Research Needs for Countering Extremist Hate

This white paper identifies open research needs that practitioners outside academia face in combating extremist hate. We group these research needs into six themes and present a list of specific projects for future research. Our goal with this document is to orient researchers across disciplines toward research questions with the potential for translational impact in countering extremist hate.

June 2022

Michael Miller Yoder
Postdoctoral Fellow
Collaboratory Against Hate
Carnegie Mellon University

Hana Habib
Postdoctoral Fellow
Collaboratory Against Hate
Carnegie Mellon University

The Collaboratory Against Hate: Research and Action Center at Carnegie Mellon University and the University of Pittsburgh aspires to develop and support innovative multidisciplinary, interdisciplinary, and cross-university research aimed at understanding how extremist hate is generated, how it circulates in online and real-life spaces, and how it polarizes society and provokes harmful and illegal acts, especially toward communities of color and other minoritized groups. We seek to develop effective interventions to inhibit every stage in the creation and growth of extremist hate groups and to minimize their destructive consequences.

www.collabagainsthate.org

1. Introduction

Countering extremist hate based on race, ethnicity, religion, gender identity, sexual orientation, and other identity-based prejudices requires input from stakeholders across society. These stakeholders include governments, technology companies, nonprofit organizations, and researchers across academic disciplines. Research in law, sociology, psychology, communication and public health is needed to understand the human and social factors that give rise to and sustain extremist hate movements. Research in computer science, data science, media studies and education is needed to track and thwart the spread of hate in online and offline contexts. However, this research is often siloed by discipline, impeding progress toward shared insights and best practices. Researchers are not always aware of the practical challenges practitioners face when working against extremist hate in non-academic contexts, such as nonprofit organizations and tech companies.

This white paper seeks to orient researchers of the Collaboratory Against Hate (CAH) toward open research needs faced by practitioners. To do so, we draw on interviews with practitioners from nonprofit organizations and tech companies who work on initiatives addressing hate. From our conversations, we identified the most pressing research needs across organizations, grouped into themes (Section 2). We also compiled a cross-disciplinary list of specific research projects that arose from these interviews (Section 3). In communicating these research ideas, we hope to spur conversation and collaboration among CAH researchers and lead to contributions with translational impact in combating extremist hate.

2. Research Themes

We identify six key themes for future research that address the challenges faced by practitioner interviewees. These cross-disciplinary themes are: 1) evaluating programs and interventions, 2) developing tools to track and detect hate activity, 3) studying the impact of hate, 4) communicating to the public and educating public audiences, 5) informing tech company policy, and 6) building practitioner resources. We present each theme below, ordered from the most commonly mentioned areas of need to more specific themes.

2.1 Program and intervention evaluation

Practitioners across tech companies and nonprofit organizations most often mentioned a need for research **evaluating whether programs or interventions designed to counter hate are actually successful in doing so**.

From nonprofit practitioners, this included measuring changes in individuals' and communities' understanding and behavior. For example, how do practitioners know if interventions have successfully prevented extremist violence? Does programming increase participants' understanding of how oppressions such as antisemitism, racism, and homophobia are connected? At the community scale, what changes can be detected after anti-hate education initiatives such as Holocaust education? Are these communities better prepared to respond to hateful propaganda? The impact of anti-hate programming on communities with marginalized identities, such as Black people, indigenous people, and people of color (BIPOC), was of particular interest. For example, studying the changes that can occur in conversations in BIPOC-only affinity spaces can inform guidelines for conversations that feel safe for participants. Given the drastic impact of the COVID-19 pandemic in switching nonprofit programming online, interviewees also had questions about how to lead effective virtual events on global issues, including extremism and antisemitism.

Similarly, practitioners at tech companies spoke to the difficulty of measuring the impact of the interventions their platforms have explored to address extremist hate. Research is needed to determine the impact of techniques beyond removing content or banning

accounts, such as changes in how algorithms prioritize content. Are these techniques more or less successful than simple account removal? How successful are interstitials that "nudge" users to modify posts that contain hateful stereotypes? Practitioners also mentioned the need for evaluating the effectiveness of counter-speech in addressing hate speech on their platforms.

Part of the difficulty of evaluating interventions comes from finding evaluation measures. This can be difficult because projects are often trying to measure intangible outcomes: how effective interventions are at *avoiding* radicalization or hateful content.

2.2 Detection and tracking

Another need is for research that produces **tools and datasets for detecting hate speech and tracking hate groups**, particularly in multimedia contexts and contexts outside the U.S.

Research is needed to accurately and comprehensively track beliefs and incidents at a large scale. This includes measuring public beliefs in misinformation and aggregating data on hate crimes across local and national organizations. Practitioners often mentioned challenges in overcoming self-report bias in collecting data on incidents of hate and bias.

Research detecting and tracking hate outside of Western, English-speaking contexts is sorely needed. There is a lack of data on hate groups outside the U.S. and Europe. Practitioners from tech companies noted challenges in finding annotators and datasets to develop tools such as hate speech detection systems outside of English and other European languages.

Difficulties in detecting hate in non-textual content were commonly noted. Tracking extremist narratives and hate speech in audio is especially crucial as podcasts often expose mainstream audiences to hateful content. Tools that identify hateful content in images and memes are also needed.

Since the in-group symbols of hate groups are constantly evolving, finding novel language use for ideologies of hate is another challenge. Practitioners from tech companies expressed that it was helpful when outside groups tracked new problematic narratives and informed them before they spread on their platforms, though practitioners at nonprofit organizations noted that this takes considerable energy and time. Academic researchers may be well positioned to support this effort. Documenting extremism in the military and law enforcement was also mentioned as a need.

2.3 Impact of hate

Practitioners often identified a need for research on **what personal and societal factors contribute to hateful extremism, and how hate speech, hateful activism, and organizing around hate impact individuals and societies.**

Interviewees in both nonprofit organizations and tech companies noted that research is needed to identify patterns that lead to extremist violence at personal and societal levels. For example, why do some members of hate communities commit violent acts while many others, consuming the same narratives, do not? What are the signs that narratives with the potential to inspire coordinated violence cross over to do so? At the extreme end of this, what are the patterns in social media activity that indicate a society on the brink of genocide?

Hate has an impact beyond violence. Some practitioners noted a need to measure other negative consequences of hate. Conversely, what are the benefits of increased prosocial behaviors on platforms?

Research on the connections between hate movements, both geographic and ideologically, is also needed. For example, one interviewee noted the lack of basic knowledge disseminated about hate groups outside the U.S. This information includes who is spreading hate in specific countries, what their main narratives are, how they are connected to violence, and how they are connected to groups outside of their own country. Other practitioners noted the need for research on connections between hate directed at different targets, such as between antisemitism and other forms of racial violence.

Finally, basic information is needed about the impact of different types of hate. For example, what is the impact of hate-filled misinformation beyond the number of impressions on social media? How does exposure to misinformation translate to belief or acceptance?

2.4 Education and communication to the public

Practitioners at nonprofit organizations often mentioned a need for **research on effective public communication and educational curricula related to hate and extremism.**

For example, what communication forms have the most impact from anti-hate nonprofit organizations (reports, statistics, social media communications, or others)? Another important research area is developing guidelines for conversations among members of the public. How can conversations be structured for people to move past their own perspectives into another's point of view? Beyond a change in understanding, practitioners were interested in how conversations can lead to action. Guidance for families speaking to members who might be drawn to extremist beliefs is also needed. Public talks about extremism often elevate privileged people as "experts." How can this be reformulated to elevate the voices of people who have not been afforded that status?

Education was noted across multiple conversations as a critical means to counter extremist hate; effective educational resources are needed. Though such topics are not often discussed in the classroom, even the youngest students are not sheltered from identity-based hatred and violence. How can educators be empowered to help students make sense of this violence? On the education policy side, how is ethnic studies education best implemented? In the university context, how are students best informed about why acts of bias, such as using racial slurs while "joking," are harmful?

2.5 Tech company policy

Research on the effectiveness, implementation, and transparency of the policies that tech companies have on addressing hateful content is also needed.

First is a need for outside groups to monitor that platforms actually address content that violates their policies, especially in languages other than English. For each piece of a platform's policy, how well are they removing or otherwise censuring content? Even if policies are carried out, how much impact does this have on reducing the spread of hateful content? There is also a tension between developing policies that reduce hateful content but allow for legitimate discussion to take place.

Practitioners at tech companies themselves were interested in how to make these policies more flexible. Policies can address specific extremist narratives such as QAnon, but how can they proactively prepare for the next hateful narrative? Others mentioned a need to limit bigotry and stereotypes from users who knew the rules so well that they could tailor their hateful messages to not explicitly violate any pieces of policy.

Some challenges regarded policies related to transparency. Interviewees wondered about the transparency reports published by tech companies. Beyond legal or self-regulatory requirements, how can these reports be useful for user experience, or as actual measures in mitigating the influence of hate on these platforms? Others called for making AI-assisted content moderation more transparent. Tools to generate explanations to users for why machine-moderated content was removed are needed. What are best practices to represent uncertainty in automatic moderation decisions?

2.6 Shared practitioner resources and definitions

Practitioners mentioned a need for **resources that synthesized terminology and best practices across various disciplines that address hate and violent extremism.**

This included the need to synthesize existing research into actionable guidelines to make them more accessible to a larger audience. Furthermore, the creation of dashboards and centralized hubs to aggregate disparate resources would be beneficial to practitioners.

A main difficulty in countering violent extremism and hate is one of definitions and taxonomies. Concepts such as "hate speech," for example, are amorphous and encompass a family of related meanings. This challenge was raised by practitioners from both nonprofit and tech companies. First is the issue of labeling what counts as "hate" or "hate groups." Practitioners generally had to define where they drew the lines themselves. Research could help explain the functional meanings of terms across organizations, sectors, and disciplines.

Identifying a common lexicon in the violence prevention space, itself multidisciplinary, was mentioned as a need. Interviewees from tech companies expressed the challenge in sorting content into useful taxonomies of abusive/hateful material that matches the needs of regulators or their own product teams. Beyond the practical issue of what content qualifies within certain categories, this difficulty in defining terms and agreeing on classifications was noted as hindering understanding between groups.

3. Initial Project List

The practitioners we spoke with described ideas across various disciplines for specific research projects that would aid their work in combating hate and extremism. The full table of these projects with categorization details (as described in Section 4) and filtering features is available at collabagainsthate.org/opportunities/project-opportunities and will be updated periodically as CAH discusses new projects with outside organizations. CAH aims to support its affiliated researchers in their work addressing these practitioner needs, such as through cultivating partnerships with relevant organizations and disseminating research findings to practitioner audiences. We provide a brief description of the initial set of projects discussed in our practitioner conversations below. If you would like to spearhead or become involved with one of these projects, please email us at CAH-info@lists.andrew.cmu.edu.

Initial list of project ideas from practitioner conversations

Extending prior misinformation research to hate/extremism	Researchers at Google recently conducted large-scale public surveys to investigate how misinformation can spread and evolve globally. While the initial surveys were focused on COVID-19 misinformation, there is potential to create similar surveys on misinformation fueling hate and extremist activity, modeled on this work.
Evaluating the framing of platform initiatives combating hate	Meta recently worked with the World Jewish Congress and UNESCO to develop a tool to direct people to accurate information about the Holocaust. As a platform intervention to increase education and counter hateful content, it would be useful to research how such initiatives can be framed for greatest impact.
Informing the Global Minds curriculum	Global Minds is the leading youth program at the World Affairs Council of Pittsburgh. As a youth-driven global education program, it aims to connect youth of immigrant backgrounds with those born in America in order to promote better cross-cultural understanding. The program includes conversation prompts and activities to help achieve this goal. Global Minds is interested in having researchers create and evaluate educational materials that can be incorporated into their curriculum.
Synthesizing prior research into guidance documents for philanthropists and violence prevention practitioners	Systematic reviews compiled by the Campbell Collaboration and the Canadian Practitioners Network provide immense insight into the effectiveness of strategies and programs for combating hate-based violence. However, further synthesis of these reports is required to better communicate to different audiences. There is a need for researchers to compile guidance documents including evidence-based approaches for practitioners in hate-based violence prevention, as well as guidance for funders supporting violence prevention efforts.
Detecting toxicity in memes	The spread of hateful messaging online increasingly occurs through visual media such as memes. There is a need for researchers to build machine learning models that can score the toxicity of text and image elements, as well as their combination. Ideally, these models should be able to "translate" the author's intended message in sharing the meme.
Identifying in-group terms	People in hateful communities quickly adopt terms and use in-group vocabularies to express themselves. A tool has been developed for matching vocabulary sets across communities, which researchers may be able to use to find novel and in-group terms in online communities.
Identifying central nodes in hate networks	Prior work has identified that misinformation content on Twitter originates from relatively few accounts; networks of hate and extremism likely function in the same way. Removing or demoting content from these accounts could have a major impact on the spread of hateful ideologies. There is a need for researchers to identify key proliferators of hate who serve as central nodes in such networks.

Identifying identity terms	Identity terms are very important in detecting hate speech and other hateful content, but should also be treated with care to avoid bias. There is a need for researchers to identify terms that refer to protected or marginalized identities across languages and contexts.
Evaluating the impact of authorship feedback interstitials	Many online platforms have integrated machine learning in moderation features to provide feedback to users when their post might contain toxic speech or be in violation of community guidelines. While authorship feedback has proven to be effective in nudging users to modify their posts, there is little understanding of what modifications are being made and if the resulting posts are any less toxic in nature. There is a need for researchers to evaluate the impact of authorship feedback interstitials on user comments, and the resulting overall toxicity on the platform.
Developing aftercare recommendations for former white supremacists and extremists	Reformed individuals who denounce hateful and extremist ideologies require a support network to prevent "relapse." However, no validated set of best practices or standardized program of care for supporting such individuals currently exists. There is a need for researchers to develop recommendations that can be incorporated into the aftercare of former white supremacists and extremists.
Evaluating the impact of information source in flagging dangerous or unverified content	Social media platforms have implemented efforts to flag dangerous or unverified content to their users. However, the information source (e.g., the platform itself, nonprofit organization, news entity, academic research) provided for producing these labels could impact users' perceptions and interactions with the labeled content. There is a need for researchers to explore which information source labels are most effective in warning users about dangerous or unverified content that they may encounter.
Creating a centralized hub for efforts in violence prevention	The field of hate-based violence prevention is transdisciplinary, involving mental health professionals, nonprofit organizations, law enforcement and other government agencies. As a result, communicating developed materials and resources across disciplines can be challenging. There is a need for a centralized hub to connect distributed efforts in violence prevention.
Identifying effective language for explaining violent extremism to families	One challenge practitioners working in violence prevention face is communicating concepts related to violent extremism to the families of individuals at risk of following dangerous ideologies. There is a need for researchers to explore effective language that practitioners can use to educate families and help them support at-risk individuals.
Performing a comparative analysis of transparency reports produced by tech companies	Major tech companies publish annual reports to provide transparency for issues impacting users of their platform, including content moderation efforts. However the utility of such reports for civil society, policymakers, and tech companies themselves has yet to be analyzed. There is a need for researchers to compare transparency reports published by different tech companies to identify what, if any, gaps exist in current reporting and how these reports could be more impactful.
Improving transparency about content moderation and algorithmic decisions	Major tech companies have policies against hate speech and actively work to remove it from their platform. However, decisions related to how and why content is removed or demoted can be nontransparent and may be perceived as silencing users' voices. There is a need for researchers to explore how to improve existing transparency efforts or identify new means of communicating content moderation decisions to users.
Analyzing the impact of dangerous content on attitudes and behavior	Despite moderation efforts, dangerous and hateful content persists on online platforms. It is unclear how interacting with this content shapes users' beliefs and interactions. For example, does interacting with this content introduce users to dangerous viewpoints or simply reaffirm their existing beliefs? How does individual behavior on the platform change if hateful content was removed? Would users actively attempt to seek out more of that content, or simply return to "mainstream" behavior?
Systematically reviewing global publications on hate and extremist groups	The academic literature on hate and extremist groups available in the public domain is largely constrained to those published in English. However, there is much to be learned from publications written in other languages about hate and extremist activity. There is a need for researchers to conduct a systematic review of the academic literature published in non-English venues to make this work globally accessible.

4. Interview details and analysis

To identify practitioners for our interviews, we initially reached out to individuals working at tech companies and nonprofit organizations who have had prior conversations with CAH leadership. Many of these individuals were available for an interview, while others connected us to other practitioners within their network. Additionally, we identified advocacy organizations that have ongoing efforts related to hate against identity groups that were not represented in our initial contact list (e.g., anti-Muslim, anti-LGBTQ+ hate). In total, we spoke with 15 individuals during 12 conversation sessions. Seven individuals we spoke with were affiliated with major tech companies, while the remaining were affiliated with nonprofit organizations. The practitioners we spoke with held varying roles within their organization, including those related to research, policy management, and executive leadership. All organizations were based in the U.S.

We developed an interview script to guide our conversations with practitioners. This script included prompts to ask about their organization's past, current, and planned future objectives related to combating hate and extremism, challenges faced in implementing those initiatives, as well as research opportunities related to extremist hate that would aid their work. Our conversations were held in February and March 2022. Each interview was attended by two CAH postdoctoral researchers (the authors of this document). One researcher guided the conversation while the other primarily took notes on the discussion.

We conducted a thematic analysis of our conversation notes to systematically identify research opportunities relevant to CAH members. We first highlighted statements in our notes that pertained to a broader research challenge and used affinity diagramming to surface common themes that emerged across our conversations. We also referenced our notes to compile a set of specific research projects that were of interest to practitioners we spoke with.

We categorized projects according to:

- **Duration**: time estimate for the project (summer/semester or multi-semester)
- **Education level:** appropriate educational background for students working on the project (undergraduate or graduate)
- **Skills/discipline:** relevant skills or disciplines for the project
- **Level of definition:** how defined the project's research questions are as described by the practitioner (low, medium, high)
- **Resources/data available:** whether data or other resources such as funding may exist for the project
- **Project partner:** the organization (if any) that is willing to partner on the project
- **Theme:** the broader research theme that the project aligns with

Projects can be filtered according to these categorizations at collabagainsthate.org/opportunities/project-opportunities.

5. Conclusion

The transdisciplinary nature of the Collaboratory Against Hate provides the potential to address real challenges faced by practitioners outside of academia combating hate and extremism. From speaking to these practitioners, we identified outstanding research opportunities that could make a translational impact. These research needs fit into six themes: 1) evaluating programs and interventions, 2) developing tools to track and detect hate activity, 3) studying the impact of hate activities, 4) communicating to and educating public audiences, 5) informing tech company policy, and 6) building shared practitioner resources and definitions. We also propose a set of research projects for students and project teams across disciplines. We hope identifying these research needs will direct CAH researchers toward work that addresses challenges faced by practitioners.

We thank the practitioners we spoke with for conversations about challenges they face in their work. This included Dr. Heidi Beirich, Co-founder of the Global Project Against Hate and Extremism; Nick Haberman, Director of the LIGHT Education Initiative; Dr. Terrence Mitchell, Vice President, Diversity, Equity, and Inclusion at PennWest University; as well as practitioners at the Holocaust Center of Pittsburgh, Google, Jigsaw, the McCain Institute for International Leadership, Meta, Stop AAPI Hate, Twitter, the World Affairs Council of Pittsburgh, and other organizations.

Thanks to Lorrie Cranor and Kathleen Blee for guidance on this project. Thanks to Tim Kelly for design work and Lydia Yoder for editing help.