Measuring and Mitigating the Impact of Platform Algorithms and Policies on Online Radicalization

PPC Summer Scholar-in-Residence Proposal for June 1 - 26, 2020

Rishab Nithyanand
Assistant Professor
Computer Science
201L MLH
rishab-nithyanand@uiowa.edu

Brian Ekdale
Associate Professor
School of Journalism & Mass Communication
E324 AJB
brian-ekdale@uiowa.edu

Abstract: Over the past decade, the shortcomings of internal policies and proprietary algorithms employed by online platforms have become apparent. These platforms can be exploited for the purpose of sowing dissent, injecting instability into our social structures, and facilitating new recruiting pipelines for extremist and violent communities. Accordingly, there is an urgent need to develop scalable methodologies to measure and mitigate the impact of platform algorithms and policies on the online radicalization process and to use empirical evidence to suggest policies that governing bodies can adopt to guide and regulate the practices of social platforms. The overall objective of this project is to understand how current regulations, internal policies, and algorithm deployments facilitate the promotion and adoption of extremist ideologies and develop interventions to mitigate their impact. Our central hypothesis is that the current policies and algorithms employed by Internet authorities and governing bodies, Internet platforms, and offline media are actively being used to facilitate online radicalization and that scalable user-level interventions are possible to mitigate the impacts of such actions. The rationale underlying the proposed research is that once user-level interventions have been demonstrated to be effective at curbing the spread of extremist discourse without harming economic incentives, online platforms will be more willing to adopt platform-wide policies and interventions to prevent mass manipulation of their users. These interventions and policy proposals are expected to have a positive impact on the development and deployment of algorithms and policies which promote civil discourse and inclusiveness.
Project Summary

Over the past decade, the shortcomings of internal policies and proprietary algorithms employed by online platforms have become apparent – social platforms have become ripe with hateful and violent speech despite efforts to moderate discourse [2-7], platforms using algorithmic personalization have been found to enforce filter bubbles [8-10] which can lead to real-world consequences [11-12], platforms with unhelpfully broad content policies have discriminated against and silenced minority communities [13-14], and the Internet as a whole is reeling from the high prevalence of disinformation and misinformation [15-19]. These developments have enabled platforms to be exploited for the purpose of sowing dissent, injecting instability into our social structures, and facilitating new recruiting pipelines for extremist and violent communities. At the heart of the problem is our lack of understanding of how internal policies and proprietary algorithmic systems geared to maximize user interaction might be exploited to encourage the adoption of extremist or radical ideologies. Accordingly, there is an urgent need (1) to develop scalable methodologies to measure and mitigate the impact of platform algorithms and policies on the online radicalization process and (2) to use empirical evidence to suggest policies that governing bodies can adopt to guide and regulate the practices of social platforms. Not meeting these needs will allow the Internet and popular online platforms to continue contributing to the promotion and adoption of extremist and violent ideologies which pose a threat to the stability of our social structures.

Our long-term goal is to facilitate the deployment of algorithms and policies which promote civil discourse and inclusiveness while simultaneously achieving social platforms’ intended economic goals of increased user interaction and creativity. The overall objective of this project is to understand how current regulations, platform policies, and algorithm deployments facilitate the promotion and adoption of extremist ideologies and develop interventions to mitigate their impact. Our central hypothesis is that: (1) the current policies and algorithms employed by Internet authorities and governing bodies (e.g., ICANN, national governments), Internet platforms (e.g., YouTube, Facebook, Reddit, 4chan), and offline media (e.g., mainstream television and radio) are actively being used to facilitate online radicalization and (2) scalable user-level interventions are possible to mitigate the impacts of such actions. This hypothesis was formulated, in part, on the basis of our own preliminary and other previous work which: (1) shows evidence of proprietary algorithms and policies being manipulated to promote extremist ideologies and (2) that the exploitation of these vulnerabilities and their impact on specific users can often be predicted ahead of time – suggesting room for proactive user-level interventions and governing body regulations that can mitigate the impact of their exploitation. The rationale underlying the proposed research is that, once the effectiveness of user-level interventions has been demonstrated to be effective at curbing the spread of extremist discourse without harming economic incentives, online platforms will be more willing to adopt platform-wide policies and interventions to prevent mass manipulation of their users.

We plan to attain the overall objective of this proposal by pursuing the following specific aims:

1. **Modeling radicalization and deradicalization processes in online forums.** Our working hypothesis is that by using the criteria laid out by law enforcement and threat assessment research, we can identify extremist users of online forums and using their public forum interaction histories, develop reasonably accurate models of their radicalization (and, if available,
their subsequent deradicalization). Such a model will allow us to understand what types of content and interactions are effective at radicalizing and deradicalizing extremists online.

2. **Developing methodologies to measure and curb radicalization caused by algorithmic personalization.** Our *working hypothesis* is that current methodologies are not effective at accurately identifying the promotion of extremist content by personalization algorithms and that methodologies need to account for the impact of a variety of factors such as: prior user-platform interaction metrics, off-platform history, and method of platform usage. Combining our comprehensive measurement methodologies, which incorporate these features, with our understanding of the online radicalization process (obtained in Aim 1) will suggest interventions to reduce the possibility of radicalization due to algorithmic personalization. Further, by actively monitoring the behavior of users of our preliminary tools, we can theorize the impact of similar platform-wide policies on user-interaction and economic metrics.

3. **Measuring and curbing the incidence of cybersquatting-based disinformation spread.** Our *working hypothesis* is that cybersquatters are actively leveraging existing trust in mainstream media brands to spread disinformation (e.g., cbsnews.com.co was actively used to spread political and extremist propaganda during the 2016 US Presidential elections by leveraging the Facebook platform). By developing mechanisms to identify and report the occurrences of such squatting to end-users and media brands, we can reduce their influence on end-users.

4. **Modeling narrative flows using provenance graphs to curb mainstream media's amplification of extremist narratives.** Our *working hypothesis* is that provenance graphs can be used to model the flow of narratives across online platforms and into the mainstream media. We anticipate the development of tools which are able to visualize these graphs will allow journalists and the public to be more informed on the origins of specific narratives associated with extremist ideologies (e.g., the framing of ‘migrant caravans’ of immigrants originated on the 4chan discussion board before being amplified by mainstream media outlets), therefore promoting more educated consumption of narratives.

The proposed research is *creative and original* because it employs a holistic approach to measure and mitigate the impact of proprietary algorithms, internal policies, and governing body regulations on user radicalization by (1) considering discussion board moderation, personalization systems, cybersquatting, and information flows from online to offline mainstream media; and (2) focusing on the development of effective and economically viable user-level interventions and governing body policy proposals. Regarding expected outcomes, in addition to developing models and methodologies to measure, understand, and mitigate the role of platform algorithms and policies on the spread of extremist ideologies, we anticipate releasing several ready-to-deploy tools (e.g., as browser extensions), usable by users of popular online platforms, aimed at reducing the spread of extremist content without compromising economic incentives for online platforms (e.g., by not resorting to content censorship). These outcomes are expected to have a *positive impact* on the development and deployment of algorithms and policies which promote civil discourse and inclusiveness.
Relevant Prior Work and Value of Collaboration

Professors Nithyanand and Ekdale currently lead an interdisciplinary research group dedicated to the study of social media algorithms. The Algorithms and Social Media research group, which is supported by the Obermann Center and the Moeller Research Lab at the University of Iowa, is composed of faculty and graduate students at the University of Iowa, Penn State University, and Gonzaga University. The group’s first research study, “Measuring Political Personalization of Google News Search,” was presented at the 2019 World Wide Web Conference (WWW’19) [19], and it has an additional manuscript currently under review at *New Media & Society*, the top journal in media and communication studies [20].

Both Nithyanand and Ekdale have published extensively in their respective fields on the subject of digital and social media. Nithyanand’s research agenda is focused on developing measurement methodologies to uncover the workings of opaque Internet ecosystems. He has published research on methodologies to uncover online discrimination, privacy-compromising tactics, and online censorship by technology corporations and state governments. Ekdale’s research agenda is focused on studying digital media culture from a global perspective. He has published research on a variety of digital media producers (e.g., bloggers, journalists, filmmakers, social media influencers, and more) working in a number of different national contexts (e.g., United States, Kenya, Haiti, and Indonesia).

Although we regularly collaborate as leaders of the Algorithms and Social Media working group, the proposed research project is much larger and more ambitious than any of our previous work. In order to submit a competitive grant proposal, we need to have dedicated time and space to work together on designing the research study and writing the grant proposal. The PPC’s Summer Scholar-in-Residence program would provide an incredible opportunity to strengthen our collaborative relationship.

Project Deliverables and Timeline

Our primary goal for the Summer Scholar-in-Residence program is to develop a proposal for submission to the National Science Foundation’s Cyber-Human Systems program (NSF: CISE: IIS: CHS). Because the final grant proposal is due in September 2020, we want to have a full draft written by the end of June so we can (1) hire an external copy editor to assist with revisions and (2) share the draft with senior colleagues at peer universities who have volunteered to review and provide feedback on the grant proposal. In August, we will make final revisions to the proposal in time for the September submission deadline.

If awarded an NSF grant, we will begin data collection immediately. Following data collection and analysis, we will compose manuscripts highlighting our methodologies and findings for conference and journal submission. Potential venues for publishing these manuscripts include the Web Conference (WWW), the Internet Measurement Conference (IMC), the Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), the International Conference on Web and Social Media (ICWSM), International Communication Association (ICA), *New Media & Society; Information, Communication & Society;* and *Journal of Computer Mediated Communication*. We will also share our results with elected officials and relevant policy makers at Internet governing bodies.
References


RISHAB NITHYANAND

https://sparta.cs.uiowa.edu/people/rishab
rishab-nithyanand@uiowa.edu
201L, MacLean Hall
The University of Iowa

RESEARCH INTERESTS

My research focuses bringing transparency to the online ecosystems which impact our sociopolitical realities. Specifically, my work seeks to discover and circumvent the entities involved in facilitating online tracking, manipulation, surveillance, and censorship of Internet users.

EDUCATION

<table>
<thead>
<tr>
<th>Degree</th>
<th>Institution</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhD</td>
<td>Computer Science, Stony Brook University</td>
<td>2017</td>
</tr>
<tr>
<td>MS</td>
<td>Computer Science, University of California at Irvine</td>
<td>2010</td>
</tr>
<tr>
<td>BTech</td>
<td>Computer Science and Engineering, SRM University</td>
<td>2008</td>
</tr>
</tbody>
</table>

SELECTED PROFESSIONAL HISTORY

<table>
<thead>
<tr>
<th>Position</th>
<th>Institution</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant Professor</td>
<td>Department of Computer Science, The University of Iowa</td>
<td>2018 – current</td>
</tr>
<tr>
<td>Ford-Mozilla Fellow</td>
<td>Data &amp; Society Research Institute</td>
<td>2017 – 2018</td>
</tr>
</tbody>
</table>

SELECTED PUBLICATIONS


**SELECTED FUNDING, HONORS, AND AWARDS**

<table>
<thead>
<tr>
<th>Grant Description</th>
<th>Amount</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLAS-IIAI Pilot Grant, University of Iowa</td>
<td>$25K</td>
<td>2020</td>
</tr>
<tr>
<td>Faculty Research Award, Google Research</td>
<td>$42K</td>
<td>2019</td>
</tr>
<tr>
<td>Open Web Fellow, Ford Foundation and Mozilla</td>
<td>$60K</td>
<td>2017</td>
</tr>
<tr>
<td>Senior Emerging Technology Fellow, Open Technology Fund</td>
<td>$55K</td>
<td>2016</td>
</tr>
<tr>
<td>Franklin Antonio Scholarship, Qualcomm Research</td>
<td>$17K</td>
<td>2014</td>
</tr>
</tbody>
</table>

**SELECTED PROFESSIONAL SERVICE**

**Program committees**

<table>
<thead>
<tr>
<th>Event</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>USENIX Security</td>
<td>2021</td>
</tr>
<tr>
<td>Privacy Enhancing Technologies Symposium (PETS)</td>
<td>2021</td>
</tr>
<tr>
<td>USENIX Symposium on Networked Systems Design and Implementation (NSDI)</td>
<td>2021</td>
</tr>
<tr>
<td>Privacy Enhancing Technologies Symposium (PETS)</td>
<td>2020</td>
</tr>
<tr>
<td>USENIX Symposium on Research in Attacks Intrusions and Defenses (RAID)</td>
<td>2019</td>
</tr>
<tr>
<td>USENIX Workshop on Free and Open Communication on the Internet (FOCI)</td>
<td>2019</td>
</tr>
<tr>
<td>ACM Workshop on Privacy in the Electronic Society (WPES)</td>
<td>2019</td>
</tr>
</tbody>
</table>

**Panels**

<table>
<thead>
<tr>
<th>Panel</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Privacy and ethics” at Department of Computer Science at U Iowa</td>
<td>2020</td>
</tr>
<tr>
<td>“Privacy and ethics” at University of Iowa Teaching with the Library</td>
<td>2019</td>
</tr>
<tr>
<td>“Privacy and ethics” at Department of Computer Science at U Iowa</td>
<td>2019</td>
</tr>
<tr>
<td>“Data privacy online” at Barnard College of Columbia University</td>
<td>2018</td>
</tr>
</tbody>
</table>

**SELECTED MEDIA COVERAGE**

<table>
<thead>
<tr>
<th>Media Source</th>
<th>Description</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Hill</td>
<td>“Reddit enlists users to combat coronavirus misinformation”</td>
<td>2020</td>
</tr>
<tr>
<td>Mother Jones</td>
<td>“Anti-Muslim Hate Has Been Rampant on Reddit Since the NZ Shooting”</td>
<td>2019</td>
</tr>
<tr>
<td>Mother Jones</td>
<td>“Users of a Major Online Trump Hub Expect They’ll Be Kicked off Reddit—and They Don’t Know Where to Go”</td>
<td>2019</td>
</tr>
<tr>
<td>El Pais</td>
<td>“Descenso en cinco clics a la madriguera del supremacismo en YouTube”</td>
<td>2019</td>
</tr>
<tr>
<td>Mother Jones</td>
<td>“Why Reddit Is Losing Its Battle with Online Hate”</td>
<td>2019</td>
</tr>
<tr>
<td>Mother Jones</td>
<td>“How Fascist Sympathizers Hijacked Reddit’s Libertarian Hangout”</td>
<td>2018</td>
</tr>
<tr>
<td>Inside Science</td>
<td>“Battling Online Bots, Trolls and People”</td>
<td>2018</td>
</tr>
<tr>
<td>CNET</td>
<td>“Reddit: Russian propaganda spread on our site before 2016 election”</td>
<td>2018</td>
</tr>
<tr>
<td>New Scientist</td>
<td>“Politics chat on Reddit reads like it was written by 6-year-olds”</td>
<td>2017</td>
</tr>
<tr>
<td>CBS News</td>
<td>“Reddit was misinformation hotspot in 2016 election, study says”</td>
<td>2017</td>
</tr>
<tr>
<td>Vox</td>
<td>“As politicians become less civil, so does the internet”</td>
<td>2017</td>
</tr>
</tbody>
</table>
Brian Robert Ekdale
Associate Professor
School of Journalism & Mass Communication
University of Iowa
brian-ekdale@uiowa.edu

ACADEMIC POSITIONS

School of Journalism & Mass Communication, University of Iowa
Associate Professor, 2018–Present.
Assistant Professor, 2011–2018.

EDUCATION

Ph.D. in Mass Communication, 2011
University of Wisconsin-Madison

M.A. in Communication Studies, 2005
Northern Illinois University

B.A. in Speech Communication and Computer Science/Mathematics, 2000
Augustana College, Rock Island, IL

SELECTED HONORS & AWARDS

Research Funding
Iowa Initiative for Artificial Intelligence. (Award amount: $8,500 equivalent support via 8 weeks of
Under review.

Research Honors
Fellow-in-Residence, Obermann Center for Advanced Studies, Spring 2020.
Third Place, Open Competition, AEJMC Newspaper and Online News Division, 2017.
African Journalism Studies Top Paper Award, AEJMC International Communication Division, 2016.
Second Place Poster Award, AEJMC Cultural and Critical Studies Division, 2012.

Research/Travel Awards
UI International Programs Summer Research Fellowship, 2020. *(deferred to 2021 due to COVID-19)*

SELECTED ARTICLES IN REFEREED JOURNALS

Wellman, M., Tully, M., Stoldt, R., & Ekdale, B. (2020). Ethics of authenticity: Influencers and the
production of sponsored content. *Journal of Media Ethics.* Advanced online publication.

Ekdale, B. (2020). Reppin’ the nation, reppin’ themselves: Nation branding and self-branding in the


Ekdale, B. (2014). “I wish they knew that we are doing this for them”: Participation and resistance in African community journalism. *Journalism Practice, 8*(2), 181–196.


**SELECTED CHAPTERS IN EDITED COLLECTIONS**


**SELECTED MANUSCRIPTS IN PROGRESS**

**Ekdale, B.** (accepted pending revisions). Engaging the academy: Confronting Eurocentrism in journalism studies. Invited chapter for forthcoming edited book *Journalism research matters: (And how yours can too)*, V. Belair-Gagnon, M. Carlson, & N. Usher (eds.).


**RESEARCH GROUPS**

Algorithms and Social Media, Obermann Center Working Group & Moeller Lab Working Group, University of Iowa. 2015–Present.

Global Media Studies, Obermann Center Working Group & Moeller Lab Working Group, University of Iowa. 2019–Present.