Restoring Healthy Online Discourse by Detecting and Reducing Controversy, Misinformation, and Toxicity Online

Shiri Dori-Hacohen, Keen Sung, JengYu Chou, and Julian Lustig-Gonzalez
AuCoDe
{firstname}@aucode.io

ABSTRACT

Healthy online discourse is becoming less and less accessible beneath the growing noise of controversy, mis- and dis-information, and toxic speech. While IR is crucial in detecting harmful speech, researchers must work across disciplines to develop interventions, and partner with industry to deploy them rapidly and effectively. In this position paper, we argue that both detecting online *information disorders* and deploying novel, real-world content moderation tools is crucial in promoting empathy in social networks, and maintaining free expression and discourse. We detail our insights in studying different social networks such as Parler and Reddit. Finally, we discuss the joys and challenges as a lab-grown startup working with both academia and other industrial partners in finding a path toward a better, more trustworthy online ecosystem.

ACM Reference Format:
Shiri Dori-Hacohen, Keen Sung, JengYu Chou, and Julian Lustig-Gonzalez. 2021. Restoring Healthy Online Discourse by Detecting and Reducing Controversy, Misinformation, and Toxicity Online. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3404835.3464926

1 INTRODUCTION

We find ourselves in a world where distrust and polarization reigns, exacerbated by amplified controversy and toxicity. On February 15, 2020, WHO Director-General Tedros Adhanom Ghebreyesus said, "We're not just fighting an epidemic; we're fighting an infodemic" [1]. *Misinformation*, or the unintentional transmitting of falsehoods, leads to harmful outcomes such as polarized public discourse and xenophobia. During COVID-19, misinformation also led to preventable deaths [19]. The European Union has listed "Mitigation of systemic risks, such as manipulation or disinformation" as a leading goal of its proposed Digital Services Act [22].

Additionally, *disinformation campaigns*, i.e. coordinated, intentional transmission of information known to be false, damage online discourse on several fronts. Not only are users bombarded with untruths, but their trust in institutions and in each other deteriorates. Controversy has a salient connection with disinformation [6, 21, 24]. Strong negative emotions are artificially amplified both by malicious actors and by approaches to maximize user time online, leading to polarization. Hate speech prevents marginalized groups from sharing perspectives in homogeneous communities. Trolling behaviors such as *sealioning* make it difficult for users to engage in a discussion of nuanced issues. This erosion of trust means that users are fearful from sharing viewpoints that differ from their communities', leading to a *spiral of silence* [15] online.

Addressing the multitude of issues requires an approach that draws on expertise from a variety of areas in computer science (e.g., natural language processing, cybersecurity, and information retrieval); and from a variety of other fields (e.g., political science, communication, and psychology). In the face of this crisis of trust, academia, industry, non-profits, and governmental agencies need to work hand-in-hand to find practical and reliable solutions that can combat this rapidly growing infodemic and restoring trust online.

2 MITIGATING THE *SPIRAL OF SILENCE*

Content moderation has itself become a controversial topic [25]. Opponents argue that it is a form of harmful censorship; proponents, that reducing toxicity and noise promotes freer discussion of ideas. As examples, two online communities tout very different moderation policies: Parler, an alt-tech social network, advertises freedom of speech, and does not remove hate speech; Reddit's ChangeMyView [20], encourages open-minded discussion while asserting a robust moderation policy. Despite touting a lack of biased moderation, reporters found Parler heavily biased and containing misinformation [3, 9]. On the other hand, Reddit's ChangeMyView has been found to facilitate effective online discourse [14, 20]. While heavy moderation has limitations, transparent policies and enforcement resulted in decreased toxicity and increased user participation.

Content moderation is invaluable in promoting empathy and preventing toxicity. It can be used to reduce disinformation campaigns, including artificial amplification of controversy. However, creating and enforcing effective and fair policies are challenging tasks that require automated analyses, the expertise of researchers from different fields, and participation from industry.

3 EFFORTS WITHIN INDUSTRY

"Big tech" companies have come under scrutiny due to misinformation proliferating on their platforms [23]. Twitter recently introduced fact-checking [4]. The Center for Humane Technology lists "Making Sense of the World: Misinformation, conspiracy theories, and fake news" as the top-most entry on their "Ledger of Harms" [7]. Additionally, there are several companies combating the infodemic, such as AuCoDe, Crisp, Blackbird.ai, FactMata, and Logically.ai as well as nonprofits such as Full Fact, Avaaz, and Meedan. Consumer concerns and pushback from civil rights groups are leading to a reduction of trust in technology companies [11, 16].

However, large tech companies funded on advertising business models are ill equipped and disincentivized to handle this problem;

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07.
https://doi.org/10.1145/3404835.3464926

self-regulation does not effectively address it [8]. Additionally, attempts to mitigate mis- and dis-information can backfire as they are perceived as censorship. These challenges intermix with issues of self-radicalization online, and are further complicated by underlying controversy over competing ideologies, propaganda, and manufactured disinformation [13].

4 SUMMARY OF CHALLENGES

In sum, multiple interlocking challenges must be solved in order to appropriately restore trust to the online information ecosystem. First, detection of misinformation, while crucial and often the focus of previous work [10, 17, 18], is only part of the equation. We must also consider an effective interventions. However, attempts to mitigate the spread of online false information can lead to a sense of policing and censorship of free speech. Second, active adversarial disinformation spread, whether done by state agents for propaganda reasons, by groups for profit motives, or by major influencers, up to and including political leadership, can occur both on- and off-line. Third, toxic speech and misinformation spread more effectively because of its appeal to emotion [2]. Manipulated images and videos are striking, and require more effort to analyze.

In the presentation at SIRIP, we will highlight these challenges as faced by academic and industry players. We will present the case that multi-layered collaborations between academia, industry and nonprofit, and across multiple disciplines such as communication [5, 13], political science [21], journalism [26], and psychology [12] just to name a few, will be necessary in order to make progress on these challenges - and, no less importantly, in avoiding solutions that may lead to harmful unintended effects. As part of this talk, we will use AuCoDe's work in this space as a case study, discussing our analysis of Parler, our efforts to decrease toxicity on the internet, as well as an industry-academic partnership we formed with the University of Massachusetts Amherst's Center for Data Science.

5 PRESENTERS

AuCoDe is an AI-based startup that detects controversies and misinformation online and turns them into actionable intelligence. AuCoDe is funded through the prestigious NSF Small Business Innovation Research (SBIR) program. **Dr. Shiri Dori-Hacohen** is the CEO & founder of AuCoDe. She has 17 years of experience, including Google and Facebook. **Dr. Keen Sung** is the VP of R&D at AuCoDe. He has 12 years of interdisciplinary research experience in cognitive neuroscience and computer science. **JengYu Chou** is a Software Developer at AuCoDe. **Julian Lustig-Gonzalez** is the Co-founder and VP Finance at AuCoDe. He has 10 years of experience as a serial entrepreneur and business strategist.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. 1951091. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Tedros Adhanom-Ghebreyesus. 2020. Munich security conference. In *Munich: World health organization. https://www. who. int/dg/speeches/detail/munich-security-conference*.

[2] Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism* 6, 2 (2018), 154–175.

[3] BBC. 2020. Parler 'free speech' app tops charts in wake of Trump defeat. (2020). https://www.bbc.com/news/technology-54873800

[4] Ryan Browne and Sam Shead. 2020. Twitter puts fact-checking label on tweets linking 5G with coronavirus. https://www.cnbc.com/2020/06/08/twitter-5g-coronavirus.html.

[5] Gonen Dori-Hacohen and Nimrod Shavit. 2013. The cultural meanings of Israeli Tokbek (talk–back online commenting) and their relevance to the online democratic public sphere. *International Journal of Electronic Governance* 6, 4 (2013), 361–379.

[6] Shiri Dori-Hacohen, Elad Yom-Tov, and James Allan. 2015. Navigating Controversy as a Complex Search Task. In *Supporting Complex Search Tasks, at ECIR'15.*

[7] Center for Humane Technology. 2021. Ledger of Harms. https://ledger.humanetech.com/.

[8] Tauel Harper. 2021. We can't trust big tech or the government to weed out fake news, but a public-led approach just might work. https://theconversation.com/we-cant-trust-big-tech-or-the-government-to-weed-out-fake-news-but-a-public-led-approach-just-might-work-155955.

[9] Ryan Mac John Paczkowski. 2021. Amazon Will Suspend Hosting For Pro-Trump Social Network Parler. (2021). https://www.buzzfeednews.com/article/johnpaczkowski/amazon-parler-aws

[10] KP Krishna Kumar and G Geethakumari. 2014. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences* 4, 1 (2014), 1–22.

[11] Micah Maidenberg. 2020.

[12] Erik C Nisbet and Olga Kamenchuk. 2019. The psychology of state-sponsored disinformation campaigns and implications for public diplomacy. *The Hague Journal of Diplomacy* 14, 1-2 (2019), 65–82.

[13] Jonathan Corpus Ong and Jason Vincent A Cabañes. 2018. Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines. *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines* (2018).

[14] John Priniski and Zachary Horne. 2018. Attitude Change on Reddit's Change My View.. In *CogSci*.

[15] Anne Schulz and Patrick Roessler. 2012. The spiral of silence and the Internet: Selection of online content and the perception of the public opinion climate in computer-mediated communication environments. *International Journal of Public Opinion Research* 24, 3 (2012), 346–367.

[16] Deepa Seetharaman. 2020. Civil Rights Groups Push for Facebook Ad Boycott. https://www.wsj.com/articles/civil-rights-groups-push-for-facebook-ad-boycott-11592379002?mod=article_inline.

[17] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. NLP-based feature extraction for the detection of COVID-19 misinformation videos on Youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.*

[18] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web.* 745–750.

[19] Marianna Spring. 2020. The human cost of virus misinformation. https://www.bbc.com/news/stories-52731624.

[20] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[21] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).

[22] The European Union. 2020. The Digital Services Act: ensuring a safe and accountable online environment. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en. Accessed: 2020-04-09.

[23] Matt Vella. 2020. Big Tech has exactly one job to do in a pandemic. https://finance.yahoo.com/news/big-tech-exactly-one-job-152409343.html. Accessed: 2020-04-09.

[24] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and Fake News: Early Warning of Potential Misinformation Targets. *ACM Trans. Web* 13, 2, Article 10 (March 2019), 22 pages. https://doi.org/10.1145/3316809

[25] Emily A. Vogels. 2020. Partisans in the U.S. increasingly divided on whether offensive content online is taken seriously enough. (2020). https://www.pewresearch.org/fact-tank/2020/10/08/partisans-in-the-u-s-increasingly-divided-on-whether-offensive-content-online-is-taken-seriously-enough/

[26] Silvio Waisbord. 2018. Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism studies* 19, 13 (2018), 1866–1878.