DOI: 10.1002/cl2.1245

PROTOCOL

Campbell Collaboration WILEY

PROTOCOL: Hate online and in traditional media: A systematic review of the evidence for associations or impacts on individuals, audiences, and communities

Ghayda Hassan[1] | Jihan Rabah[2] | Pablo Madriaza[3] | Sebastien Brouillette-Alarie[1] | Eugene Borokhovski[2] | David Pickup[4] | Wynnpaul Varela[5] | Melina Girard[5] | Loïc Durocher-Corfa[5] | Emmanuel Danis[5]

[1]Department of Psychology, Université du Québec à Montréal, Montréal, Quebec, Canada

[2]Concordia University, Montréal, Quebec, Canada

[3]Université du Québec à Montréal, Montréal, Quebec, Canada

[4]Centre for the Study of Learning and Performance, Concordia University, Montréal, Quebec, Canada

[5]Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence, Department of Psychology, Université du Québec à Montréal, Montréal, Quebec, Canada

Correspondence
Ghayda Hassan, Department of Psychology, Université du Québec à Montréal, Adrien-Pinard Bldg (SU), PO Box 888, Downtown Station, Montréal, QC, H3C 3P8, Canada.
Email: hassan.ghayda@uqam.ca

Abstract

This is the protocol for a Campbell systematic review: The objectives are as follows: (1) to critically and systematically synthesize the empirical evidence on the effects or impacts of exposure to or consumption, active search, or promotion of hate content online or in traditional media; (2) to describe how the characteristics of hate (e.g., type of content, ideologies, severity, type of platform) impact the documented effects; (3) to collect and identify the role of contextual variables (e.g., individual traits, age, gender, socio-economic background) on the documented effects; (4) to collect and produce a meaningful classification of outcomes; and (5) to identify gaps and limitations in the research and related policy documents.

1 | BACKGROUND

1.1 | The problem

With the boom of interactive social media, hostile and offensive hate content has increased exponentially around the world (Weber et al., 2020). In surveys of young people between the ages of 15 and 30, as many as 53% of American, 48% of Finnish, and 39% of British respondents report having been exposed to hateful online material. In some cases, this exposure is also accompanied by victimization. For example, in the United Kingdom, 10%–20% also report being an actual target of abuse (Vidgen et al., 2019). In New Zealand, the same is true for 11% of the adult population, while in the United States, as many as 41% of adults recount experiences of being victimized (Waqas et al., 2019).

Technological advancements such as the use of social networking to chat, search and exchange knowledge, express thoughts, and engage with others have rendered social media a convenient and effective platform of interaction (Rabah, 2014). However, the accessibility of popular social network sites like Facebook and

[Correction added on 30 June 2022, after first online publication: "Loïc Durocher-Corfa" Author name has been corrected in this version.]

Twitter (Mossie & Wang, 2019), the elective anonymity of cyberspace, and the ease with which divisive opinions can be expressed online (Davidson et al., 2019) are at least partly responsible for the spike in online hate content. Although the precise determinants of such content can vary, research suggests the offline implications of digital hate speech are significant, with detrimental effects at the individual, community, and society level (Del Vigna et al., 2017). Studies point to a clear association between digital hate speech and actual hate crime (Mossie & Wang, 2019), as well as online hate speech and offline violence against targeted communities (Weber et al., 2020). Moreover, according to a recent systematic review, exposure to radical violent online material is linked to extremist beliefs and possibly a higher risk of violent behaviors (Hassan et al., 2018).

Although hate content has been mostly associated with the Internet, traditional media have also been implicated in the transmission of hate content, particularly in regions where the Internet has had a lower penetration, as well as in rural sectors. In Africa, for example, hate content broadcast on radio has been linked to violence and destabilization (Pate & Ibrahim, 2020; Somerville, 2011). The most extreme case of this connection has probably been Rwanda, where several researchers have pointed out the influence that hate content broadcast on the radio had on the genocide in the country (Adelman & Suhrke, 2017; Straus, 2007). In Western countries, local radio stations have also been associated with hate content. This is the case of hate speech on commercial radio in the United States in which Latino communities are stigmatized and stereotyped (Noriega & Iribarren, 2012) or shock radio in the province of Quebec in Canada, which have been the source of the transmission of Islamophobic rhetoric in the country (Perry, 2019; Perry et al., 2017). Even though newspapers have had less attention on these issues, some research has emphasized their role in the production of this content. According to Merklejn and Wiślicki (2020), traditional newspapers in Japan, for example, have played an important role in the development of right-wing hate content targeting the Korean community in the country.

Preventing and countering hate speech is a hugely complex undertaking that requires a nuanced appreciation of multiple interacting dimensions and actors. To counteract the phenomenon of hate that is publicly accessible via media, governments are enacting laws to limit the spread of hate speech and pushing social media sites to build strategies that reduce the dissemination of digital hatred (Mossie & Wang, 2019; Perry, 2017; Ross et al., 2017). Despite the recent spate of policies and legislation, there is a lack of consensus on what combatting hate means (Brown & Sinclair, 2019). Much available research focuses on the nature of harmful content and activity in sharing and using such content via media. However, research that empirically measures the impact on individuals and specific audience segments and the wider social impact on communities is limited. In addition, understanding the prevalence of hate speech in digital and not digital media is crucial for addressing more complex issues, such as its causes, manifestations, societal impacts, and effective solutions (Vidgen et al., 2019).

Therefore, the proposed review aims to gather, analyze, critically appraise, and synthesize empirical research about the impacts or associations of exposure to, consumption of, or active search or promotion of hate online and in traditional media, specifically on individuals, communities, and society. Results will inform policymakers and professionals working in this field about strategic countermeasures to deal with the phenomenon, identifying gaps in the literature and helping to determine future research needs.

1.2 | Defining hate speech and exposure to hate

Agreeing upon a clear and exhaustive definition of hate speech is a necessary step toward a better understanding of the phenomenon. It will also help guide this review and ensure its audience is adequately prepared to receive and interpret the findings. While there is no internationally reached consensus of what hate speech is, different entities have devised their own definitions. For example, the monitoring body of the European Commission against Racism and Intolerance (ECRI)—which has published individual country and cross-country recommendations about the phenomenon's complex nature—states that hate speech entails:

> the use of one or more particular forms of expression—namely, the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatization or threat of such person or persons and any justification of all these forms of expression—that is based on a non-exhaustive list of characteristics or status that includes "race," color, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation (ECRI, 2015, p. 16).

Meanwhile, the International Convention on the Elimination of All Forms of Racial Discrimination (CERD, 2013) understands hate speech as an utterance in direct disregard of human dignity and the core principles of human rights which seeks to undermine both individuals and societies. For the United Nations (2019), however, hate speech can be:

> any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor (p. 2).

Also important to our understanding of hate speech are the voices of experts in the field and academics who highlight the

difficulty of defining the phenomenon and identifying the conditions that allow its existence. Among these, Parekh's theorization (2012) stands out as the most relevant. According to him, hate speech is a form of communication "directed against a specified or easily identifiable individual or…a group of individuals based on an arbitrary and normatively irrelevant feature," which "stigmatizes the target group by implicitly or explicitly ascribing to it qualities widely regarded as highly undesirable," and portraying it as "an undesirable presence and a legitimate object of hostility" (pp. 40–41).

Against the numerous interpretations of hate speech, our definition draws from the UN's conception but lies in most agreement with Parekh's words. Referring to the UN definition ensures this review's alignment with the global consensus on hate speech. It also helps avoid confusion with other phenomena (e.g., legal hate speech vs. illegal hate speech). Nevertheless, recognizing the multifaceted nature and myriad manifestations of hate speech is crucial for this review as it will greatly affect the inclusion/exclusion rationale and process.

The following definitions will guide our systematic review:

Hate: "Hate" is defined by the urge to damage, humiliate, or destroy a targeted group or individual (White, 1996). Expressing hate toward the other has the objective not only to harm or be aggressive toward them but to eventually destroy the other either psychologically (e.g., via humiliation) or by literally getting rid of them (e.g., via killing and torturing) with the ultimate intention of damaging the target because of what they represent and not what or how they behave (Ben-Ze'ev, 2008; Fischer et al., 2018).

Hate Speech: While hate speech is addressed in many international and regional standard-setting documents, no internationally agreed-upon definition of hate speech currently exists. "Hate speech" in this review refers to any type of communication in speech, writing, behavior, or multimedia, that attacks or uses pejorative or discriminatory language with reference to a person or a group based on their protected characteristics, in other words, their religion, race, ethnicity, nationality, color, descent, and gender.[1] Importantly, hateful rhetoric does not target individuals themselves—as may be the case with cyberbullying—but rather expresses feelings of disdain toward a collective, even if this speech is directly directed against an individual (Blazak, 2009; Hawdon et al., 2017).

Interactions with Hate Speech: In this review, "interactions with hate speech" encompasses four descriptors: "exposure to," "consumption of," "active search for," and "promotion of" hate speech. Differentiating between these is important because some authors believe the effect differs depending on the type of interaction. For example, according to Schils and Pauwels (2014), active search leads to different outcomes than passive exposure. Therefore, for this review, "exposure to hate speech" relates to coming into contact with online or traditional media content that promotes any type of communication that attacks or uses pejorative or discriminatory utterances with reference to a person or a group, as defined above. However, "consumption of hate speech" refers to reading or viewing such content. In contrast, "active search for hate speech" refers to conducting searches of such content and "promotion of hate speech" refers to initiating, inciting, or otherwise actively supporting such content.

Hate via Media: For the purposes of this review, "hate via media" means hate or hateful content accessible by any medium of communication designed to reach the general public. Moreover, the definition of "media" comprises traditional media (i.e., newspapers, radio, or television) and online media. The latter will cover media appearing in either Web 1.0 (i.e., static HTML websites with minimal opportunities for users to interact or contribute content) or Web 2.0 (i.e., participative websites where users can interact with each other and contribute user-generated content). Furthermore, Web 2.0 will include the following types of sites or apps: social networking sites with a social, professional, business, or ideological orientation; video- or image-sharing sites; online discussion forums; wikis; blogs; multimedia messaging apps; search-and-discovery apps; any hybrids of the preceding; and sites within the deep web and dark web.

1.3 | How exposure to hate may be linked to the outcomes

The risks associated with the pervasiveness of hate materials, both on online platforms and in traditional media, have been of growing concern among scholars, decision-makers, and practitioners (Hawdon et al., 2017; Merklejn & Wiślicki, 2020; Perry, 2019; Straus, 2007; Tynes et al., 2008). Individuals may be daily exposed to hate content through media and actively consume, seek it out, or promote it. Therefore, engagement with hate could have multiple impacts or associated effects on the exposed individuals and the communities and societies to which they belong. These impacts include victimization and psychological distress (Lee & Leets, 2002; Leets, 2002; Tynes, 2006; Ybarra et al., 2008) and risk of violence (Foxman & Wolf, 2013; Hawdon et al., 2017; Kiilakoski & Oksanen, 2011). To address the complexity of these impacts, we have developed a logic model which exemplifies how hate rhetoric operates and provides insights on the potential impacts of this and related content on individuals, communities, and societies (see Figure 1). Although the logic model will inform our systematic review, we will not be bound by it. Indeed, the model may be subject to iterative refinement depending on the dynamics of the relationships between hate and its impacts as identified during the development of the systematic review.

According to research, exposure to or consumption of hate speech can have multiple consequences for individuals, particularly associated with changes in behaviors, attitudes, and emotions. These changes may involve experiencing different forms of harm and increasing the risk of engaging in hate speech or acts of hatred, including violent radicalization. From this point of view, hate speech is not only a set of utterances but also an action carrying real-world outcomes (Salminen et al., 2020). Regarding emotional impacts,

[1] https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

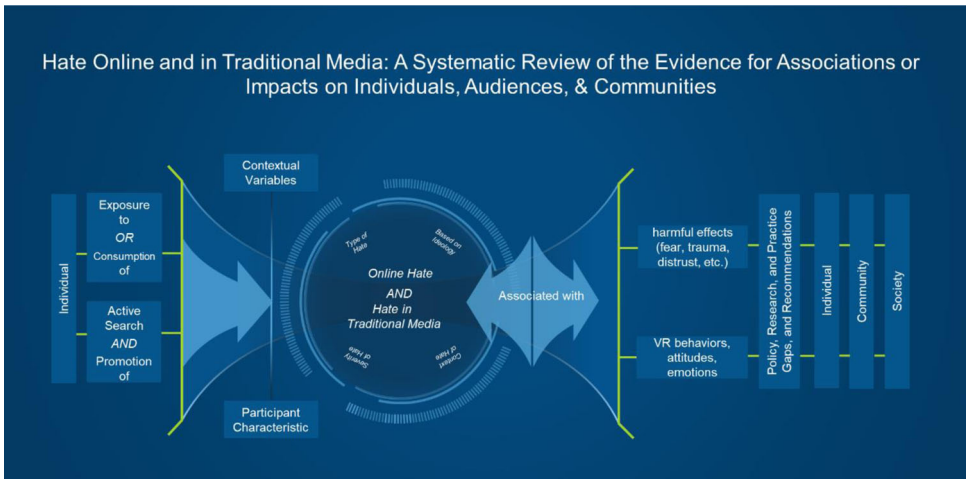

FIGURE 1 Logic model.

studies talk of anxiety and depression as being associated with exposure to online hate material. Specifically, experiences of being a target of online hate speech are linked to depression among young African Americans (Tynes et al., 2008). Also, some research claims that Muslims living in non-Muslim western societies are likely to suffer from anxiety as they anticipate online threats against them becoming a reality (Awan & Zempi, 2015). Exposure to or consumption of hate speech could also lead to changes in attitudes and then targeted actions. These may include violent acts offline resulting from the hateful rhetoric's capacity to spread antagonisms and influence motivations to take action (Moule et al., 2017). For example, some studies suggest the increased use of Facebook is associated with increased acts of violence against refugees (Müller & Schwarz, 2020). As mentioned above, hate content broadcast through traditional media such as radio has had an impact on the development and support of violence in Africa (Adelman & Suhrke, 2017; Pate & Ibrahim, 2020; Straus, 2007). Violent radicalization has also been associated with the consumption or exposure to this type of content. According to one African study, 42.5% of participants consider that hate messages could lead people to extremist positions and 27.5% to radicalization (Fayoyin, 2019). Thus, the data seems to indicate that traditional and social media could act as a breeding ground for real-world violence, often targeting vulnerable minorities (Siegel, 2020).

Over time, the individuals and the communities they identify with may become associated with the above-mentioned outcomes, potentially affecting how societies function. Hate speech is known to compromise society-level intergroup cohesion (Izsák, 2015; Tynes et al., 2008) and result in mounting feelings of hostility and unease in intergroup contexts (Adelman & Suhrke, 2017; Straus, 2007). Prolonged exposure to hateful material is known to weaken the ability to recognize hate for what it is while allowing prejudice toward the outgroup to grow (Soral et al., 2018). Whether directly or indirectly, hate speech can encourage violent acts. This is because hate speech allows individuals and groups to express negative views toward others and coordinate efforts to act them out (Siegel, 2020). Particularly in the online space, the anonymity of individuals facilitates the expression of hateful sentiments with greater zeal than in the real world. (Cohen-Almagor, 2017; Gagliardone et al., 2014; Vollhardt et al., 2007). In more extreme circumstances, hate speech can even lead to extreme violence on a societal level. Hate speeches broadcast through the radio station RTLM in Rwanda have, for example, been identified as one of the factors behind the genocide in the country (Adelman & Suhrke, 2017; Straus, 2007).

Apart from being multifaceted in terms of impacts, interactions with hate may be mediated or moderated by characteristics of hate, including the type of content, ideologies, severity, frequency, and platform type. Costello and Hawdon (2018) report a correlation between users' presence on Reddit and Tumblr and the frequency with which they may produce hateful online content. The latter is also linked to whether or not the users belong to an online community, particularly one already linked to hate speech. Furthermore, individual and contextual variables (e.g., individual traits, age, gender, socioeconomic background) can act as mediating factors when encounters with online hate are concerned. Hawdon et al. (2017), for example, cite studies in which (1) individuals who are less guarded and easily trust others are more prone to experiencing online hate than individuals with other behavioral patterns and (2) women appear to be victimized online more often than men. Unfortunately, evidence on the prevalence of hate-speech-related abuse is not always readily available (e.g., due to elective anonymity on social media or individual and group characteristics of the perpetrators being largely unknown). Moreover, it often lacks the important contextual information necessary to better understand the phenomenon of hate speech (Vidgen et al., 2019).

1.4 | Outcomes

We will analyze and synthesize data from studies retained on how people respond to interactions with hate speech and attempt to

create meaningful categories of such responses. Given the importance but relative novelty (and hence limited sources) of research on the issue, our review will be open to descriptions of all types of consequences of exposure to hate speech. However, it will also make systematic efforts to categorize them using a variety of criteria. First and foremost, we will separate objective (i.e., observable) outcomes from subjective (i.e., self-reported) ones. We also anticipate multiple categories based on who was affected by exposure to hate speech and how the reaction to this exposure manifested, as described through our logic model. The latter may include emotions (e.g., emotional unrest, fear, anger, insecurity, feeling victimized), attitudes (e.g., prejudicial attitudes, extremist positions or ideologies, normalization of violent action and rhetoric), behaviors (e.g., the perpetuation of violent action and or speech), social fracturing (discrimination, ostracization, deepening of vulnerabilities, polarization of divisions and grievances, reduction of trust in others), and other harms.

Outcomes will be assessed at the individual, community, and society levels. They will be categorized according to the measures via which the data was collected, for example, self-reported (i.e., when individuals who are interviewed report their own outcomes or perceptions of their outcomes), peer-, family-, or practitioner-reported, and the source from which it was obtained (e.g., social media, the government, and the police).

To do this, we will first identify all the outcomes found in the retained studies. Outcomes that are identified by different names but belong to the same theoretical construct will be considered to be the same outcome. This classification will be based on the operational definitions provided by the studies' authors. In case of uncertainty, the authors will be contacted to clarify these definitions. Once these primary outcomes are defined, they will be classified according to our conceptual framework explained above. This classification will be carried out by the research assistants through a coding process in which they will apply an analysis grid containing an operational definition of our preliminary categories. Most of this analysis will, however, be inductive and therefore based on the information found in the studies. As we have previously stated, our initial classification is only hypothetical and, therefore, we are open to new ways of interpreting these results from the data collected.

Our preliminary research of existing literature revealed diverse outcomes associated with people's reactions to different manifestations of hate speech. Examples of eligible outcomes identified in primary research include:

Individual-level outcomes

- Emotional impact: Saha et al. (2019) found that exposure to online hate was causally related to the expression of stress in the study's participants. In other research, Pacheco and Melhuish (2018) discovered that this type of content can give way to negative emotions such as feeling “angry,” “anxious,” “sad,” or “humiliated.”
- Behavioral changes: Pacheco and Melhuish (2018) also found that exposure to online hate speech can give way to behavioral changes, e.g., not wanting to leave one's house, inability to sleep, excessive fear of talking to strangers, and social withdrawal. Blaya and Audrin (2019) found evidence of a causal relationship between cyberhate victimization and perpetration, as well as a strong correlation between exposure to online hate and perpetration.
- Attitudinal changes: In Gaudette et al. (2020) study, extreme right-wing group members claimed that hate-related content on the Internet had been critical in reaffirming their extremist views and facilitating involvement in hate. Participants in other research stated that online hate consumption represented a gateway to offline socialization (e.g., organizing rallies and events that promoted extremist content) that, in turn, led to action (Koehler, 2014).

Community-level outcomes

- Näsi et al. (2015) measured the social trust of participants who have experienced hate speech. According to their findings, witnessing negative images and writings reduces social trust in co-workers or schoolmates, neighbors, and people in general, as well as people one has only met online.

1.5 | Why it is important to do the review

The phenomenon of hate speech, its relationship with media, and its capacity to negatively affect offline contexts have all been the focus of a wide range of studies. Researchers have demonstrated clear links between this discourse and hate crimes targeting vulnerable communities (Mossie & Wang, 2019). However, to date, there have been far fewer efforts to systematically collate and analyze existing evidence of the impacts, associations, and side effects that hate speech may have on individuals, communities, and societies. This statement is true in the case of studies that focus on the dissemination of this content online, but it is even more true in the case of studies that address its dissemination through traditional media, where there is no specific review on the subject.

This type of analysis has become particularly relevant on account of several government initiatives to restrain media platforms considered a breeding ground for violent extremism and other forms of violence (Gagliardone et al., 2016; Hussain & Saltman, 2014). In Europe, for example, the European Commission reached agreements in 2016 with the largest IT companies regarding a code of conduct obliging operators to promptly remove hate content online. SM (social media) giants such as Facebook, Twitter, and Google are also being pushed to introduce some form of effective content control (Poletti & Michieli, 2018). In a similar vein, France's proposed “Avia law” will require social media companies to promptly remove hate speech or else face fines. Meanwhile, New Zealand's “Christchurch Call” promises further development of tools that prevent users from uploading hateful content and greater transparency about how such content is being monitored and eliminated (New Zealand Government, 2019). Within the Canadian context, the government is also

standardizing the definition of hate crimes, increasing funds for training related to online hate, and engaging online platforms and service providers to better monitor and address online hate speech and promptly remove illegal or criminal content (The Standing Committee on Justice and Human Rights, 2019).

Understanding the issue and impacts of hate speech is also of utmost importance because of their relationship to freedom of speech. If free speech is to be protected, it is crucial to build a deep and accurate appreciation of the links between hate speech and the actual harm it inflicts, an approach that will help present persuasive arguments in favor of new preventive measures (Barendt, 2019). For example, Canada's Digital Charter (2019) pledges to defend the freedom of expression while holding social media platforms accountable for hate speech online at the risk of incurring penalties (Government of Canada, 2019). The United Nations' Strategy and Plan of Action on Hate Speech (2020) urges nations to provide their digital citizens with the tools to identify, reject, and defend themselves against hate speech. At the same time, it emphasizes the right to freedom of expression and the importance of education when combatting hate speech online.

In this context, a comprehensive and systematic review of the impacts and associations of these discourses could be of great value. Indeed, the undertaking could provide a fulcrum on which governments, policymakers, and tech companies might base their efforts. As such, it could help them draw policies, reframe laws and provide tools to prevent the negative consequences of hate speech.

To properly contextualize the upcoming work, we searched for prior relevant systematic reviews and meta-analyses in ERIC, Academic Search Complete, and Google Scholar, using variations of keywords and concepts relevant to our study. Similarly, we searched for existing published reviews on the subject in the Campbell Library, the Cochrane Library, and the PROSPERO registry. Our searches yielded several studies relevant to our proposed review. However, we failed to identify any other systematic review that aggregated studies posing similar research questions.

Although our research team found reviews investigating specific outcomes of hate speech among specific populations and contexts, their understandings of hate speech were narrower than ours. For example, Bliuc et al. (2018) examined ten years of research—conducted between 2005 and 2015—on cyber-racism as perpetrated by groups and individuals, synthesizing a broad range of findings, methodologies, and key research areas. Our proposed review differs from Bluic et al. in several ways. First, ours measures the effects (outcomes or associations) of these interactions on perpetrators *and* victims. Also, it is not limited to racism as a form of hate speech but includes all forms of hate speech targeted toward other protected characteristics such as gender, religion, or nationality, via both traditional and online media.

Likewise, Hassan et al. (2018) synthesized specific empirical research about the relation and effects of exposure to radical violent online material on related associations with extremist online and offline attitudes, emotions, and the risk of committing political violence. However, our proposed review will include studies on terrorist extremist speech and consumption and interaction with violent extremist content *only* when accompanied by hate speech and harmful communication toward protected characteristics and across different media types (i.e., online and traditional media channels).

Furthermore, Samari et al. (2018) systematically reviewed empirical research on Islamophobia as a form of discrimination and resulting associations between Islamophobia, health, and socio-ecological determinants of health. However, as mentioned above, our proposed review is broader in terms of targeted groups. Also, while our review is limited to hate speech delivered via media, the resulting associations will not be limited to health and socioecological determinants of health. Instead, we will include all outcomes reporting participants' resulting behaviors, attitudes, and emotions.

In conducting our research, we were able to find reviews whose understanding of hate speech was similar to ours but whose overall aims, inclusion criteria, and linguistic reach differed significantly. One such study is that of Tontodimamma et al. (2020), who carried out a broad mapping study of the conceptual structure of hate speech literature and the interactions of evolving themes over the last 30 years. To achieve these objectives, the researchers applied bibliometric measures, tools for mapping knowledge, and 'mining' techniques to identify themes via recurring patterns of words. While the authors' definition of hate speech is close to ours, their main aim was to create a bibliometric overview of the breadth and limitations of current research related to hate speech. Such themes are directly pertinent to our review. However, we aim to offer a critical and in-depth appraisal of the impacts and associations that interactions with hate—both online and through traditional media—can have on individuals and societies. Notwithstanding, Tontodimamma et al.'s study represents an important installment to the body of literature on the issue of hate speech. It also provides valuable insights from which our proposed review may benefit.

The systematic review by Paz et al. (2020) offers a critical look at the evolution and current state of English and Spanish hate speech scholarship within the fields of communication studies and legal sciences. The review analyzes the studies according to impact factor, field of study, and language. However, it focuses primarily on the studies' objectives and methodologies. The authors stress that to be comprehensive and successful, approaches to countering the effects of hate speech need to include interdisciplinary and transversal collaborations. While the review represents a starting point for mapping hate speech scholarship in different countries and within varying disciplinary, thematic, and methodological fields, its reach is significantly narrower than that of our proposed review. Thanks to the multilingual nature of our research team, we will include studies published in French, English, Arabic, Spanish, and Russian—giving our review a broader linguistic *and* cultural perspective—and will not impose a publication date limit. In addition, while we will exclude other reviews, we will broaden our reach by including mass-media sources. More importantly, our review will go beyond aggregation and comparison, examining instead a wide range of *measurable* outcomes that encounters with hate can yield in an individual or group.

Finally, the review by Poletto et al. (2020) examines how hate speech text is analyzed in the literature. The authors categorize studies on dimensions such as where the data was obtained, what type of behavior was represented, and what the annotation framework comprised. One of their key findings was a lack of consistency in hate speech data sets and taxonomies for harmful content aggregation. Given the field's heterogeneity, Poletto et al.'s review provides the research community with an up-to-date, multilingual, and multidimensional hate speech recognition resource. Our proposed review shares Poletto et al.'s concern over the need to support the research community by delivering a comprehensive, unbiased, and broad-reaching review. However, its fundamental goal in our case is to produce a synthesis of empirical evidence on the effects and impacts of hate speech on individuals and societies.

The search process has also been a path of learning for our team. For example, Waqas et al.'s (2019) mapping and scientometric analysis of research trends and hotspots in online hate research will help contextualize and tighten our own search strategy, especially with respect to hate speech in cyberspace. Moreover, Waqas et al.'s findings regarding prevalent themes of research in the field, the focus of these publications, which countries and journals seem to publish the most on this topic, and which organizations fund them will be closely looked at, taken note of, and reviewed to make sure that our inclusion criteria are sufficiently comprehensive to cover the many different analyses of online hate research publications. Employing the thus far described approach promises to result in a systematic review that is exhaustive and precise.

In summary, our review aims to give practitioners and researchers a better understanding of how hateful rhetoric operates. It will also provide evidence that may help those tasked with creating preventive measures to efficiently counter hate in an integrated manner at a time when both independently and in collaboration, nations are making concerted efforts to contain the phenomenon of hate speech. Furthermore, the review will inform policy-makers and professionals working in the field about existing strategic countermeasures to deal with the phenomenon, identify gaps in the literature, and help determine future research needs.

2 | OBJECTIVES

The current review will gather, critically appraise, and synthesize empirical research about the impacts or associations of exposure to, consumption of, active search for, or promotion of hate speech via media, at the individual, community, and societal levels.

The general objectives of this review are as follows:

1. to critically and systematically synthesize the empirical evidence on the effects or impacts of exposure to or consumption, active search, or promotion of hate content online or in traditional media;
2. to describe how the characteristics of hate (e.g., type of content, ideologies, severity, type of platform) impact the documented effects;
3. to collect and identify the role of contextual variables (e.g., individual traits, age, gender, socioeconomic background) on the documented effects;
4. to collect and produce a meaningful classification of outcomes; and
5. to identify gaps and limitations in the research and related policy documents.

3 | METHODS

3.1 | Criteria for considering studies for this review

3.1.1 | Types of studies

We will include any empirical study published up to 31 December 2021 using primary data and quantitative measures that let us establish an impact and/or association relationship between exposure to or consumption of hate content on different platforms (online and traditional media) and the consequences it may have on individuals, communities, and society. For a study to be included, it has to analyze hateful communication that stigmatizes individuals or groups because of their protected characteristics (i.e., religion, ethnicity, nationality, race, color, descent, or gender). We will also include quantitative sections of studies using mixed methods that allow us to establish these relationships. Therefore, any qualitative study or any section of a study using qualitative methods, as well as studies based exclusively on basic descriptive statistics for single variables, opinion-based, or theoretical studies or studies without primary data will be excluded from this study.

We will collect data from studies employing experimental or quasi-experimental (non-randomized) designs as well as cross-sectional and longitudinal studies with or without control groups. Given that we will gather and analyze data from studies that measure the impact of these contents at the individual, community, and societal levels and that these contents are publicly available in different media, the relationship between these contents and their impacts are not always studied through experimental designs especially in a recent field such as the case of hate studies. Given the latter point, Caudy et al. (2016) recommend, for example, in the case of criminological studies, the use of experimental and observational designs in meta-analyses. The impact of online hate speech is usually studied in natural environments such as social media or through surveys of people already exposed to this type of content. Therefore, focusing only on experimental designs could restrict and exclude studies that analyze data relevant to this systematic review. Thus, in the case of cross-sectional designs we will include high-quality studies that employ suitable statistics such as multivariate regression and bivariate correlation models, and provide sufficient information to calculate effect sizes. Studies based on a single sample

may include cross-sectional and longitudinal correlational studies in which an association is clearly established between observed variability in exposure to or consumption of hate content and one or more outcomes of interest, as well as studies comparing a group before and after exposure to or consumption of hate content (intervention group only with pretest and posttest design, interrupted-time-series studies, etc.). We will also include non-randomized studies with two or more comparative groups (groups exposed to or consuming this content and control groups) in which at least one outcome is compared between the two groups (prospective and retrospective case-control studies, comparative studies with posttest only design, controlled before-and-after study, etc.).

We will not impose any other restrictions on study design or date because the state of the literature is such that doing so could lead to the inclusion of only a very small number of studies that do not give a clear picture of what is being done in the field.

Regarding sampling methods used in the primary studies included in this systematic review, a wide range of methods will be considered acceptable as long as the sample provides enough information to make inferences about the study's intended population.

3.1.2 | Types of participants

To better comprehend the heterogeneous impacts of hateful media content across individual, community, and society levels, we will place no limit on population or individual characteristics of study participants.

3.1.3 | Types of exposure to hate

The studies included in this review will examine hate speech delivered through any media designed to reach the general public. We will include two types of media: traditional mass media (i.e., newspapers, radio, or television) and online media. The latter will cover media appearing in either Web 1.0 (i.e., static HTML websites with minimal opportunities for users to interact or contribute content) or Web 2.0 (i.e., participative websites where users can interact with each other and contribute user-generated content). Furthermore, Web 2.0 will include the following types of sites or apps: social networking sites with a social, professional, business, or ideological orientation; video- or image-sharing sites; online discussion forums; wikis; blogs; multimedia messaging apps; search-and-discovery apps; any hybrids of the preceding; and sites within the deep web and dark web.

3.1.4 | Types of outcome measures

The outcomes of interest in this review are the measurable effects of individuals' interactions with hate speech via the media. They may include self-reported measures (i.e., when individuals interviewed report their own outcomes or their perception of them) and measures reported by peers, family, or professionals, along with measures reported by governments, law enforcement agencies, and open-source generated data. Attention will also be paid to where the data was sourced. This list is, however, not exhaustive. Other outcome measures may be discovered in the literature and will be included to the extent that they meet the inclusion criteria of the studies.

3.1.5 | Types of settings

We will not limit the inclusion of studies with regard to settings.

3.1.6 | Language of studies

We will include any documents written in English, French, Arabic, Spanish, and Russian (languages spoken by the research team members).

3.1.7 | Exclusion criteria

The following types of studies will be excluded:

- Any qualitative study or any section of a mixed-method study using qualitative methods.
- Studies based exclusively on basic descriptive statistics.
- Systematic reviews and literature reviews.
- Descriptive, opinion, and theoretical documents on the subject.
- Studies on programs that aim to prevent hate online or offline.
- Cyberbullying studies that do not analyze hate communication that stigmatizes an individual or group based on their protected characteristics (i.e., religion, ethnicity, nationality, race, color, ancestry, or gender).

3.1.8 | Example of studies that might be eligible for inclusion in the review

Our preliminary research of existing literature revealed several studies to be included in this systematic review. Lee-Won et al. (2020), for example, investigated the effects of hate messages from multiple sources on a group of 172 African Americans on Twitter. To do so, they devised an online experiment in which these participants were randomly divided into three groups: those who were exposed to hate tweets from multiple sources, those who were exposed to hate tweets (of identical content) from a single source, and those who were exposed to non-hate tweets (control). Emotional distress was measured after being

exposed to this content. The data was analyzed using an analysis of covariance. The results showed that hate tweets from multiple sources, compared to identical single-source messages and non-hate tweets, elicited greater emotional distress.

Saha et al. (2019) is another example of a study that can be included in this systematic review. This study quantified the degree of hate exposure of individuals across 6 million comments on 174 college Reddit communities and the degree of online stress of these same individuals. A causal inference framework was utilized to better understand the psychological effects online hate speech has. These comments were temporally partitioned in 2016 so that before this date the measures were considered as baseline stress and a baseline of hate exposure. On this basis the authors divided individuals into two cohorts of college Reddit users, where one had a history of exposure to hate speech (treatment group), and one did not (control group). They observed that compared to their baseline stress, the stress level of individuals in the treatment group was higher than those in the control group.

3.2 | Search methods for identification of studies

3.2.1 | Electronic searches

To locate relevant research literature, we will employ a professional librarian to assist in developing an effective search. The searches will target online hate, with groups of keywords formed around the concepts of (1) Hate, (2) Expression of Hate, (3) Hate Source and (4) Media Environment. Where possible, proximity operators will be used to closely link the concepts of hate and expression of hate. Here, for example, is the final search that will be tested in the PsycINFO database:

HATE[2]: ((Hatred OR Hate OR Dangerous OR Fanatic* OR Prejudic* OR Intoleran* OR Bias OR Violen* OR Negative OR Stigmat* OR Discriminat* OR Bigot* OR Hostil* OR "desire to destroy" OR "desire to damage" OR oppress* OR Abuse OR Abusi* OR "desire to kill" OR humiliat* OR Intimidat* OR Terrori*)

NEAR/2

EXPRESSIONS OF HATE: (Crime* OR Speech OR Incident* OR Conduct OR Act OR Acts OR Abuse* OR Vilif* OR Language OR Harass* OR word* OR express* OR comment* OR defam* OR slur* OR troll* OR flaming OR "flame war" OR Incite* OR Insult* OR "call* for" OR derog*))

AND

HATE SOURCE: (Radical* OR Indoctrinat* OR Fundamentalis* OR "Homegrown Terror" OR Terror* OR Eco-terror* OR Al Qaida OR ISIS OR Anti-Capitalis* OR Extremis* OR Supremacis* OR Nationali* OR "Homegrown Threat*" OR Jihad* OR "White Power" OR "Neo-Nazi" OR "Right Wing" OR "righ-wing" OR "Left Wing" OR "left-wing" OR Nativis* OR "Anti-Immigra*" OR "Ecological Violence" OR "Anti-Capitalis*" OR Islamophob* OR "Alt-Right" OR Antifa* OR Incel OR Homophob* OR discriminati* OR Transphobi* OR Antisemiti* OR Anti-Semiti* OR Mysogyn* OR Xenophob* OR homophob* OR anifemini* OR racis* OR stereotyp* OR sexis* OR gender OR "sex* identit*" OR LGBT)

AND

MEDIA ENVIRONMENT: AB (Website* OR "Information System*" OR "Electronic Communication*" OR Online OR "Social Media" OR "World Wide Web" OR "Web 2.0" OR Internet OR Virtual OR Cyber OR Website* OR Digital OR "Computer Media*" OR Bebo OR Facebook OR Flickr OR Foursquare OR Friendster OR Hulu OR Instagram OR LinkedIn OR Meetup OR Pinterest OR Reddit OR Snapchat OR Tumblr OR Xing OR Twitter OR Yelp OR Youtube OR TikTok OR Photolog OR Telegram OR WhatsApp OR Messenger OR Twitch OR Discord OR Gab OR "Chat Room*" OR "Online Forum*" OR "Discussion Forum*" OR "Videogame Communit*" OR "Dark Web" OR "Dark Net" OR "Deep Web" OR "Computer Crime" OR "Anonymity Network*" OR "Image Board*" OR 4chan OR 8chan OR meme* OR Media OR Communication* OR Telecommunication* OR Radio OR Television OR Newspaper* OR News OR Cinema OR Movie* OR Journalis* OR Theatre OR Music OR Video*)

Most keywords will be searched for in any field, except for groups of keywords related to the concept of Media Environment. In this case, we will search for media environment only in the Abstract field. There are two reasons for this. First, the use of social media in research is increasing in several areas of knowledge, and the potential number of studies identified in the initial step of the search could make this systematic review impracticable due to the limited resources available to us. Second, the media environment is a key aspect of this study, and the studies to be included in this review should emphasize at the outset that they examine hate speech delivered through any media designed to reach the general public.

Next, we will conduct searches in a variety of bibliographic databases, both subject-specific databases and general multidisciplinary databases. The proposed list is as follows: Academic Search Complete (EBSCO), Communication Abstracts (EBSCO), Communication and Mass Media Complete (EBSCO), Criminal Justice Abstracts (EBSCO), ERIC (EBSCO), Medline (EBSCO), NCJRS, ProQuest Central, ProQuest Dissertations and Theses Global, PsycINFO (APA PsycNet), Social Services Abstracts (ProQuest), Sociological Abstracts (ProQuest), SocINDEX (EBSCO), and the Web of Science platform's core collection (SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, and ESCI). While the searches will employ standard Boolean logic, they will be tailored to the features of each database, making use of available controlled vocabulary and employing proximity operators where possible.

[2]In some contexts, such as the UN Human Rights Council resolutions, the concept of hate speech is avoided and replaced by terms such as stigmatization, discrimination, incitement, and spread of discrimination and prejudice, or incitement of hatred due to its ambiguity and multifaceted nature.

TABLE 1 List of gray literature sources.

ADL (Anti-Defamation League)
Alternative to Violence Project
Article 19
Bricks Against Hate Speech
CHRC (Canadian Human Rights Commission)
Council on Foreign Relations
Counter Narratives
Dangerous Speech
eMORE (Monitoring and Reporting Online Hate Speech in Europe)
Equality and Human Rights Commission
German National Center for Crime Prevention
Global center on cooperative security
Global Kids Online
Hatebase
Hedayah
HRMI (Human Resource Management Institute)
Human Rights Watch
IFEX (International Freedom of Expression Exchange)
ILGA Europe (The International Lesbian, Gay, Bisexual, Trans and Intersex Association)
INACH (International Network Against Cyber Hate)
INAR (Irish Network Against Racism)
International Centre for Counter-Terrorism—The Hague (ICCT)
International network for hate studies
ISD (The Institute for Strategic Dialogue)
ISTSS (International Society for Traumatic Stress)
Kaichid Dialogue Centre
MANDOLA (Monitoring and Detecting Online Hate)
MediaSmarts
Minority Rights Group International
Moonshot
OHCHR (Office of the United Nations High Commissioner for Human Rights)
Online Antisemitism Task Force
OPHI (The Online Hate Prevention Institute)
OSCE (Organization for Security and Cooperation in Europe)
Partners against hate
Report It
Research Outreach
Tech Transparency Project
The Alan Turing Institute
The Council of Europe
UiO C-REX—Center for Research on Extremism
UK Home Office Research Database
UK Safer Internet Centre
UNAOC (United Nations Alliance of Civilizations)
UNCCT
UNESDOC
UNICRI (United Nations Interregional Crime and Justice Research Institute)
UNODC
US National Criminal Justice Reference Service
Project SOMEONE
No hate speech movement

3.2.2 | Searching other resources

The searches of bibliographic databases will be supplemented with a targeted search for gray literature, using the Google search engine as the primary tool. We will also check the OpenGrey collection, as well as the websites of various governments and nongovernmental organizations (see Table 1). International groups/publications will be a particular focus of this stage, and we will use the Google advanced search form to limit the results by target regions.

The websites and proceedings of any identified academic conferences considered relevant will be scanned. We will also conduct citation searches of prior reviews in the subject area (e.g., Bliuc et al., 2018; Hassan et al., 2018; Waqas et al., 2019).

Finally, to track down possible remaining studies, we will employ a "branching" procedure in which we carefully examine the reference sections of articles we have already retrieved.

3.3 | Data collection and analysis

3.3.1 | Selection of studies

Selecting admissible evidence studies will be performed by two research assistants who will independently screen the abstracts of the total number of studies identified in the literature search, compare decisions, then document, discuss and resolve disagreements. In this first step, research assistants will answer four questions to assess the eligibility of studies:

1. Do the studies directly address hate speech as it has been defined in this protocol?
2. Do the studies measure exposure to or consumption of hate speech through a media source?
3. Do the studies use any empirical primary data?

4. Do the studies analyze any impact or association of this hate content on individuals, communities or society?

All questions can be answered through one of three alternatives: "yes," "no," or "maybe." If one of the criteria is not met (alternative "no"), the study will be excluded at this stage. Thus, all studies that are at least considered to *maybe* meet all criteria will be selected for full-text screening. Research assistants will also check for duplicate sources. In this first step, Fleiss' kappa (Fleiss, 1971) will be computed to ensure adequate inter-rater agreement. Once the inter-rater agreement for the selection of studies is confirmed, two coders will also review and cross-review the rest of the documents for final eligibility. During this initial coding, if the study abstract is not available, the assistants will make the initial decision by reading the study introduction.

During full-text screening, the assistants will confirm that the studies meet the four initial criteria, which is often not easy at the initial stage, as well as the full eligibility criteria described earlier in the protocol. All studies categorized as "maybe" should be transformed in this way into "yes" or "no." In addition, they will also have to confirm that the hate content addresses protected characteristics of individuals or groups, that the studies are based on eligible study designs, that the studies provide a bivariate or multivariate analysis of this association, and that it is not a duplicate source. If all of these are confirmed, the selected studies will be coded in their entirety.

Lastly, the PRISMA (http://www.prisma-statement.org) template will be used to record the results of the literature searches in a flowchart.

3.3.2 | Data extraction and management

A coding sheet (see Supporting Information: Appendix B) including the following criteria for data extraction will be created for each study:

- Reference information: Document ID, Study Title, Study Author(s), Publication Year, Place published or accessed with URL, Reference Type, Coding References.
- Study details: Country of Study, Language, Date of Research, Peer reviewed, Funded research, Conflicts of interest, Ethical Issues.
- Methodology: Type of study, Sample constitution procedure, Country/place of recruitment, Sample characteristics, Source of hate speech measure, Source of outcome measure, Quantitative measures on the link between exposure and outcome, Qualitative measures on the link between exposure and outcome.
- Independent Variable Details: Perpetrator of Hate Speech, Target of Hate Speech, Type of Hate Speech, Participants' interaction with hate speech (Exposure to hate, Consumption of hate, Active search of hate, Promotion of hate), Hate Speech medium (Traditional media, Online media).
- Dependent variable details: Participants' outcome after interaction with hate speech (Type of measured outcome: Mental health symptoms, Emotions, Attitudes. Behavior, etc.).
- Interaction variables measured: that is, interaction, confounding, or moderating variables that influence the relationship between the independent and dependent variables. Quantitative results on the link between exposure and outcome: Statistical results such as effect size or any other statistic that allows us to calculate this effect size as well as unexpected outcomes including study harms.
- Qualitative results on the link between exposure and outcome: Relevant qualitative results that allow the analysis, interpretation, or contextualization of the results obtained by the studies. Authors' Conclusion and Recommendations (policy, research, practice).
- Study limitations and ethical issues.

A "summary of evidence" table will be drawn once all the data is gathered.

3.3.3 | Measures of association

We will conduct a meta-analysis of the results for each association between exposure to or consumption of hate content and possible outcome categories only if a minimum of two studies yield effect sizes of the same outcome construct. Studies using experimental designs and observational designs will be meta-analyzed separately as well as those establishing causal relationships and those analyzing other types of associations. Most of the studies examined in this review will consider one independent variable (exposure to or consumption of hate content) and one or plus dependant variables. To calculate effect sizes that adequately indicate the strength of relationships between two comparable variables of interest, summary statistics (predominantly, correlation coefficients, but also relevant information from regression models, etc.) will be extracted (Borenstein, Hedges, et al., 2009). This systematic review involves several potential measures of association that need to be clarified to carry out a meta-analysis of their findings.

- For studies that use continuous variables to compare independent groups or pre-post scores or matched groups, we will calculate standardized mean differences (Cohen's *d* or Hedge's *g*), which will mostly be calculated using means, standard deviations, and sample sizes (Borenstein, Hedges, et al., 2009).
- In the case where we find binary dependent variables, we will use a log odds ratio as the measure of effect size (Lipsey & Wilson, 2001).
- For studies using bivariate correlations and where the correlation matrix derived from multivariate models could be obtained, we will convert Pearson's r to Fisher's z to calculate the effect

size. As pointed out by Lipsey and Wilson (2001), the absence of the correlation matrix could limit this analysis. In such a case we will first contact the authors of the study to obtain it or we will use the available information to calculate the effect sizes, provided that the study includes adequate descriptive statistics (Aloe & Thompson, 2013). This information will be calculated using the formulas proposed by Lipsey and Wilson (2001).

Following the Campbell Collaboration guidelines (Polanin & Snilstveit, 2016), if multiple effect size metrics are identified among the retained studies, we will use the most common effect size metric and convert all other effect sizes to that metric, using conventional formulas. Finally, these effect sizes will be analyzed with the Comprehensive Meta-Analysis™ program according to the random effects analytical model.

3.3.4 | Assessment of risk of bias in included studies

Both Cochrane and the Campbell Collaboration caution against inadequate quality assessment of studies that originate in areas of research characterized by heterogeneity of design, tools, samples, and outcomes. Accordingly, this review will use a risk of bias assessment tool with which each study will be coded separately (see Supporting Information: Appendix A for a description of this tool). Given that this systematic review will include both experimental and nonexperimental studies, we have chosen the Mixed Methods Appraisal Tool (Hong et al., 2018). This tool was initially developed to evaluate mixed methods studies; however, it includes two sections to individually evaluate quantitative randomized controlled trials and quantitative non-randomized studies, which are the two main design types included in this review. This tool will allow a thorough exploration of the issues (properly executed randomization, comparability of groups at baseline, representativeness of the sample, taking into account confounding factors, etc.) that the selection of studies in this review may present. The quality assessment of the studies will be carried out by two assistants who have been previously trained in this assessment. In case of disagreement in the evaluations, the researcher in charge of the process will make the final decision. The assessment of the risk of bias in the included studies will be summarized in a table including the results obtained per study (Yes, No, I can't tell) for each variable of the tool used. These results will then be outlined in detail.

3.3.5 | Unit of analysis issues

Units of analysis in this review will include individuals, communities, or groups who have had interactions with hate speech via media. All units of analysis may be considered for inclusion and be classified accordingly. Before synthesizing our findings, any article reporting on the same data set will be clearly linked. In this case, only the most recent and, if possible, the peer-reviewed version will be kept in the systematic review. For multi-arm studies, groups will be merged or divided into separate associations, all the while avoiding any double-counting of participants. Data of outcomes provided in multiple metrics will be pooled into independent categories. Given the broad range of studies that will likely be included, papers will be grouped to account for any potential dependencies. Furthermore, measures will be taken to rule out any double-counting of evidence. In rare cases when the repeated use of the same sample in the same outcome category is unavoidable, we will use robust variance estimate adjustment as recommended in Hedges et al. (2010).

A meta-analysis requires that the effect sizes of the studies be independent of each other. However, this situation is not always met, particularly in the case of more than one outcome and more than one measure over time. In this case, following Borenstein, Hedges, et al. (2009), our first choice will be to consider only one effect size per study or to aggregate the effect sizes of the outcomes based on the same construct (mean of the outcomes as the unit of analysis). In case there is more than one outcome at different time points, the procedure is relatively similar, with the difference that, instead of the mean, we will consider the difference between the effect sizes of the different time points.

3.3.6 | Dealing with missing data

In the absence of metrics mentioned in the "Measures of association" section, we will use, sample sizes, *t*- or *F*-test scores, and associated *p* values. In case missing data prevent us from calculating the effect size, we will try to contact the authors of the studies.

3.3.7 | Assessment of heterogeneity

As proposed by Borenstein, Cooper, et al. (2009), we will assess heterogeneity using I^2, τ^2, and *Q* statistics. Some moderators could also have an impact on the relation between exposure to or consumption of hate through the media and the effects on participants. Previous research has pointed, for example, to variables such as age, gender or sexual orientation, and ethnicity as moderators of this association (Baines et al., 2010; Saha et al., 2019). We will perform moderator analysis using these variables to measure their effect on heterogeneity. This list of moderators is not exhaustive. If, during the analysis, other moderators become relevant, they may also be used.

3.3.8 | Assessment of reporting biases

If enough quantitative data are collected to enable a meta-analysis, we will carry out all standard procedures of dealing with the dependency issue (see "Unit of analysis issues" for this topic), publication bias, and sensitivity analyses as outlined in the methodology literature (e.g., R. M. Bernard, Borokhovski, et al., 2014;

Borenstein, Hedges, et al., 2009; Cooper, 2017; Hedges et al., 2010; Rothstein et al., 2005) using *Comprehensive Meta-Analysis* software (Borenstein et al., 2014).

Publication bias analysis within a meta-analysis deals with the concern that some nonsignificant effects, null-effects, or even effects on the opposite from the observed weighted average effect size pole of the distribution are missing from the analysis. If studies with these effects were to be found and included, they would nullify the observed effect.

To reduce "publication bias", we will include both articles published in peer-reviewed scientific journals and gray literature. We will find the latter by searching the Web using branching techniques for studies, reports from government and nongovernment organizations, conference proceedings, and other relevant documents. Furthermore, the websites of organizations working in the subject area of hateful speech via media will be manually searched for additional materials (see "Search methods for identification of studies" section). Documents written in English, French, Arabic, Spanish, and Russian—all languages spoken by the research team members—will be included. Of note, our systematic review will exclude other systematic reviews and literature reviews. This review will also exclude descriptive, opinion, and theoretical documents on the subject, including mass-media sources (even if they refer to secondary scientific findings) as well as studies on programs or interventions that aim to counter hate online or offline.

We will analyze publication bias using Duval and Tweedie's Trim and Fill analysis and funnel plots, assuming that studies with small sample sizes will be less published (Borenstein, Hedges, et al., 2009).

Sensitivity analysis will be described in a specific section below.

3.3.9 | Data synthesis

If one or more meta-analyses are possible, then the data synthesis will be done by presenting the results of the calculation of effect sizes for each association between exposure to or consumption of hate speech and each expected outcome or construct that synthesizes several similar outcomes. First, the outcomes found will be listed and then, if there are at least two effect sizes, a forest plot will be used for each association analyzed, showing the confidence interval (95%) and the results of the effect size metrics used by each study, as well as the overall results.

In the case of studies using bivariate correlations, a frequent problem in a meta-analysis is synthesizing bivariate and partial effect sizes from different multiple regression studies since both effect sizes estimate different parameters (Aloe et al., 2016). We expect that different studies will use different sets of covariates and thus estimate the partial effects of a predictor variable under different conditions. One option is to conduct two separate meta-analyses for bivariate and partial effect sizes (Aloe et al., 2016). However, following Wolfowicz et al. (2019), we chose to synthesize them into a single meta-analysis and perform a moderator analysis, if data is available, to evaluate these effects. In this case, we will give preference to bivariate estimates in our analysis, and use multivariate effect sizes only if the former are not available, which is the usual practice in criminological studies (Wolfowicz et al., 2019). As recommended by Aloe et al. (2016), we will perform a meta-regression that includes covariates reflecting the differences in complexity and structure of the model.

If a meta-analysis is not possible, we will draw preliminary recommendations and integration of the accumulated findings using a narrative synthesis method (Moher et al., 2009). Attention will be given to the potential intensity of the association between hate speech exposure, hate speech precursors, and its outcomes. The collected data will be synthesized based on the following dimensions:

1. The main findings of the literature search;
2. The reliability of extracted data (i.e., the Quality of Study Assessment tools will be used to qualitatively ascertain the data's robustness);
3. The relevancy and generalizability of findings

The gaps in existing research and the limitations of knowledge will be presented in both cases.

3.3.10 | Sensitivity analysis

For quantitative studies within a meta-analysis (if warranted), standard procedures of sensitivity analysis (e.g., *CMA* routine "One study removed") will be implemented to identify those effects that disproportionally affect the overall findings (i.e., to locate outliers). Usually, those are the effects of atypically high magnitude (either positive or negative) that also are based on much larger than average in size samples and need to be treated to prevent misinterpretation of population parameters. We anticipate removing these atypical studies.

ROLES AND RESPONSIBILITIES

- Content: Pablo Madriaza, Ghayda Hassan, Sébastien Brouillette-Alarie
- Systematic review methods: Pablo Madriaza, Eugene Borokhovski, Ghayda Hassan, Sébastien Brouillette-Alarie,
- Statistical analysis: Pablo Madriaza, Sébastien Brouillette-Alarie, Eugene Borokhovski,
- Information retrieval: David Pickup, Eugene Borokhovski
- Scientific Writing and Editing: Wynnpaul Varela
- Data Coding: Wynnpaul Varela, Melina Girard, Loïc Durocher-Corfa, Emmanuel Danis

SOURCES OF SUPPORT

The research team has received funding from the Public Safety Canada and approval to move forward with the review.

DECLARATIONS OF INTEREST

The research team has no potential conflict of interest for this review.

PRELIMINARY TIMEFRAME

Approximate date for submission of the systematic review is 31 March 2022*.

*The Date is subject to revision pending Campbell peer-review approval for protocol and progress of the systematic review team.

PLANS FOR UPDATING THE REVIEW

Dr. Ghayda Hassan and her research team will be responsible for updating this review, every 2 years after publication date.

REFERENCES

Adelman, H., & Suhrke, A. (Eds.). (2017). *The path of a genocide: The Rwanda crisis from Uganda to Zaire*. Transaction Publishers.

Aloe, A. M., Tanner-Smith, E. E., Becker, B. J., & Wilson, D. B. (2016). Synthesizing bivariate and partial effect sizes. *Campbell Systematic Reviews*, *12*(1), 1–9. https://doi.org/10.4073/cmpn.2016.2

Aloe, A. M., & Thompson, C. G. (2013). The synthesis of partial effect sizes. *Journal of the Society for Social Work and Research*, *4*(4), 390–405. https://doi.org/10.5243/jsswr.2013.24

Awan, I., & Zempi, I. (2015). *We fear for our lives: Offline and online experiences of anti-Muslim hostility*. https://www.tandis.odihr.pl/bitstream/20.500.12389/22288/1/08624.pdf

Baines, P. R., O'Shaughnessy, N. J., Moloney, K., Richards, B., Butler, S., & Gill, M. (2010). The dark side of political marketing: Islamist propaganda, Reversal Theory and British Muslims. *European Journal of Marketing*, *44*(3–4), 478–495.

Barendt, E. (2019). What is the harm of hate speech? *Ethical Theory and Moral Practice*, *22*(3), 539–553. https://doi.org/10.1007/s10677-019-10002-0

Ben-Ze'ev, A. (2008). Hating the one you love. *Philosophia*, *36*(3), 277–283. https://doi.org/10.1007/s11406-007-9108-2

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education*, *26*(1), 87–122. https://doi.org/10.1007/s12528-013-9077-3

Blaya, C., & Audrin, C. (2019). Toward an understanding of the characteristics of secondary school cyberhate perpetrators. *Frontiers in Education*, *4*, 46. https://doi.org/10.3389/feduc.2019.00046

Blazak, R. (2009). Toward a working definition of hate groups. *Hate Crimes*, *3*, 133–162.

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

Borenstein, M., Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 279–293). Russell Sage, Inc.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Wiley.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2014). *Comprehensive meta-analysis Version 3.3.070*. Biostat.

Brown, A., & Sinclair, A. (2019). *The politics of hate speech laws*. Routledge.

Caudy, M. S., Taxman, F. S., Tang, L., & Watson, C. (2016). Evidence mapping to advance justice practice. In D. Weisburd, D. Farrington, & C. Gill (Eds.), *What works in crime prevention and rehabilitation: Lessons from systematic reviews* (pp. 261–290). Springer.

Cohen-Almagor, R. (2017). Why confronting the internet's dark side? *Philosophia*, *45*(3), 919–929. https://doi.org/10.1007/s11406-015-9658-7

Cooper, H. (2017). *Research synthesis and meta-analysis* (5th ed.). Sage Publications.

Costello, M., & Hawdon, J. (2018). Who are the online extremists among us? Sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. *Violence and Gender*, *5*(1), 55–60. https://doi.org/10.1089/vio.2017.0048

Davidson, J., Livingstone, S., Jenkins, S., Gekoski, A., Choak, C., Ike, T., & Phillips, K. (2019). *Adult online hate, harassment and abuse: A rapid evidence assessment*. UK Council for Internet Safety. http://eprints.lse.ac.uk/103230/1/Livingstone_adult_online_hate_published.pdf

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. Proceedings of the First Italian Conference on Cybersecurity (ITASEC. 17), pp. 86–95. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/811450/Adult_Online_Harms_Report_2019.pdf

European Commission against Racism and Intolerance (2015). *ECRI General Policy Recommendation No. 15*, 16. https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01

Fayoyin, A. (2019). Online radicalisation and Africa's youth: Implications for peacebuilding programmes. In E. K. Ngwainmbi (Ed.), *Media in the global context* (pp. 23–46). Springer International Publishing. https://doi.org/10.1007/978-3-030-26450-5_2

Fischer, A., Halperin, E., Canetti, D., & Jasini, A. (2018). Why we hate. *Emotion Review*, *10*(4), 309–320. https://doi.org/10.1177/1754073917751229

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Foxman, A. H., & Wolf, C. (2013). *Viral hate: Containing its spread on the Internet*. Macmillan.

Gagliardone, I., Patel, A., & Pohjonen, M. (2014). *Mapping and analysing hate speech online: Opportunities and challenges for Ethiopia*. https://eprints.soas.ac.uk/id/eprint/30573

Gagliardone, I., Pohjonen, M., Beyene, Z., Zerai, A., Aynekulu, G., Bekalu, M., Bright, J., Moges, M. A., Seifu, M., Stremlau, N., Taflan, P., Gebrewolde, T. M., & Teferra, Z. (2016). *Mechachal: Online debates and elections in Ethiopia—From hate speech to engagement in social media*. SSRN. https://doi.org/10.2139/ssrn.2831369

Gaudette, T., Scrivens, R., & Venkatesh, V. (2020). The role of the internet in facilitating violent extremism: Insights from former right-wing extremists. *Terrorism and Political Violence*, 1–18. https://doi.org/10.1080/09546553.2020.1784147

Government of Canada. (2019). *Canada's Digital Charter in Plan by Canadians, for Canadians Action: A*. https://www.ic.gc.ca/eic/site/062.nsf/vwapj/Digitalcharter_Report_EN.pdf/$file/Digitalcharter_Report_EN.pdf

Hassan, G., Brouillette-Alarie, S., Alava, S., Frau-Meigs, D., Lavoie, L., Fetiu, A., Varela, W., Borokhovski, E., Venkatesh, V., Rousseau, C., & Sieckelinck, S. (2018). Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence. *International Journal of Developmental Science*, *12*(1–2), 71–88. https://doi.org/10.3233/DEV-170233

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. https://doi.org/10.1002/jrsm.5

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., & Nicolau, B. (2018). *Mixed methods appraisal tool (MMAT) version 2018—User guide*. McGill University.

Hussain, G., & Saltman, E. M. (2014). *Jihad trending: A comprehensive analysis of online extremism and how to counter it*. Quilliam. https://preventviolentextremism.info/jihad-trending-comprehensive-analysis-online-extremism-and-how-counter-it

Izsák, R. (2015). *Hate speech and incitement to hatred against minorities in the media*. UN Humans Rights Council. A/HRC/28/64. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/000/32/PDF/G1500032.pdf?OpenElement

Kiilakoski, T., & Oksanen, A. (2011). Soundtrack of the school shootings: Cultural script, music and male rage. *Young*, *19*(3), 247–269. https://doi.org/10.1177/110330881101900301

Koehler, D. (2014). The radical online: Individual radicalization processes and the role of the Internet. *Journal for Deradicalization*, *1*, 116–134. https://journals.sfu.ca/jd/index.php/jd/article/view/8

Lee, E., & Leets, E. (2002). Persuasive storytelling by hate groups online. *American Behavioral Scientist*, *45*(6), 927–957. https://doi.org/10.1177/0002764202045006003

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-Semitism and anti-gay speech. *Journal of Social Issues*, *58*(2), 341–361. https://doi.org/10.1111/1540-4560.00264

Lee-Won, R. J., White, T. N., Song, H., Lee, J. Y., & Smith, M. R. (2020). Source magnification of cyberhate: Affective and cognitive effects of multiple-source hate messages on target group members. *Media Psychology*, *23*(5), 603–624. https://doi.org/10.1080/15213269.2019.1612760

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications.

United Nations. (2019). *United Nations strategy and plan of action on hate speech*. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

Merklejn, I., & Wiślicki, J. (2020). Hate speech and the polarization of Japanese National Newspapers. *Social Science Japan Journal*, *23*(2), 259–279. https://doi.org/10.1093/ssjj/jyaa015

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mossie, Z., & Wang, J. H. (2019). Vulnerable community identification using hate speech detection on social media. *Information Processing and Management*, *57*(3). https://doi.org/10.1016/j.ipm.2019.102087

Moule, R. K., Decker, S. H., & Pyrooz, D. C. (2017). Technology and conflict: Group processes and collective violence in the internet era. *Crime, law, and social change*, *68*(1), 47–73. https://doi.org/10.1007/s10611-016-9661-3

Müller, K., & Schwarz, C. (2020). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*. https://doi.org/10.1093/jeea/jvaa045

Näsi, M., Räsänen, P., Hawdon, J., Holkeri, E., & Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. *Information Technology & People*, *28*(3), 607–622. https://doi.org/10.1108/ITP-09-2014-0198

New Zealand Government. (2019). *Christchurch Call to eliminate terrorist and violent extremist online content adopted*. https://www.beehive.govt.nz/release/christchurch-call-eliminate-terrorist-and-violent-extremist-online-content-adopted

Noriega, C. A., & Iribarren, F. J. (2012). Toward an empirical analysis of hate speech on commercial talk radio. *Harvard Journal of Hispanic Policy*, *25*, 69.

Pacheco, E., & Melhuish, N. (2018). *Online hate speech: A survey on personal experiences and exposure among adult New Zealanders*. Retrieved from SSRN https://doi.org/10.2139/ssrn.3272148

Parekh, B. (2012). Is there a case for banning hate speech? In M. Herz, P. Molnar (Eds.), (pp. 37–56). Cambridge University Press.

Pate, U. A., & Ibrahim, A. M. (2020). Fake news, hate speech and Nigeria's struggle for democratic consolidation: A conceptual review. In A. M. G. Solo (Ed.), *Advances in human and social aspects of technology* (pp. 89–112). IGI Global. https://doi.org/10.4018/978-1-7998-0377-5.ch006

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. *Sage Open*, *10*(4). https://doi.org/10.1177/2158244020973022

Perry, J. (2017). Ireland in an international comparative context. In A. Haynes, J. Schweppe, & S. Taylor (Eds.), *Critical perspectives on hate crime* (pp. 93–107). Palgrave Macmillan.

Perry, B. (2019). Breaking the peace: The Quebec City terrorist attack. In I. Zempi & I. Awan (Eds.), *The Routledge International Handbook of Islamophobia* (pp. 275–285). Routledge.

Perry, B., Mirrlees, T., & Scrivens, R. (2017). The dangers of porous borders: The "Trump Effect" in Canada. *Journal of Hate Studies*, *14*, 53–76

Polanin, J. R., & Snilstveit, B. (2016). Campbell methods policy note on converting between effect sizes *(Version 1.1, updated December 2016)*. The Campbell Collaboration.

Poletti, C., & Michieli, M. (2018). Smart cities, social media platforms and security: Online content regulation as a site of controversy and conflict. *City, Territory and Architecture*, *5*(1), 1–14. https://doi.org/10.1186/s40410-018-0096-2

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 2020. https://doi.org/10.1007/s10579-020-09502-8

Rabah, J. (2014). Gendered subjectivities and cyberspace dialogues in Lebanon: A critical discourse analysis. In V. Venkatesh, J. Wallin, J. C. Castro, & J. E. Lewis (Eds.), *Educational, psychological, and behavioral considerations in niche online communities* (pp. 101–111). IGI Global. https://doi.org/10.4018/978-1-4666-5206-4.ch007

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated CommunicationBochumer Linguistische Arbeitsberichte*, *17*(6–9). https://doi.org/10.17185/duepublico/42132

Rothstein, H. R., Sutton, A. J. & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and psychological effects of hateful speech in online college communities. *Proceedings of the 10th ACM Conference on Web Science*, *2019*, 255–264.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, *10*(1), 1–34. https://doi.org/10.1186/s13673-019-0205-6

Samari, G., Alcala, H., & Sharif, M. Z. (2018). Islamophobia, health, and public health: A systematic literature review. *AJPH Research*, *108*(6), 1–9. https://doi.org/10.2105/AJPH.2018.304402

Schils, N., & Pauwels, L. (2014). *How invariant is the interaction between extremist propensity and exposure to extremist moral settings in sub groups by gender and immigrant background? Testing a leading hypothesis of Situation Action Theory*. Presented at the ECS Meeting, Prague, the Czech Republic. http://hdl.handle.net/1854/LU-5969780

Siegel, A. (2020). Online Hate Speech. In N. Persily, & J. Tucker (Eds.), *Social media and democracy: The state of the field, prospects for reform* (pp. 56–88). Cambridge University Press.

Somerville, K. (2011). Violence, hate speech and inflammatory broadcasting in Kenya: The problems of definition and identification. *Ecquid Novi:*

African Journalism Studies, *32*(1), 82–101. https://doi.org/10.1080/02560054.2011.545568

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Straus, S. (2007). What Is the relationship between hate radio and violence? Rethinking rwanda's "Radio Machete." *Politics & Society*, *35*(4), 609–637. https://doi.org/10.1177/0032329207308181

The Standing Committee on Justice and Human Rights. (2019). *Taking action to end hate: Report of the Standing Committee on Justice and Human Rights*. https://www.ourcommons.ca/Content/Committee/421/JUST/Reports/RP10581008/justrp29/justrp29-e.pdf

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2020). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*, 157–179. https://doi.org/10.1007/s11192-020-03737-6

Tynes, B. M. (2006). Children, adolescents, and the culture of online hate. In N. Dowd, D. Singer, & R. F. Wilson (Eds.), *Handbook of children, culture, and violence* (pp. 267–289). Sage.

Tynes, B. M., Giang, M. T., Williams, D. R., & Thompson, G. N. (2008). Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health*, *43*(6), 565–569. https://doi.org/10.1016/j.jadohealth.2008.08.021

UN Committee on the Elimination of Racial Discrimination (CERD). (2013). *General recommendation No. 35: Combating racist hate speech*. https://www.refworld.org/docid/53f457db4.html

Vidgen, B., Margetts, H., & Hassis, A. (2019). *How much online abuse is there? A systematic review of evidence for the UK*. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf

Vollhardt, J., Coutin, M., Staub, E., Weiss, G., & Deflander, J. (2007). Deconstructing hate speech in the DRC: A psychological media sensitization campaign. *Journal of Hate Studies*, *5*(15), 15–35. https://doi.org/10.33972/jhs.40

Waqas, A., Salminen, J., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS One*, *14*(9). https://doi.org/10.1371/journal.pone.0222194

Weber, M., Viehmann, C., Ziegele, M., & Schemer, C. (2020). Online hate does not stay online—How implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior. *Computers in Human Behavior*, *104*(106–192). https://doi.org/10.1016/j.chb.2019.106192

White, R. S. (1996). Psychoanalytic process and interactive phenomena. *Journal of the American Psychoanalytic Association*, *44*(3), 699–722. https://doi.org/10.1177/000306519604400303

Wolfowicz, M., Litmanovitz, Y., Weisburd, D., & Hasisi, B. (2019). A field-wide systematic review and meta-analysis of putative risk and protective factors for radicalization outcomes. *Journal of Quantitative Criminology*. https://doi.org/10.1007/s10940-019-09439-4

Ybarra, M. L., Diener-West, M., Markow, D., Leaf, P. J., Hamburger, M., & Boxer, P. (2008). Linkages between internet and other media violence with seriously violent behavior by youth. *Pediatrics*, *122*(5), 929–937. https://doi.org/10.1542/peds.2007-3377

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hassan, G., Rabah, J., Madriaza, P., Brouillette-Alarie, S., Borokhovski, E., Pickup, D., Varela, W., Girard, M., Durocher-Costa, L., & Danis, E. (2022). PROTOCOL: Hate online and in traditional media: A systematic review of the evidence for associations or impacts on individuals, audiences and communities. *Campbell Systematic Reviews*, 18, e1245. https://doi.org/10.1002/cl2.1245