 2020. 'Racism and the COVID-19 Epidemic: Recommendations for Health Care Workers'. *American Journal of Public Health* 110(7): 956–957. As of 4 August 2020: https://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.2020.305698

Liz, J. 2020. 'State Racism, Social Justice, and the COVID-19 Pandemic'. *Public Philosophy Journal*. As of 4 August 2020: https://publicphilosophyjournal.org/full-record/?amplificationid=2146

Lowles, N. & Levene, J. 2019. 'State of Hate 2019: People vs the Elite?'. As of 8 October 2020: https://www.hopenothate.org.uk/wp-content/uploads/2019/02/state-of-hate-2019-final-1.pdf

Lu, R. & Sheng, Y. 2020. 'From Fear to Hate: How the Covid-19 Pandemic Sparks Racial Animus in the United States'. As of 4 August 2020: https://arxiv.org/ftp/arxiv/papers/2007/2007.01448.pdf

Lyu, H., Chen, L., Wang, Y. & Luo, J. 2020. 'Sense and sensibility: Characterizing social media users regarding the use of controversial terms for covid-19'. IEEE Transactions on Big Data.

Machete, P. & Turpin, M. 2020. 'The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review'. In: Hattingh M., Matthee M., Smuts H., Pappas I., Dwivedi Y., Mäntymäki M. (eds). 'Responsible Design, Implementation and Use of Information and Communication Technology'. *Lecture Notes in Computer Science 12067*. Springer, Cham. As of 14 September 2020: https://link.springer.com/chapter/10.1007/978-3-030-45002-1_20

Malik, Nikita. 2020. 'Self-isolation might stop Coronavirus, but it will speed the spread of extremism'. Foreign Policy. As of 4 August 2020: https://foreignpolicy.com/2020/03/26/self-isolation-might-stop-coronavirus-but-spread-extremism/

Manalo, Ricky. 2020. 'When the Racist Response to Covid-19 Hits Home'. America Magazine. As of 4 August 2020: https://www.americamagazine.org/faith/2020/03/30/when-racist-response-covid-19-hits-home

Manavis, Sarah. 2020. 'Covid-19 has caused a major spike in anti-Chinese and anti-Semitic hate speech'. *New Statesman*. 29 April 2020. As of 19 August 2020: https://www.newstatesman.com/science-tech/social-media/2020/04/covid-19-coronavirus-anti-chinese-antisemitic-hate-speech-5g-conspiracy-theory

Marwick, A. & Lewis, R. 2017. 'Media manipulation and disinformation online'. As of 8 October 2020: https://datasociety.net/library/media-manipulation-and-disinfo-online/

McDowell-Smith, A., Speckhard, A. & Yayla, A.S. 2017. 'Beating ISIS in the digital space: Focus testing ISIS defector counter-narrative videos with American college students'. *Journal for Deradicalization* 10: 50–76. As of 30 September 2020: https://journals.sfu.ca/jd/index.php/jd/article/view/83/73

McNeil-Willson, Richard. 2020a. 'Framing in times of crisis: Responses to COVID-19 amongst Far Right movements and organisations'. ICCT. As of 8 October 2020: https://icct.nl/publication/framing-in-times-of-crisis/

———. 2020b. 'What the "war on terror" can teach us about the fight against COVID-19'. Open Democracy. As of 8 October 2020: https://www.opendemocracy.net/en/global-extremes/what-the-war-on-terror-can-teach-us-about-the-fight-against-covid-19/

Moonshot. 2020a. 'COVID-19: conspiracy theories, hate speech and incitements to violence on Twitter'. As of 14 September 2020: http://moonshotcve.com/covid-19-conspiracy-theories-hate-speech-twitter/

———. 2020b. 'The Impact of COVID-19 on Canadian Search Traffic: Double-digit increases in engagement with extremist content in Canada's six largest cities'.

Mueller, K., & Schwarz, C. 2020. 'Fanning the Flames of Hate: Social Media and Hate Crime'. https://warwick.ac.uk/fac/soc/economics/staff/crschwarz/fanning-flames-hate.pdf

Muzzatti, S.L. 2005. 'Bits of Falling Sky and Global Pandemics: Moral Panic and Severe Acute Respiratory Syndrome (SARS)'. *Illness Crisis and Loss* 13: 117–128. doi:10.1177/105413730501300203

Naseer, M.S. 2020. 'COVID 19: Incubator for Online Extremism'. NCTC. As of 4 August 2020: https://www.researchgate.net/profile/Shahuneeza_Naseer/publication/342170161_COVID-19_Incubator_for_Online_Extremism/links/5ee71302458515814a5e9afd/COVID-19-Incubator-for-Online-Extremism.pdf?origin=publication_detail

National Literacy Trust. 2018. 'Fake News and Critical Literacy: The final report of the Commission on Fake News and the Teaching of Critical Literacy in Schools'. As of 15 September 2020:

https://cdn.literacytrust.org.uk/media/documents/Fake_news_and_critical_literacy_-_final_report.pdf

Ng, Edmond. 2020. 'The Pandemic of Hate is Giving COVID-19 a Helping Hand'. *The American Journal of Tropical Medicine and Hygiene* 102(6): 1158. As of 4 August 2020: http://www.ajtmh.org/content/journals/10.4269/ajtmh.20-0285

Nyhan, B. & Zeitzoff, T. 2018. 'Conspiracy and misperception belief in the Middle East and North Africa'. *Journal of Politics* 80(4): 1400–1404. As of 8 October 2020: https://doi.org/10.1086/698663

Ockenden, Sasha. 2020. 'Misinformation, Disinformation and the Coronavirus'. Stop Funding Hate. As of 14 August 2020: https://stopfundinghate.info/2020/06/08/guest-post-tactical-tech-misinformation-disinformation-and-the-coronavirus/

Ofcom. 2020. 'COVID-19 news and information: consumption and attitudes'. Ofcom. As of 8 October 2020: https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/coronavirus-news-consumption-attitudes-behaviour

Parker, D. & Lindekilde, L. 2020. 'Preventing Extremism with Extremists: A Double-Edged Sword? An Analysis of the Impact of Using Former Extremists in Danish Schools'. *Education Science* 10(4) 111.

Parlakkılıç, Alaattin. 2018. 'Cyber Terrorism Through Social Media: A Categorical Based Preventive Approach'. *International Journal of Information Security Science* 7(4): 172–178. As of 4 August 2020: https://www.researchgate.net/publication/331178142_Cyber_Terrorism_Through_Social_Media_A_Categorical_Based_Preventive_Approach

Pei, X. & Mehta, D. 2020. '#Coronavirus or #Chinesevirus?!: Understanding the negative sentiment reflected in Tweets with racist hashtags across the development of COVID-19'. As of 4 August 2020: https://arxiv.org/abs/2005.08224

Polyakova, A. & Fried, D. 2019. 'Democratic Defense Against Disinformation 2.0'. Brookings. As of 8 October 2020: https://www.brookings.edu/research/democratic-defense-against-disinformation-2-0/

Popat, Rajiv. 2020. 'Rise in racism following Leicester lockdown'. ITV. As of 14 August 2020: https://www.itv.com/news/central/2020-07-02/rise-in-racism-following-leicester-lockdown

Poynter. 2018. Here's how close automated fact-checking is to reality. As of 14 September 2020: https://www.poynter.org/fact-checking/2018/heres-how-close-automated-fact-checking-is-to-reality/

———. 2019. ‘Commit to transparency — sign up for the International Fact-Checking Network's code of principles’. As of 14 September 2020: https://ifcncodeofprinciples.poynter.org/

Priest, N., Thurber, K.A., Maddox, R., Jones, R. & Truong, M. 2020. ‘COVID-19 racism is making kids sick’. As of 4 August 2020: https://insightplus.mja.com.au/2020/18/covid-19-racism-is-making-kids-sick/

Pronin, E. 2007. ‘Perception and misperception of bias in human judgment’. *Trends in Cognitive Sciences* 11(1). January 2007. 37–43. As of 14 September 2020: https://www.sciencedirect.com/science/article/pii/S1364661306002993?via%3Dihub

Radnitz, S. & Underwood, P. 2017. ‘Is Belief in Conspiracy Theories Pathological? A Survey Experiment on the Cognitive Roots of Extreme Suspicion’. *British Journal of Political Science* 47(1): 113–129.

Readfearn, G. 2020. ‘How did coronavirus start and where did it come from? Was it really Wuhan’s animal market?’ *The Guardian*. 27 April 2020. As of 15 September 2020: https://www.theguardian.com/world/2020/apr/28/how-did-the-coronavirus-start-where-did-it-come-from-how-did-it-spread-humans-was-it-really-bats-pangolins-wuhan-animal-market

Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A. & Meira, W. 2019. ‘Auditing Radicalization Pathways on YouTube’. As of 13 August 2020: https://arxiv.org/abs/1908.08313

Russell, Anna. 2020. ‘The rise of coronavirus hate crimes’. *The New Yorker*. As of 5 August 2020: https://www.newyorker.com/news/letter-from-the-uk/the-rise-of-coronavirus-hate-crimes

Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y. & Zannettou, S. 2020. ‘“Go eat a bat, chang!”: An early look on the emergence of sinophobic behavior on web communities in the face of covid-19’. *Computer Science*. As of 8 October 2020: https://arxiv.org/pdf/2004.04046.pdf

Schwarz, K. & Holnburger, J. 2019. ‘Disinformation: what role does disinformation play for hate speech and extremism on the internet and what measures have social media companies taken to combat it?’. In Baldauf et al. (eds), ‘Hate Speech and Radicalisation Online: The OCCI Research Report’, 35. As of 5 August 2020: https://www.isdglobal.org/wp-content/uploads/2019/06/ISD-Hate-Speech-and-Radicalisation-Online-English-Draft-2.pdf

Silver, Laura. R. 2016. ‘China in the Media: Effects on American Opinion’. *Publicly Accessible Penn Dissertations.* 2017. As of 14 August 2020: https://repository.upenn.edu/cgi/viewcontent.cgi?article=3803&context=edissertations

Smith, V. & Wanless, A. 2020. ‘Unmasking the Truth: Public Health Experts, the Coronavirus, and the Raucous Marketplace of Ideas, Carnegie Endowment for International Peace’. As of 8 October 2020: https://carnegieendowment.org/2020/07/16/unmasking-truth-public-health-experts-coronavirus-and-raucous-marketplace-of-ideas-pub-82314

Spencer, L., Ritchie, J., Lewis, J. & Dillon, L. 2003. 'Quality in qualitative evaluation: a framework for assessing research evidence'. As of 19 August 2020: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/498321/Quality-in-qualitative-evaulation_tcm6-38739.pdf

Sunstein, C.R. & Vermeule, A. 2009. 'Conspiracy theories: Causes and cures'. *Journal of Political Philosophy* 17, 202–227. As of 13 August 2020: https://doi.org/10.1111/j.1467-9760.2008.00325.x

Stop Funding Fake News. 2020. 'More wins: Less money for Fake News'. As of 16 August 2020: https://www.stopfundingfakenews.com/successes

Stop Hate for Profit. 2020. Partners. As of 19 October 2020: https://www.stophateforprofit.org/

Tekiroglu, S., Chung, Y. & Guerini, M. 2020. 'Generating Counter Narratives against Online Hate Speech: Data and Strategies'. As of 16 September 2020: https://arxiv.org/abs/2004.04216

Tobin, A., Varner, M. & Angwin, J. 2017. 'Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up'. ProPublica. 28 December 2017. As of 14 September 2020: https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes

Tran, Elaine. 2020. 'Warning: Symptoms May Include Racism: A Content Analysis of Anti-Asian Racism and Sentiment amid the COVID-19 Pandemic in Digital Media'. As of 5 August 2020: https://roam.macewan.ca/islandora/object/gm:2134

Thorne, J. & Vlachos, A. 2018. 'Automated Fact Checking: Task Formulations, methods and future directions'. As of 14 September 2020: https://arxiv.org/pdf/1806.07687.pdf

Thorson, Emily. 2016. 'Belief Echoes: The Persistent Effects of Corrected Misinformation'. *Political Communication* 33.

Tuck, H. & Silverman, T. 2016. 'The Counter-Narrative Handbook'. Institute for Strategic Dialogue. As of 30 September 2020: https://www.isdglobal.org/wp-content/uploads/2016/06/Counter-narrative-Handbook_1.pdf

United Nations (UN). 2020. 'The Impact of the COVID-19 Pandemic on Counter-Terrorism and Countering Violent Extremism'. CTED. June 2020. As of 19 October 2020: https://www.un.org/sc/ctc/wp-content/uploads/2020/06/CTED-Paper%E2%80%93-The-impact-of-the-COVID-19-pandemic-on-counter-terrorism-and-countering-violent-extremism.pdf

United Nations Educational, Scientific, and Cultural Organisation (UNESCO). 2020. 'Youth Engagement is Key to Counter the Rise of Antisemitism spurred on by COVID-19'. 30 June 2020. As of 19 October 2020: https://en.unesco.org/news/youth-engagement-key-counter-rise-antisemitism-spurred-covid-19

University of Oxford. 2020. 'Conspiracy beliefs reduce the following of government coronavirus guidance'. University of Oxford. 22 May 2020. As of 18 August 2020: https://www.ox.ac.uk/news/2020-05-22-conspiracy-beliefs-reduces-following-government-coronavirus-guidance

Uscinski, J. 2016. 'How playing on conspiracy theories can be key to electoral success'. LSE online. As of 14 September 2020: http://blogs.lse.ac.uk/usappblog/2016/06/07/how-playing-on-conspiracy-theories-can-be-key-to-electoral-success/

Vegetti, F. & L. Levente. 2020. 'Belief in conspiracy theories, aggression, and attitudes towards political violence'. As of 8 October 2020: https://federicovegetti.github.io/pdfs/paper_conspiracy_2020.pdf

Velásquez, N., Leahy, R., Restrepo, N.J., Lupu, Y., Sear, R., Gabriel, N., Jha, O. & Johnson, N. 2020. 'Hate multiverse spreads malicious COVID-19 content online beyond individual platform control'. As of 5 August 2020: https://arxiv.org/abs/2004.00673

Vidgen, B., H. Margetts & A. Harris. 2019a. 'How much online abuse is there? A systematic review of evidence for the UK. Policy Briefing: Full Report'. Alan Turing Institute.

Vidgen, Bertie, Yasseri, T. & Margetts, H. 2019b. 'Trajectories of Islamophobic hate amongst far right actors on Twitter'. University of Oxford. As of 20 August 2020: https://arxiv.org/ftp/arxiv/papers/1910/1910.05794.pdf

Vosoughi, S., Roy, D. & Aral, S. 2018. 'The spread of true and false news online'. *Science* 1151(March), 1146–1151. As of 14 August 2020: https://science.sciencemag.org/content/359/6380/1146

Wallner, Claudia. 2020. 'Preventing and Countering Violent Extremism through Education Initiatives: Assessing the Evidence Base'. RUSI. As of 8 October 2020: https://rusi.org/sites/default/files/pcve_education_final_web_version.pdf

Wang, Cynthia. 2020. 'Uncertainty. Loss of Control. Why COVID-19 Is a Perfect Storm for Conspiracy Theories'. As of 8 October 2020: https://insight.kellogg.northwestern.edu/article/uncertainty-loss-control-covid-19-conspiracy-theories

Wardle, Clair. 2017. 'Fake news. It's complicated'. First Draft. As of 14 August 2020: https://firstdraftnews.org/latest/fake-news-complicated/

Wilson Centre. 2020. 'What Islamists Are Doing and Saying on COVID-19 Crisis'. Wilson Centre. As of 14 August 2020: https://www.wilsoncenter.org/article/what-islamists-are-doing-and-saying-covid-19-crisis

Wilson, J. 2020. 'Disinformation and blame: how America's far right is capitalizing on coronavirus'. *The Guardian*. 19 March 2020. As of 13 August 2020:

https://www.theguardian.com/world/2020/mar/19/america-far-right-coronavirus-outbreak-trump-alex-jones

Ziems, C., He, B., Soni, S. & Kumar, S. 2020. 'Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis'.

Zine, Jasmine. 2020. 'Unmasking the racial politics of the coronavirus pandemic'. The conversation. 3 June 2020. As of 8 October 2020: https://theconversation.com/unmasking-the-racial-politics-of-the-coronavirus-pandemic-139011

# Annex A. Glossary of key terms

Table A-1 provides expanded definitions of the key terms identified in the literature review.

**Table A-1: Glossary of key terms**

| Term | Definition | Source |
|---|---|---|
| **Abusive content** | 'Abusive content' encompasses harmful content, which often does not abide by platforms' regulation guidelines. | Vidgen et al. (2019a) |
| **Activation** | 'Activation' refers to the first publication of either a hate or a counter-hate message on Twitter. | Ziems et al. (2020) |
| **Alt-right** | 'Alt-right' refers to a neologism coined in 2008 to define right-wing political supporters misaligned with conservatives. It is sometimes understood as a new way to designate 'white nationalists' and 'white supremacists' to introduce their discourse into mainstream media. | Marwick & Lewis (2017) |
| **Antisemitism** | 'Antisemitism' refers to a perception in relation to the Jewish community that can be expressed as hate against Jews, Jewish institutions and religious facilities. | Allington (2020); UNESCO (2020) |
| **Belief Persistence** | 'Belief persistence' refers to an insistence in maintaining misperceptions even after they have been proved to be false. | Thorson (2016) |
| **Civil monitoring** | 'Civil monitoring' refers to voluntary activities conducted online by individuals or groups to monitor online actions they consider to be harmful to societies. | Hatakka (2019) |
| **Confirmation bias** | 'Confirmation bias' is a tendency for individuals to favour information that aligns with their pre-existing views rather than be exposed to other points of view. | Mueller & Schwarz (2020) |
| **Conspiracy belief** | 'Conspiracy beliefs' are held by those who recognise a conspiracy theory (see below) as being true. | Douglas et al. (2019a) |
| **Conspiracy theory** | 'Conspiracy theories' are narratives created to infer that an event or | Bolsen & Druckman (2018); Connolly et al. (2019); Douglas et al. |

| | | |
|---|---|---|
| | situation is the result of a secret plan made by powerful individuals or groups. | (2019a); Vegetti & Levente (2020); ISD (2020); Fangen & Holter (2020); Holbrook (2020) |
| **Dangerous speech** | 'Dangerous speech' refers to forms of expression that are likely to prompt an audience to support hateful activities and commit violence. Dangerous speech, unlike hate speech (see below) does not target a particular individual or group. | Benesch et al. (2020) |
| **Disinformation** | 'Disinformation' refers to false information that is intentionally shared and disseminated in an organised fashion to mislead the audience and satisfy political, financial, psychological or social motivations. Coined by Stalin, the term comes from the Russian 'dezinformatsiya' to label Soviet propaganda campaigns. Disinformation is different from misinformation and malinformation (see below). | ISD (2020); Ball & Maxmen (2020); Ockenden (2020); Digital, Culture, Media and Sports Committee (2019); European Commission (2018b); Cherilyn & Posetti (2018); Jones (2020); Schwarz & Holnburger (2019) |
| **Ethnonationalism** | 'Ethnonationalism' refers to a form of nationalism where ethnicity is the link between all nation members in addition to a shared heritage and culture. | Davey & Hart (2020) |
| **Extremism** | 'Extremism' is a system of belief that opposes democratic values, establishes a hierarchy between members of a group that are superior over those who do not belong and wish to remain separated from them. Extremists advocate for societal change that aligns with their beliefs and are willing to use violent means to achieve their goals. | Baldauf et al. (2019); Lowles & Levene (2019) |
| **Fact-checking** | 'Fact-checking' refers to journalistic activities focused on the assessment of public claims, including political claims, to identify whether they are true or false/misleading. | Amazeen et al. (2018) |
| **False amplifier** | 'False amplifier' refers to coordinated action online by inauthentic accounts, primarily on social media, to distort and manipulate political debates. | Council of Europe (2017a) |
| **False information** | 'False information' is used in this report as a catch-all term to refer collectively to online misinformation, disinformation and conspiracy theories (see definitions of these terms). | Kumar and Shah (2018); Brennen et al. (2020) |
| **Fake news** | 'Fake news' refers to a type of false information that can contain misinformation (see below) or disinformation (see above), and can be | Marwick & Lewis (2017); Digital, Culture, Media and Sports Committee (2019); Jones (2020); |

| | | |
|---|---|---|
| | spread in the media and online. | Schwarz & Holnburger (2019) |
| **Far-right** | 'Far-right' refers to those on the political spectrum that gather around at least three of the following features: nationalism, racism, xenophobia, antidemocracy and strong-state advocacy. They are likely to support white supremacy and adhere to conspiracy theories (see above). | ISD (2020); Fangen & Holter (2020) |
| **Hate crime** | 'Hate crime' consists of a criminal act motivated by the perpetrator's bias against their real or perceived identity, e.g. in relation to race, religion, nationality, origin, gender or sexual orientation. | Gover et al. (2020) |
| **Hate speech** | 'Hate speech' refers to those forms of expression that aim to discriminate against one group of people to alienate them based on their identity – real or perceived by the author according to their bias – according to their religion, nationality, ethnicity, gender or sexual orientation. Hate speech should be differentiated from dangerous speech (see above). | Baldauf et al. (2019); DePaula et al. (2018) |
| **Hateful extremism** | 'Hateful extremism' refers to hateful language, narratives and behaviours that 'incite and amplify hate, or engage in persistent hatred, or equivocate about and make the moral case for violence', drawing on hateful, hostile or supremacist beliefs directed at an out-group, 'who are perceived as a threat to the wellbeing, survival or success of an in-group'; which can cause harm to individuals, communities or members of that out-group or wider society as a whole. | CCE (2019) |
| **Hyper partisan sites** | 'Hyper partisan sites' are those sites deeply rooted in an ideology used as a lens through which they disseminate false information and decontextualised information to create misleading views. | Marwick & Lewis (2017) |
| **Malinformation** | 'Malinformation' refers to the publication of private information to deliberately cause harm to an individual or an organisation. Malinformation is different from disinformation (see above) and misinformation (see below). | Council of Europe (2017a); Cherilyn & Posetti (2018) |
| **Misinformation** | 'Misinformation' refers to false or incorrect information that is spread | Ball & Maxmen (2020); Ockenden (2020); Digital, Culture, Media and |

| | | |
|---|---|---|
| | without necessarily involving the intention to mislead an audience. Misinformation is different from disinformation (see above) and malinformation (see above). The term is sometimes used to define any type of false information regardless of motivations. | Sports Committee (2019); Brennen et al. (2020); Cherilyn & Posetti (2018); Jones (2020) |
| **Misperceptions** | 'Misperceptions' are associated with errors in human judgements and decision making, particularly when these are caused by bias due to existing beliefs, expectations, context, needs, motives and desires. | Pronin (2007) |
| **Prejudices** | 'Prejudices' refer to those negative unjustified attitudes held against members of a designated group. | Barderi (2018) |
| **Preventing violent extremism (PVE)** | 'Preventing violent extremism' refers to preventative approaches to reduce the likelihood of individual radicalisation and of violent extremist beliefs and behaviours (see below). | Wallner (2020) |
| **Radicalisation** | 'Radicalisation' is a process through which an individual is drawn to violent extremism (see below). | Avis (2020) |
| **Religious hate speech** | 'Religious hate speech' (also referred to as religious hatred) refers to a type of hate speech (see above) that specifically targets and discriminates against individuals or groups with different or no religious believes. | Albadi et al. (2019) |
| **Violent extremism** | 'Violent extremism' refers to the adoption and the use of violence as a means to achieve political, economic or social goals. | Avis (2020) |
| **Zoombombing** | 'Zoombombing' refers to the intrusion of virtual meetings held on video conferencing platforms to disseminate hateful speech. | CST (2020) |

Source: RAND Europe (2020) review of sources listed in Table A-1.

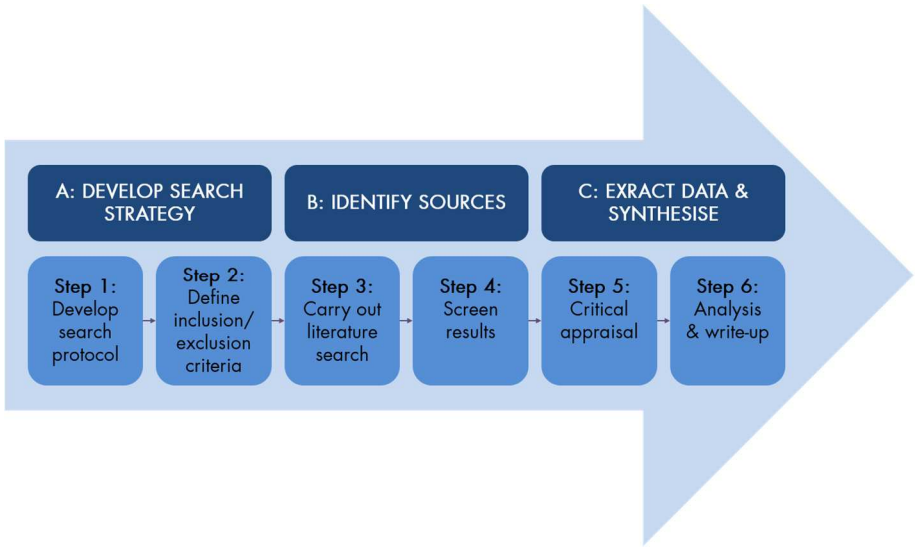
# Annex B. Research approach

This annex describes the research methods used to undertake this study, a summary of which is presented in Section 1.3.

## B.1. Overview of approach

The study team used a Rapid Evidence Assessment (REA) approach to deliver the literature review. As presented in Figure B-1, the REA approach followed three stages: (A) Developing a search strategy; (B) Identifying sources; and C) Data extraction and final synthesis.

**Figure B-1: REA approach**




## B.2. Develop search strategy

The first task involved developing a search strategy in order to set out the parameters of the REA. This was developed with the input and guidance of a RAND Librarian and in consultation with CCE. The search protocol included details regarding search terms, selected databases and inclusion and exclusion criteria.

To develop the search strategy, a number of indicative search terms were piloted and relevant databases were identified. Pilot testing of search terms was conducted to help ensure that the terms were broad enough to include a range of relevant studies, but narrow enough that the search citation numbers were manageable.

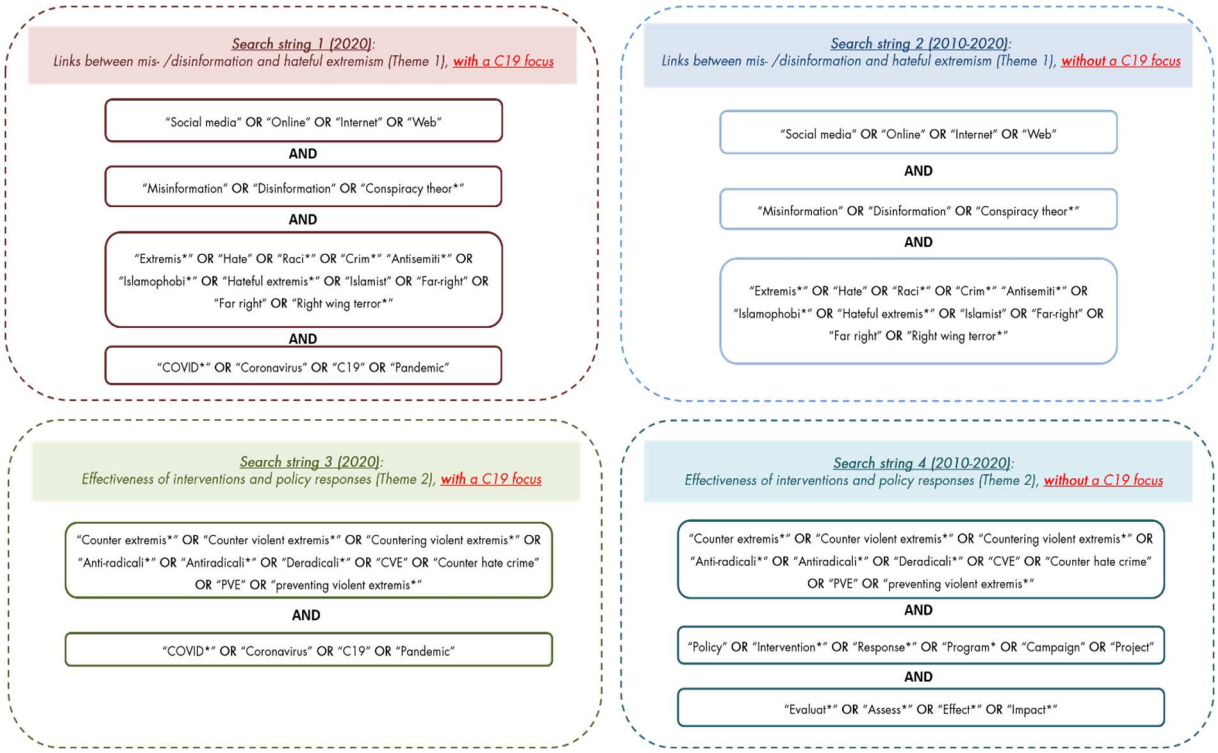
The protocol involved the use of four search strings, covering the two review themes from both a general standpoint, and with a focus on sources that considered the effect of COVID-19:

- **Search string 1**: Theme 1 (links between hateful extremism and false information), with a COVID-19 focus.

- **Search string 2**: Theme 1 (links between hateful extremism and false information), without a COVID-19 focus.
- **Search string 3**: Theme 2 (associated online interventions and policy responses), with a COVID-19 focus.
- **Search string 4**: Theme 2 (associated online interventions and policy responses), without a COVID-19 focus.

The final search strings are presented in Figure B-2 below.

**Figure B-2: Search strings**




A set of inclusion and exclusion criteria was developed and applied to the REA. These criteria are presented in Table B-1.

**Table B-1: Inclusion and exclusion criteria**

| | Inclusion criteria | Exclusion criteria | Rationale |
|---|---|---|---|
| **Research focus** | Literature that focuses on the two review themes:<br><br>1. Links between hateful extremism and false information.<br><br>2. Associated interventions and policy responses. | Research not concerned with the review themes specified in the inclusion criteria. | Focusing on the thematic areas discussed and agreed with CCE at the beginning of the study. |
| **Geographic location** | UK (primary focus)<br><br>Rest of world (secondary). | None. | Focusing on the UK as the country of primary interest; including other country-focused sources to identify practices, insights and lessons from other contexts. |
| **Source type** | Academic research.<br><br>Grey literature (research papers, evaluations, policy documentation, polling data).<br><br>[*Media sources – secondary focus*]. | Documents without clear organisational authorship, letters, editorials, comments, book reviews. | Optimising the quality of sources in the literature search. |
| **Language** | English-language sources. | Non-English language sources. | Expectation that English-language sources would yield a sufficient volume of literature within the study time constraints. |
| **Publication date** | 2020 (searches 1; 3).<br><br>2010-20 (searches 2; 4). | Research published before 2020 (1; 3); before 2010 (2; 4). | 2020 selected for 1 & 3 as C19 is specific to this year; 2010–2020 selected for 2 & 4 to ensure relevance and timeliness of papers while keeping the range broad enough to capture sufficient results. |

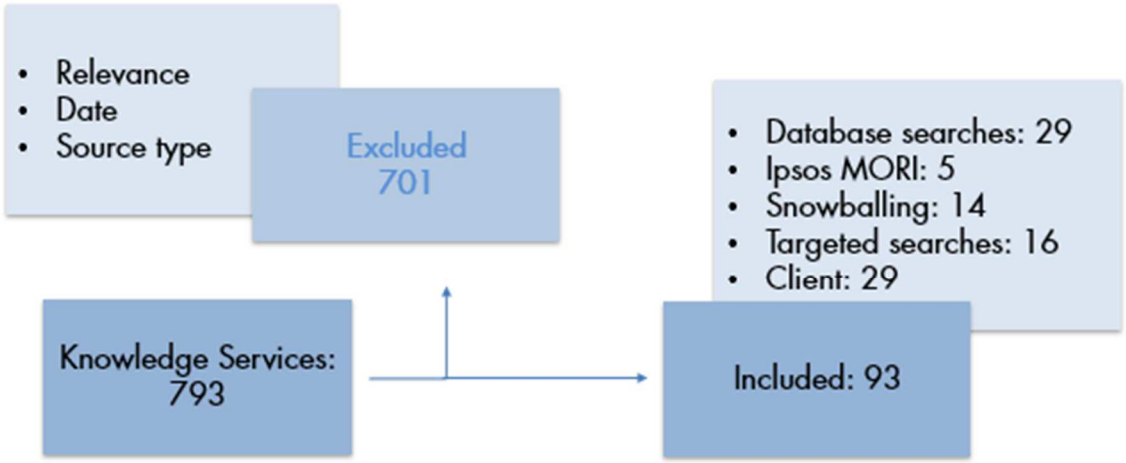
The literature search was conducted in July 2020 in the following databases:

- Academic Search Complete;
- Policy File Index;
- Scopus;
- Google Scholar.

The search strings presented in Figure B-2 were used in the Academic Search Complete, Policy File Index and Scopus searches, with searches applied to the title, abstract or subject fields of source records. A simplified approach was used in Google Scholar given the more limited search combinations the latter database is able to process, and the searches were applied to source titles only.

Figure B-3 presents the number of resulting papers from the initial database search.

**Figure B-3: REA study selection**




With duplicates removed across databases and search strings, the initial literature review identified 793 sources. After the titles and abstracts of these records were scanned for relevance (see 'Step 3'), 701 were removed. Records were excluded due to their lack of relevance to the study scope, and sources were also excluded where the publication data preceded 2010 (for searches 2 and 4) or preceded 2020 (for searches 1 and 3), or where the search type did not meet the inclusion criteria (see Table B-1).

Further to the database searches, the RAND team undertook targeted searches of organisational publications and 'snowball' searches; that is, searching the reference lists of the academic papers and grey literature identified through the database searches. The targeted searches focused on webpages and publications from the following organisations:

- Brookings Institution;
- Carnegie Endowment for International Peace;
- Chatham House;
- Council of Europe;
- European Commission;
- EU vs. Disinfo;
- IGLA Europe;

- Social Observatory for Disinformation and Social Media Analysis;
- Stop Funding Fake News;
- Stop Hate UK;
- Teaching Tolerance;
- UNESCO.

Finally, as presented in Figure B-3 RAND Europe included 33 sources recommended by CCE (28) and Ipsos MORI (5). As Figure B-3 illustrates, 93 sources were found to be relevant to the themes of the review and the inclusion criteria, and were therefore included in the full-text review.[266]

## B.3. Identify sources

Applying the final search terms presented in Figure B-2, the study team performed the searches on the titles, abstracts and subject fields of records within Academic Search Complete, Policy File Index and Scopus; and on the source titles of records in Google Scholar. All search results were then loaded into EndNote bibliographic software before titles and abstracts of these records were screened for relevance against the criteria specified in Table B-1. When screening the titles and abstracts of records, a researcher coded whether records should be included or excluded from the full text review. Records were excluded if the titles and abstracts did not meet the inclusion criteria.

## B.4. Data extraction and synthesis

A full-text review of the 93 sources was then undertaken. For each source, data was extracted into an Excel spreadsheet, with content mapped against the following categories:

- Bibliographic information;
- Country focus;
- COVID-19 vs. non-COVID-19 focus;
- Type of hateful extremism;
- Target population (i.e. for disinformation efforts and/or interventions);

[266] It should be noted that an additional 10 sources were included in the review, following submission of the Draft Final Report, in order to address final areas of interest highlighted by CCE and to add further supporting evidence. These papers were identified through a targeted Google search: Banjo et al. (2019); Barclay (2011); CCDH & Restless Development (2020); CoE (2017b); GIFCT (2020); Gibbs (2017); Hedayah & ICCT (2014); KHN (2020); McDowell-Smith et al. (2017); Tuck & Silverman (2016). Please note that these sources are *additional to* (rather than included in) the source numbers quoted in this annex, Chapter 2 and the additional Annex C.

- Focus on mis- and/or disinformation;
- Key definitions (e.g. misinformation, disinformation, hateful extremism);
- Key findings for review theme 1 (nature of links between hateful extremism and false information; impact of these links; insights from neighbouring policy areas);
- Key findings for review theme 2 (nature of existing online interventions and policy responses; effectiveness of these counter-measures; any source recommendations for future policy and practice);
- Underlying research methods and evidence gaps identified by source;
- Additional content of relevance for review write-up.

To collate and summarise the REA findings, an Internal Synthesis Workshop was held on 11 August 2020, at which researchers from RAND Europe and Ipsos MORI discussed the emerging findings from the literature review and their implications for primary research design. Findings were then written up in a narrative synthesis that was structured in relation to the two review themes and the supporting research questions, and then integrated into this report. A quantitative overview of reviewed sources was also produced through an analysis of the included and excluded studies at the initial searching, screening and data-extraction stages.