ARTIFICIAL INTELLIGENCE AND DISINFORMATION:

STATE-ALIGNED INFORMATION OPERATIONS AND THE DISTORTION OF THE PUBLIC SPHERE

Author: **Dr. Courtney Radsch**
Report Coordinator: Julia Haas and Deniz Wagner

This report was commissioned by the OSCE Representative on Freedom of the Media as part of the project to put a spotlight on AI and freedom of expression, #SAIFE. The views, findings, interpretations, recommendations and conclusions expressed herein are those of the author and do not necessarily represent the official position of the OSCE and/or its participating States.

Wallnerstrasse 6
A-1010 Vienna, Austria
Tel.: +43-1 514 36 68 00
E-mail: pm-fom@osce.org
http://www.osce.org/fom/sofjo

TABLE OF CONTENTS

INTRODUCTION

The past few years have seen exponential increases in the capabilities of artificial intelligence (AI) and machine learning alongside the use of increasingly sophisticated information operations to manipulate the public sphere. Weaponized and gendered disinformation, influence campaigns, and online harassment have compounded existing press freedom and human rights challenges awhile creating new ones.

State-sponsored disinformation campaigns are increasingly common in political systems of all types. These campaigns leverage the design of social media platforms and the AI systems that power them to pursue a strategy of undermining, drowning out, and delegitimizing real news through coordinated efforts to silence critics and manipulate public opinion. They often include online harassment targeted at those reporting on information operations or engaged in fact-checking those in power. From loosely coordinated to tightly choreographed campaigns, information operations leverage state and/or party resources to manipulate public opinion by leveraging the AI systems that govern the platform-mediated public sphere and/or propel harassment campaigns. This paper analyses the dynamics of state-aligned disinformation campaigns and the role that AI plays in this context. It specifically examines coordinated campaigns deployed against journalists and media outlets, their gendered dimension, and how they leverage and manipulate AI systems to contort the public sphere.

Information operations have become a central element of domestic politics and geopolitics in countries around the world. State-aligned information operations refer to concerted efforts to manipulate AI systems and psychology to influence public opinion, attitudes, and actions in order to support the political goals of the state's leaders or influence elections. Disinformation refers to false, fabricated, misleading, or manipulated information disseminated with malign intent[1] to influence or deceive.[2] This is distinct from misinformation, which lacks the deceptive or malign intentionality aspect.[3] An exponential increase in the budgets, personnel, and attention devoted by states, governments, and political parties to information operations has occurred amid a decline in economic stability of independent media. This trend has become even more acute during the

[1] Tucker, Joshua, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." Hewlett Foundation, 2018. *https://www.ssrn.com/abstract=3144139*.; Guess, Andrew, and Benjamin Lyons. "Misinformation, Disinformation, and Online Propaganda." In *Social Media and Democracy*, edited by Nathaniel Persily and Joshua Tucker, 10–33. Cambridge University Press, 2020. *https://doi.org/10.1017/9781108890960.003*.

[2] See for example: Bradshaw and Howard; Jankowicz, Nina, Jillian Hunchak, Alexandra Pavliuc, Celia Davies, Shannon Pierson, and Zoë Kaufmann. "Malign Creativity: How Gender, Sex, and Lies Are Weaponized against Women Online." Science and Technology Innovation Program. Wilson Center, January 2021.

[3] Wardle, Claire, and Hossein Derakhshan. "Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making." Council of Europe, September 27, 2017. *https://firstdraftnews.org/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-de%CC%81sinformation-1.pdf?x80491*; Tumber, Howard, Silvio Waisbord, Rachel Armitage, and Cristian Vaccari, eds. "Misinformation and Disinformation." In *The Routledge Companion to Media Disinformation and Populism*, 38–48. London: Routledge, 2021. *https://doi.org/10.4324/9781003004431*.

COVID-19 pandemic,[4] contributing to what the World Health Organization labeled an "infodemic".[5] Concurrently, the prevalence and sophistication of disinformation-fueled online harassment campaigns is increasing, threatening not only the journalists and media outlets that are targeted, but also eroding public trust in journalism and fact-based reality while undermining the public's right to be informed.

These dynamics, in turn, make it more difficult for the press to hold those in power to account. Women journalists and those reporting on politics, corruption, and crime are increasingly finding themselves in the crosshairs of harassment and disinformation campaigns, often instigated by state entities and government officials, and fueled by state media and PR firms who know how to manipulate AI systems. Coordination leverages the underlying logic of network effects—where the value of a product or service to a user increases with the number of other users using the same product or service—and how various AI processes shape content circulation on social media. The strategy of targeting journalists is aimed at denying the news media agenda-setting power, framing journalists and their profession as "fake news", and decertifying journalists, particularly women, as legitimate actors in the public sphere. The outcome is to undermine the viability and sustainability of independent journalism, reinforce distrust in the media, and delegitimize reporting on social media manipulation and other types of information operations. This, in turn, reduces pluralism and the diversity of voices and perspectives. The impact on democratic governance and public health are profound.

Gendered disinformation is a defining component of state-aligned information operations. I define gendered disinformation as a form as weaponized information operations primarily aimed at women in public life, such as journalists and politicians, that leverage social media and algorithmic amplification to perpetuate misogynistic tropes and reputational threats aimed at dissuading women from participating in public life and expressing themselves freely.[6] Gendered disinformation creates a perception that such prejudiced and intolerant views and vitriol are common and justified, creating a vicious circle of distrust and demonization. Similarly, a recent empirical study and expert consultation characterized gendered disinformation as information activities (e.g. creating, sharing, disseminating content) that attack or undermine people on the basis of their gender and exploit gendered narratives to promote political, social or economic objectives.[7] Jankowicz et al. conceptualize gendered disinformation as a subset of online gender

[4] Posetti, Julie, Emily Bell, and Pete Brown. "Journalism and the Pandemic:A Global Snapshot of Impacts." ICFJ and The Tow Center for Journalism, 2020.; Reuters Institute for the Study of Journalism. "Few Winners, Many Losers: The COVID-19 Pandemic's Dramatic and Unequal Impact on Independent News Media." Accessed July 6, 2021. *https://reutersinstitute.politics.ox.ac.uk/few-winners-many-losers-covid-19-pandemics-dramatic-and-unequal-impact-independent-news-media*.

[5] Adhanom Ghebreyesus, Tedros. "Munich Security Conference: World Health Organization," February 15, 2020. *https://www.who.int/director-general/speeches/detail/munich-security-conference*.

[6] LGBTQI and trans people also are affected by gendered disinformation, which similarly plays on misogynistic and exclusionary tropes.

[7] Judson, Ellen, Asli Atay, Alex Krasodomski-Jones, Rose Lasko-Skinner, and Josh Smith. "Engendering Hate: The Contours of

abuse that is defined by the use of false or misleading gender and sex-based narratives against women with malign intent and often with some degree of coordination that is aimed at deterring women from participating in the public sphere.[8]

AI, CONTENT MODERATION AND DISINFORMATION

From the processes that automate various aspects of content moderation to machine learning and neural networks, AI is deployed throughout the online publication and content governance process.[9] Social media and other internet platforms depend on AI systems to enforce their Terms of Service and rules governing acceptable speech and behavior on their platforms. For example, automated detecting, filtering and blocking of child sexual abuse material, terrorist and violent extremist content.

Algorithms provide instructions and organize information based on a vast number of signals, and are embedded in content moderation and curation systems, search engines, and the underlying ad tech infrastructure of the contemporary internet. Algorithms and AI systems more broadly are built and trained on data, the prevalence or lack of data can have a significant influence on training models, machine learning, which in turn underpin content governance systems that impact the spread of disinformation. For example, AI systems on major social media platforms work better in some languages than others.[10]

Automation through algorithmic instructions and progressive machine learning influence content and account removals or shadow banning,[11] prioritization and de-prioritization, promotion and demotion, curation, and monetization of content. AI can also be deployed prior to publication and thus in advance of moderation, through upload filters and hash databases.[12]

Network effects make virality possible, giving rise to information cascades[13] that machine learning

State-Aligned Gendered Disinformation Online." Demos, October 2021. *https://demos.co.uk/project/engendering-hate-the-contours-of-state-aligned-gendered-disinformation-online/*.

[8] Jankowicz, Nina, Jillian Hunchak, Alexandra Pavliuc, Celia Davies, Shannon Pierson, and Zoë Kaufmann. "Malign Creativity: How Gender, Sex, and Lies Are Weaponized against Women Online." Science and Technology Innovation Program. Wilson Center, January 2021.

[9] Bukovska, Barbora. *Spotlight on Artificial Intelligence and Freedom of Expression*. Edited by Julia Haas. Vienna. Austria: OSCE Representative on Freedom of the Media, 2020. *https://www.osce.org/files/f/documents/9/f/456319_0.pdf*.

[10] On Natural Language Processing see Luccioni and Viviano 2021; Joshi et al. 2020. On Facebook see Haugen 2021; Allen

[11] Radsch, Courtney. "Shadowban/Shadow Banning" p. 295. In Belli, Luca, Nicolo Zingales, and Yasmin Curzi. *Glossary of Platform Law and Policy Terms.* Official Outcome of the IGF Coalition on Platform Responsibility, December 2021. *https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/31365/0.%20MIOLO_Glossary%20of%20Platform%20Law_digital.pdf?sequence=1&isAllowed=y*.

[12] Radsch, Courtney. "Hash/Hash Database." In *Glossary of Platform Law and Policy Terms*, edited by Luca Belli, Nicolo Zingales, and Yasmin Curzi, 157–58, 2021. *https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/31365/0.%20MIOLO_Glossary%20of%20Platform%20Law_digital.pdf?sequence=1&isAllowed=y*.; "Content-Sharing Algorithms, Processes, and Positive Interventions Working Group Part 1: Content-Sharing Algorithms & Processes." Global Internet Forum to Counter Terrorism (GIFCT), July 2021. *https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI1-2021.pdf*.

[13] Radsch, Courtney C. *Cyberactivism and Citizen Journalism in Egypt: Digital Dissidence and Political Change*. Information

in algorithmic content moderation and collaborative filtering systems will tend to reinforce. Coordinated campaigns take advantage of these properties of AI and the internet, for example, by targeting popular pages in hopes of reaching their audience and intimidating those with different views into silencing themselves. Coordinated anti-vaxxer disinformation campaigns often used mass comments from authentic, duplicate and fake accounts and deployed smears against its targets (which included journalists) to spread disinformation, for example.[14]

In various OSCE states, information operations have targeted independent media and journalists as well as opposition figures, including those who live in exile to avoid retaliation. Coordinated campaigns for example drown Facebook posts in pro-government comments and manipulate Facebook's engagement algorithms to threaten and harass their targets.[15] Sometimes, official accounts directly control fake assets without any obfuscation, a former Facebook data scientist wrote in a scathing memo before she quit. "Perhaps they thought they were clever; the truth was, we simply didn't care enough to stop them."[16]

In such scenarios, the sentiments emanating from inauthentic accounts are often picked up by real people and have led to temporary account shutdowns that force the targets off Facebook, one of the few outlets where people living in restricted press freedom environments can express themselves freely. Facebook regularly is popular and important among political activists and journalists.[17] Social media are sometimes the only platform left for people to have a conversation and for blocked news platforms to actually share it, so they should devote more resources to countries that may be small markets but where there are limited domestic alternatives for free expression and journalism.[18]

At the same time, however, despite years of reporting on pro-government information operations

---

Technology and Global Governance. Palgrave Macmillan US, 2016. *https://doi.org/10.1057/978-1-137-48069-9*.; Sunstein, Cass R. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press, 2018. *https://doi.org/doi:10.1515/9781400890521*.

[14] For examples from France and Italy, see Gleicher, Nathaniel, Ben Nimmo, David Agranovich, and Mike Dvilyanski. "Adversarial Threat Report." Meta (Facebook), December 1, 2021. *https://about.fb.com/wp-content/uploads/2021/12/Metas-Adversarial-Threat-Report.pdf*.

[15] For the example of the New Azerbaijan Party (YAP), see "April 2021 Coordinated Inauthentic Behavior Report." Detailed Report: CIB. Meta (Facebook), April 2021. *https://about.fb.com/news/2021/05/april-2021-coordinated-inauthentic-behavior-report/*.;
Geybullayeva, Arzu. "Azerbaijan's troll factory revealed – Azerbaijan Internet Watch." *Azerbaijan Internet Watch* (blog), September 4, 2021. *https://www.az-netwatch.org/news/azerbaijans-troll-factory-revealed/*; "September 2020 Coordinated Inauthentic Behavior Report." Facebook, September 2020. *https://about.fb.com/wp-content/uploads/2020/10/September-2020-CIB-Report.pdf*.; Geybullayeva, Arzu. "Inauthentic pages target independent news platform – will Facebook take notice [part 2, the case of Mikroskop Media] –." Azerbaijan Internet Watch, April 14, 2021. *https://www.az-netwatch.org/news/inauthentic-pages-target-independent-news-platform-will-facebook-take-notice-part-2-the-case-of-mikroskop-media/*.

[16] Silverman, Craig, Ryan Mac, and Pranav Dixit. "'I Have Blood On My Hands': A Whistleblower Says Facebook Ignored Global Political Manipulation." *BuzzFeed News*, September 14, 2020. *https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo*.

[17] For example, interview with journalist and digital rights expert Arzu Geybullayeva, July 15, 2021.

[18] For example, Facebook only implemented Azeri language review in 2020. See Kim Malfacini, Manager of Product Policy at Facebook, email reply to author, Jan. 14, 2022.

by journalists and academics, Facebook does not necessarily increase resources or deploy local operations staff, even when internal and journalistic investigations expose ongoing state-backed trolling.[19]

Collaborative filtering relies on algorithms making automatic personalized predictions based on compiled preferences from other users and their interactions. Recommendation algorithms include a range of signals and data points that are continuously updated and reviewed. Typically, social media recommendation algorithms are premised on keeping a user engaged with the content and the platform for as long as possible.

Not all algorithms work the same, however. Google's search algorithms, for example, explicitly incorporate graph authority, that is, signals of quality such as the name and originality of the publisher. On Facebook, one of the world's most important internet platforms, the News Feed determines what is shown or omitted via a ranking algorithm that seeks to show users content that they will find most relevant and engaging, and as such, it is being continually developed and tested.[20] However, the Feed ranking models do not include consideration about the authority of the content producers, meaning that troll farms, low quality clickbait, and plagiarized material can do just as well as quality, original content or journalism.[21] As former Facebook data scientist Jeff Allen put it in his exit memo, "basically all publishers are competing at the content level." That is one reason why information operations target this platform and why disinformation is so challenging to address at scale in the Facebook ecosystem specifically.

From training hate speech classifiers and integrity labeling, to content moderation and recommendation systems, language plays a significant role in AI systems. Content moderation, search, information integrity, and machine-learning systems are only as good as the data sources and the company's artificial-intelligence capabilities in a given language.[22] Limited data and training sets lead to limited moderation and poor algorithmic sophistication in low priority languages. Most NLP models are declared "language agnostic" but are in fact trained and tested

[19] Wong, Julia Carrie, and Luke Harding. "'Facebook Isn't Interested in Countries like Ours': Azerbaijan Troll Network Returns Months after Ban." *The Guardian*, April 13, 2021, sec. Technology. *https://www.theguardian.com/technology/2021/apr/13/facebook-azerbaijan-ilham-aliyev*; Geybullayeva, Arzu. "Inauthentic pages target independent news platform – will Facebook take notice [part 2, the case of Mikroskop Media] –." Azerbaijan Internet Watch, April 14, 2021. *https://www.az-netwatch.org/news/inauthentic-pages-target-independent-news-platform-will-facebook-take-notice-part-2-the-case-of-mikroskop-media/*
See also Facebook/Meta's monthly reports on Coordinated Inauthentic Behavior https://about.fb.com/news/tag/coordinated-inauthentic-behavior/

[20] Kramer, A. D. I., J. E. Guillory, and J. T. Hancock. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111, no. 24 (June 17, 2014): 8788–90. *https://doi.org/10.1073/pnas.1320040111*.

[21] Ibid Allen

[22] D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. Strong Ideas. MIT Press, 2020. *https://books.google.com/books?id=zZnSDwAAQBAJ*; Hao, Karen. "How Facebook and Google Fund Global Misinformation." *MIT Technology Review*, November 20, 2021. *https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/*.

on languages from a few "wealthy language families". The skewed distribution of resources in AI modeling means that 90 percent of the world's 7000 languages used by more than a billion people have minimal, if any, support.[23] Languages and countries not prioritized by major platforms or the AI research community receive far less attention, if any at all, with limits the ability to leverage AI to effectively counteract disinformation or online harassment. For example, Ethiopia's 100 million population speaks six languages, but Facebook integrity systems only support two of them, and for several years the company lacked expertise in two of India's most popular languages spoken by 600 million people.[24] At a more basic level, Amharic, the second most spoken Semitic language after Arabic, is disadvantaged in AI systems because many of its typological features are "ignored" in key classifications databases.[25] At the very least, this data void impedes machine learning, automated detection of disinformation or abuse, and means that many people around the world are communicating on platforms that lack sufficient linguistic support.

NEURAL NETWORKS, MANUFACTURED NEWS, AND SYNTHETIC/DEEP FAKE TECHNOLOGY

Computer vision and natural language processing (NLP) have become increasingly accurate as the amount of data online used to develop and train deep-learning models continues to expand (at least in dominant languages). As trained neural networks get exponentially larger they are getting better at making accurate predictions, such as how a person would look or sound, or sequencing words or sentences, resulting in more nuanced images and more realistic-sounding text and audio.

Natural language processing is being used to manufacture news. While this has been a benefit to news organizations that have outsourced some of their formulaic story generation to AI systems and journalist bots—coverage of simple business or sports news, for example—it also exacerbates the existing challenges of identifying real news and journalism online.[26] The ability of humans to detect news generated by a model versus a human decreases, to the point that it is near chance in some of the most popular and widespread models.[27]

---

[23] Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–93. Online: Association for Computational Linguistics, 2020. *https://doi.org/10.18653/v1/2020.acl-main.560*. P 6282

[24] Subramaniam, Tara. "The Big Takeaways from the Facebook Papers." *CNN*, October 26, 2021, sec. Business. *https://www.cnn.com/2021/10/26/tech/facebook-papers-takeaways/index.html*.

[25] Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–93. Online: Association for Computational Linguistics, 2020. *https://doi.org/10.18653/v1/2020.acl-main.560*.

[26] Peiser, Jaclyn. "The Rise of the Robot Reporter." *The New York Times*, February 5, 2019, sec. Business. *https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html*.

[27] 500-800 words is a typical length for a standard news article in English. The GPT-3 model produced news articles of around 500 words that humans had difficulty distinguishing from human-written news articles. Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. "Language Models Are Few-Shot Learners," 25. Vancouver, Canada, 2020.

The increasing prevalence of automatically generated text as well as fake and manipulated multimedia creates new perils for journalism and the spread of disinformation. AI and machine-learning techniques including NLP models, neural networks, and "generative adversarial networks" that can learn and improve their functioning has led to cheaper, more sophisticated and widespread use of AI tools, for example to create deep fake technologies. These machine-learning models enable the creation of increasingly realistic synthetic media with diminishing amounts of data; in fact, some models can create realistic talking heads with as few as one image, though with 32 images the model can achieve perfect realism and personalization.[28] But perfection is often not necessary in disinformation operations. Manipulated images, video and audio[29] are referred to as "deep fakes" and "shallow fakes" or "cheap fakes" depending in its level of sophistication.[30]

Deep fakes are manipulated video, audio, or other digital representations produced by sophisticated and often experimental machine-learning techniques that yield seemingly realistic, but fabricated, images and sounds. They can be difficult for even sophisticated users and technology to determine authenticity, and thus user-level media and information literacy initiatives have limited impact. Shallow fakes, which are poorly produced and only require easily accessible software or none at all, are more easily identified as inauthentic, but are also easy to make and turn into memes that enhance virality. Both are types of audiovisual manipulation that can diminish trust in news on social media by increasing uncertainty, even if people are not necessarily misled, by contributing to "generalized indeterminacy and cynicism".[31] Fakes pose verification challenges for journalists (and others!) even as they are also weaponized against them.

Broadcast journalists are particularly susceptible to being targeted and impersonated by deep fake technology since there is so much source material to train deep fake AI systems with, making for more realistic impersonation.[32] "The backbone of deep fake technology is deep learning neural networks trained on facial images to map the facial expressions of the source to the target," according to experts, and the amount of material shared by so many individuals on social media and the internet more broadly, means that there is a lot of source material to train on.

---

[28] Zakharov, Egor, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models." *ArXiv:1905.08233 [Cs]*, September 25, 2019. *http://arxiv.org/abs/1905.08233*. Pp. 1, 8.

[29] As the costs of sophisticated technology decrease, concerns increased sophistication and a growing concern about audio fakes, as the rise of Clubhouse and other audio-centric platforms increase in popularity among journalists.

[30] Paris, Britt, and Joan Donovan. "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence." Data & Society, September 2019. *https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf*.; Brown, Nina I. "Deepfakes and the Weaponization of Disinformation." *Virginia Journal of Law & Technology* 23, no. 1 (2020). *https://heinonline.org/HOL/Page?handle=hein.journals/vjolt23&id=1&div=&collection=*.; Born, Kelly, and Neil Edgington. "Analysis of Philanthropic Opportunities to Mitigate the Disinformation/Propaganda Problem." Hewlett Foundation, 2017. *https://www.hewlett.org/wp-content/uploads/2017/11/Hewlett-Disinformation-Propaganda-Report.pdf*.

[31] Vaccari, Cristian, and Andrew Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media + Society* 6, no. 1 (January 1, 2020): 2056305120903408. *https://doi.org/10.1177/2056305120903408*.

[32] Botha, Johnny, and Heloise Pieterse. *Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security*, 2020.; Lyu, Siwei. "Detecting 'deepfake' Videos in the Blink of an Eye." The Conversation. *http://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072*.

Creating content that looks real has become a convincing way to smear or spread false information about someone, with journalists as both targets of the manipulated content themselves but also targeted with fake content in order to entice them to cover it as news, which would then prompt accusations of reporting "fake news". From a strategic perspective, smearing is a recurring tactic in framing contestations and can trigger decertification of the target journalist and/or their media outlet, and delegitimize the broader perspective or issue more generally.

When women journalists are targeted with deep or shallow fakes they are often sexualized or sexually explicit, which can lead to more harassment, dogpiling, and physical reprisals, especially in contexts where the falsified behavior is framed as culturally inappropriate. One study found that 96 percent of all existing deep fakes circulating online featured women in acts of nonconsensual pornography.[33] Leveraging such manipulated media to defame or discredit women journalists is a common strategy, and likely to become an even more powerful tactic that will be deployed in framing competitions as it becomes easier to manipulate audio and video.

Smear campaigns are typically coordinated to imply a general consensus, for example that a journalist is a peddler of "fake news", a spy, a slut, etc. Smearing often has a gendered or sexualized component, particularly when aimed at women, and gendered tropes and disinformation are common attributes of smear campaigns. These are often coupled with accusations that journalists are not abiding by professional standards or journalistic norms. The term "presstitutes", for example, is specifically aimed at women journalists, framing them in a single word as both morally questionable and willing to trade sex for stories. The popularity of this term has propelled it to other parts of the internet ecosystem.

Behavioral science has established how negative narratives resonate and achieve stickiness, which, when coupled with the collaborative filtering and ranking algorithms, can lead to "emotional contagion" and cause people to experience the same emotions without their awareness or alter their behavior to match the new apparent reality.[34] Smear campaigns stir anger, another potent emotion that is also one of the most interactive and contagious.

EVOLUTION AND PREVALENCE OF STATE-ALIGNED INFORMATION OPERATIONS

The weaponization of information operations has evolved over the past several years to become

[33]

[34] Kramer, A. D. I., J. E. Guillory, and J. T. Hancock. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111, no. 24 (June 17, 2014): 8788–90. *https://doi.org/10.1073/pnas.1320040111*Singer, P. W. and Emerson T. Brooking. *Likewar: The Weaponization of Social Media*. *Likewar : The Weaponization of Social Media*. Boston: Eamon Dolan/Houghton Mifflin Harcourt, 2018. P. 162.

far more widespread, prolific, and organized. Moreover, information warfare and domestic influence operations are big business. No longer relegated to the shadows or the opaque budgets of authoritarian governments, content moderation and manipulation services are available to anyone with the resources to pay and are offered by a widening array of firms globally.

According to a recent study by the Oxford Internet Institute, government and political actors in at least 80 countries have deployed cyber troops, leveraging social media platforms, messaging apps, and state-aligned media along with a rising industry of public relations and political operatives who sell their services in an increasingly lucrative market.[35] Facebook, for example, reported that it took down more than 150 information operations from 50 countries around the world between 2017 and 2020, many of which included journalists as targets.[36] The company said that these "InfoOps" were becoming more sophisticated and more widely available, and often included state-affiliated media. Twitter has removed tens of thousands of accounts, as well as content, implicated in disinformation operations.[37]

While the use of cyber troops and paid commentators has a well-known history in countries with poor press freedom records and high levels of repression,[38] the expansive use of these strategies by liberal democracies and populist leaders has exponentially increased in the past five years since the U.S. election of Donald Trump as president, the UK's "Brexit" vote to leave the European Union, and revelations about how Facebook and other social media companies fuel the spread of disinformation, propaganda, and violence.[39]

The commercialization of information and influence operations has risen exponentially since 2016, giving rise to what I term "moderation mercenaries", firms or individuals who sell their social

[35] Bradshaw, Samantha, Hannah Bailey, and Philip N. Howard. "2020 Industrialized Disinformation 2020 Global Inventory of Organized Social Media Manipulation." Working Paper. Oxford Internet Institute, 2021. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/01/CyberTroop-Report-2020-v.2.pdf.

[36] Facebook: The State of Influence Operations 2017-2020. about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf

[37] Twitter Transparency Reports on Information Operations. https://transparency.twitter.com/en/reports/information-operations.html

[38] Shahbaz, Adrian, and Allie Funk. "Freedom on the Net 2020. The Pandemic's Digital Shadow." Freedom House, 2020. See other years as well.

[39] Opinio Juris. "Renewed Impetus for Accountability: Implications of the Myanmar Fact-Finding Mission Report," September 25, 2018. *https://opiniojuris.org/2018/09/25/renewed-impetus-for-accountability-implications-of-the-myanmar-fact-finding-mission-report/*.; Tumber, Howard, and Silvio Waisbord, eds. *The Routledge Companion to Media Disinformation and Populism*. London: Routledge, 2021. *https://doi.org/10.4324/9781003004431*.; Tucker, Joshua, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." Hewlett Foundation, 2018. *https://www.ssrn.com/abstract=3144139*.; Ferrara, Emilio, Stefano Cresci, and Luca Luceri. "Misinformation, Manipulation, and Abuse on Social Media in the Era of COVID-19." *Journal of Computational Social Science* 3, no. 2 (November 1, 2020): 271–77. *https://doi.org/10.1007/s42001-020-00094-5*.; Bradshaw, Samantha, Hannah Bailey, and Philip N. Howard. "2020 Industrialized Disinformation 2020 Global Inventory of Organized Social Media Manipulation." Working Paper 2021. Oxford Internet Institute, 2021. *https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/01/CyberTroop-Report-2020-v.2.pdf*. Bradshaw, Samantha, and Philip N Howard. "Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation," 2017.

media manipulation skills to whoever can pay. This includes for-hire public relations (what Bradshaw and Howard call "black PR firms")[40] and research and analytics firms that design and carry out information operations regardless of their detrimental impact on the public sphere, human rights, or democracy. A Buzzfeed investigation found at least 27 online information operations attributed to PR or marketing firms between 2011 and 2019, with 70 percent of them occurring in 2019.[41] They included domestic and foreign firms working in countries with a history of online harassment of journalists, particularly women, like the Philippines and Saudi Arabia, as discussed further below. According to the Oxford Internet Institute, the industry grew to at least $68 million by the end of the last decade, though this is very likely a vast understatement.[42]

Some moderation mercenaries provide fake, misleading, or plagiarized content that looks like journalism, further obfuscating quality independent news and contributing to the erosion of trust in the media. The use of such PR firms and coordinated disinformation by political candidates has become a standard part of electoral campaign repertoires around the world. Other disinformation actors are motivated purely by profit motive, and seek to boost content monetization through plagiarism and algorithmic manipulation. Much of this material plagiarizes or distorts existing journalistic material. Although they may not be intentionally state-aligned, their work has many similar impacts.

Increasingly, those who can afford to pay for these services include politicians and state entities, with information operations having become a standard element of campaigning and electioneering around the world.[43] The main social media platforms play an important role in making such information operations possible and profitable. For example, a former Facebook employee noted the company's ranking algorithms rather than user choice were responsible for the most significant portion the disinformation operations' ability to reach Facebook users, with 75 percent of the reach non-direct, in other words, algorithmically-driven.[44] Similarly, as much as 60 percent of engagement with Instant Articles, a fast-loading format specifically for publishers, were taking place on scraped content—much of it plagarized from real news outlets—at one point.[45] Facebook

[40] Nyst, Carly, and Nick Monaco. "State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns." Institute for the Future, 2018. *https://www.iftf.org/fileadmin/user_upload/images/DigIntel/IFTF_State_sponsored_trolling_report.pdf*.

[41] Silverman, Craig, Jane Lytvynenko, and William Kung. "Disinformation For Hire: How A New Breed of PR Firms Is Selling Lies Online." *Buzzfeed*, June 6, 2020. *https://www.buzzfeednews.com/article/craigsilverman/disinformation-for-hire-black-pr-firms*.

[42] Bradshaw and Howard, "The Global Disinformation Order 2019 Global Inventory of Organised Social Media Manipulation."

[43] See for example Apuke, Oberiri. "The Role of Social Media and Computational Propaganda in Political Campaign Communication." *Language & Communication* 5 (November 25, 2018): 225–51. Also Bradshaw, Samantha, Hannah Bailey, and Philip N. Howard. "2020 Industrialized Disinformation 2020 Global Inventory of Organized Social Media Manipulation." Working Paper. Oxford Internet Institute, 2021. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/01/CyberTroop-Report-2020-v.2.pdf.

[44] Allen, Jeff. "How Communities Are Exploited on Our Platforms: A Final Look at the 'Troll Farm' Pages." MIT Technology Review, October 4, 2019. *https://s3.documentcloud.org/documents/21063547/oct-2019-facebook-troll-farms-report.pdf*.

[45] Instant Article a native publication format for Facebook that loads more quickly than others, see *https://www.facebook.com/formedia/tools/instant-articles*. Allen, Jeff. "How Communities Are Exploited on Our Platforms: A Final Look at the 'Troll Farm' Pages." MIT Technology Review, October 4, 2019.

(which owns Instagram) is a primary vector given its popularity and availability: it has more than 3 billion users across the globe, growing popularity in most countries, and widespread availability, especially through data subsidization provided by programs like Facebook Free Basics, which provided free access to a pared-down mobile experience including a selection of media outlets.[46] An MIT Technology Review investigation found that Google and Facebook were bankrolling disinformation operations by paying millions in advertising dollars to clickbait actors, particularly in countries where these payouts "provide a larger and steadier source of income than other forms of available work."[47] Because social media algorithms boost engagement content that goes viral on one platform is likely to do well on another, and will often be "recycled"—via plagiarism—to maximize distribution and revenue.

Before the Cambridge Analytica scandal that revealed the vast information operations of the 2016 Trump campaign, Rappler CEO and Editor-In-Chief Maria Ressa uncovered how the Duterte campaign in the Philippines used fake accounts, disinformation, and coordinated propaganda to leverage Facebook's algorithms and manipulate public opinion.[48]

Ressa and her team at the news website *Rappler* pioneered disinformation reporting, turning it into a beat and developing new approaches to tracking and documenting online propaganda techniques while fending off increasingly virulent online attacks. These attacks targeted Ressa, smearing the Nobel Prize winning journalist, threatening her with rape, and relentlessly inundating her with abusive messages. The online violence increased following her reporting and commentary on disinformation and Duterte, according to a study that examined five years worth of online harassment directed at Ressa.[49] It revealed in forensic detail the dynamics of the abuse against her, noting that 60 percent of the attacks used disinformation and accusations of "fake news" to undermine her journalistic integrity and credibility. Ressa, a Philippines and American citizen, is also facing nine lawsuits and up to one hundred years in prison in retaliation for her reporting. The heightened danger of state-aligned disinformation campaigns is their linkage with the rest of the state apparatus, such as law enforcement and the judiciary, which can be leveraged as part of the campaign. Strategic lawsuits against public participation, SLAPPs, such as those facing Ressa and Caruana Galizia before her death, are a case in point.

https://s3.documentcloud.org/documents/21063547/oct-2019-facebook-troll-farms-report.pdf.

[46] "Free Basics in Real Life Six Case Studies on Facebook's Internet 'On Ramp' Initiative from Africa, Asia and Latin America." Advox. Global Voices, July 27, 2017.

[47] Hao, Karen. "How Facebook and Google Fund Global Misinformation." MIT Technology Review, November 20, 2021. https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/.

[48] Ressa, Maria. "Propaganda War: Weaponizing the Internet." *Rappler*, October 3, 2016. *https://www.rappler.com/nation/propaganda-war-weaponizing-internet*.

[49] Posetti, Julie, Diana Maynard, and Dylan Bontcheva. "Maria Ressa- Fighting an Onslaught of Online Violence, A Big Data Analysis." International Center for Journalists, March 2021. *https://www.icfj.org/sites/default/files/2021-03/Maria%20Ressa-%20Fighting%20an%20Onslaught%20of%20Online%20Violence_0.pdf*.

THE EXAMPLE OF INVESTIGATING COORDINATED DISINFORMATION

Often the planning and coordination of disinformation and harassment campaigns take place on private platforms such as Telegram or WhatsApp, secret Facebook groups, or the Dark Web. When messages then go into public forums there appears to be an authentic groundswell, which can trigger further amplification and visibility via content-sharing algorithms.

The case of investigative journalist and blogger Daphne Caruana Galizia is emblematic and serves as a yet-unheeded warning. Galizia was vilified in a campaign of both online and offline harassment in the months before her assassination in October 2017. Secret pro-government Facebook groups were reportedly used to rally supporters and denigrate Galizia and other "problematic" journalists, while fake accounts were used to spread "fake news" and counteract journalistic reporting.[50]

"This has been going on for years, this vitriol," said Caroline Muscat, editor of *The Shift* news, who led a six-month investigation into the social networks of the Labour party following its electoral victory and Galizia's murder. "They drown you out with fake news and spend money to drown out real news." This makes it more difficult for independent media to play an agenda-setting role because recommendation and search algorithms are inundated with alternative signals of popularity and interest.

On the one-year anniversary of Galizia's assassination, Muscat described one photo that circulated showing a picture of her and Daphne with the caption "We got rid of one witch and another one appeared" followed by a comment in Maltese "this one deserves bombs too." Muscat had spent six months observing the way that government-affiliated private Facebook groups organized coordinated attacks on political opponents and journalists like Galizia, whose dogged reporting on corruption and cronyism made her few friends among those in power. After infiltrating six secret Facebook groups totaling more than 60,000 members, some of which were administered by people working in government ministries according to her reporting, she observed how the targeting of activists, journalists and difficult politicians was synchronized in these closed groups before spreading into public channels.[51]

And while the groups have changed over the intervening years, the tactics have remained the same and are being refined, Muscat said in a 2021 interview.[52] "The government whip sets the narrative,

[50] "Investigating Joseph Muscat's Online Hate Machine." Accessed September 29, 2021. *https://theshiftnews.com/2018/05/14/investigating-joseph-muscats-online-hate-machine/.*
[51] The Shift Team. "Investigating Joseph Muscat's Online Hate Machine." *The Shift*, May 14, 2018. *https://theshiftnews.com/2018/05/14/investigating-joseph-muscats-online-hate-machine/.*
[52] Author interview with Caroline Muscat. Virtual, July 15, 2021.

and the trolls start to pick it up." Over the past several years, *The Shift* has conducted a series of investigations into the use of state-aligned online hate groups, and the use of coordinated harassment and propaganda campaigns.[53] Yet little has changed in how Information Operations are conducted.

Although Muscat reported the hateful and threatening comments to Facebook, she said the company replied that it did not amount to hate speech and therefore did not remove the post. “I can guarantee they don’t have anyone in Malta,” she said in 2018. Facebook said it had “Maltese language review” in place since 2018.[54]

Malta is a country of less than half a million people who speak a language that is under-represented digitally among European languages[55] and is a poor-resourced language in AI. Malta seems to represent a market too small for most global companies to care about. Maltese is among the groups of languages that currently lack NLP tools but "fight on with [its] gasping breath" for resources.[56]

THE DISINFORMATION BEAT

Journalists have naturally turned their reporting to how the state and political parties use social media and the dynamics of the broader information ecosystem. Reporting on AI, algorithms, and disinformation have increased exponentially in the past five years, with new beats emerging along with new types of reporting that relies on data journalism, programming expertise, and access to the big tech platforms that govern so much of the public sphere. Information operations are newsworthy, not least of all when they are so central to contemporary elections and politics. Specifically since 2016, journalists around the world have developed new beats covering mis-/disinformation, information operations, and conspiracy theories.[57] Much of what we know about information operations and how this form of social media propaganda takes place in practice initially came from reporting done by journalists like Maria Ressa and Rappler in the Philippines, Buzzfeed and the Markup in the United States, the Organized Crime and Corruption Reporting Project (OCCRP) in Europe, and countless other outlets and freelancers.

These journalists, in turn, have become targets of disinformation and online harassment campaigns themselves. Journalists who cover information operations and investigative journalists scrutinizing

[53] See The Shift landing page "Labour's Secret Online Groups" that collect this reporting at https://theshiftnews.com/category/investigations/labour-online-hate-groups/
[54] Kim Malfacini, Manager of Product Policy at Facebook, email reply to author, Jan. 14, 2022.
[55] Camilleri, John J. “Digitizing the Grammar and Vocabulary of Maltese.” In *Digitizing the Grammar and Vocabulary of Maltese*, 359–86. De Gruyter Mouton, 2016. *https://doi.org/10.1515/9783110496376-014*.
[56] Joshi et al P 6284
[57] See for example Napoli, Philip M. “The Platform Beat: Algorithmic Watchdogs in the Disinformation Age.” *European Journal of Communication* 36, no. 4 (August 1, 2021): 376–90. *https://doi.org/10.1177/02673231211028359*.; McClure Haughey, Melinda, Meena Devii Muralikumar, Cameron A. Wood, and Kate Starbird. “On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online.” *Proceedings of the ACM on Human-Computer Interaction* 4, no. CSCW2 (October 14, 2020): 1–22. *https://doi.org/10.1145/3415204*.

those in power are particularly at risk for coordinated disinformation and harassment campaigns because of their reporting of matters that officials would rather keep quiet. Gendered disinformation and its weaponization against journalists who cover information operations has become a central aspect of how these campaigns work. Organizations that track online harassment have identified a significant increase year over year,[58] alongside the expansion of these new reporting beats.

In various cases, journalists become the target of disinformation and harassment after publishing investigations into individual countries' information operations. This also happens when they are based in another country, revealed how foreign governments are deploying strategies and using propaganda apparats along with trolling and social media manipulation to influence public opinion. Such campaigns regularly dig up and use incriminating or objectionable information, often including false allegations of promiscuity, drug use, or mental illness, plastering it across social media to discredit the journalist, especially women. Such campaigns aim to discredit the target, make journalistic work seem unreliable and to ultimately stop journalists from reporting, particularly from disclosing facts about such social media propagandists.[59] Experiences show how framing journalists as unreliable hacks is common but more difficult in contexts where the state or foreign states exert little influence over the news media, or where there are high levels of trust in news media,[60] and access to legal remedy.

Amid the myriad reasons for the rise of digitally-inflected propaganda campaigns in countries that had a free press in the 21st century are their integration into electoral politics, the emergence of an industry devoted to supporting and designing them, and few laws or regulations restricting their use.

AGENDA-SETTING, FRAMING, AND CERTIFICATION

The concepts of agenda-setting, framing, and certification from communications theory and journalism studies help explain how strategies of harassment and disinformation work. Agenda-setting is about the power to shape and determine what to think about, specifically in the public sphere, thus influencing how priorities are set. Framing affects the basic frameworks of understanding available to society that enable them to organize and make sense of events and experiences. Frames structure experience and expectations, and cultivate conceptions of what is important, correct, or relevant.[61] Agenda-setting and framing are the processes by which

[58] See for example Bedoya, Daniel, Michael Carbone, and Sage Cheng. "Strengthening Civil Society's Defenses: What Access Now's Digital Security Helpline Has Learned From Its First 10,000 Cases." Access Now, June 7, 2021. *https://www.accessnow.org/cms/assets/uploads/2021/06/Helpline-10000-cases-report.pdf*.
[59] For example, see Aro, Jessikka. "The Cyberspace War: Propaganda and Trolling as Warfare Tools." *European View* 15, no. 1 (June 2016): 121–32. *https://doi.org/10.1007/s12290-016-0395-5*.
[60] For example, Reunanen, Esa. "Digital News Report: Finland." Reuters Institute Digital News Reports 2016-2020. *https://www.digitalnewsreport.org/*
[61] Morgan, Michael, ed. *Against the Mainstream : The Selected Works of George Gerbner*. New York: P. Lang, 2002.; Goffman,

institutions like the media or policymakers compete to shape what people pay attention to, how they interpret everyday life, and how they assign meaning to, or reorient their thinking about, a particular issue.[62]

Telling the public where to focus its attention is a source of power that accrues to those who occupy a strategic position in the communication process.[63] State and political actors as well as journalists occupy such positions. But today, their positions and the processes of agenda-setting, framing and certification are algorithmically mediated by tech platforms, meaning that those who can influence visibility online can garner greater attention and power. A beat, for example, indicates devotion of resources to a specific topic or institution, and contributes to agenda-setting.

Agenda-setting is a competitive process, typically led by elites and journalists, though social media has complicated what was often conceived of as a linear process.[64] Research has shown that media and journalist accounts do indeed set agendas on social media,[65] meaning that when the interests of those in power diverge from those in the media, it may be in their interest to limit the visibility of the alternate perspective or reframe it as lacking credibility. This is a common framing in gendered disinformation and online harassment campaigns, which includes smearing the target as a recurring tactic. Furthermore, frames conveyed by generally trusted sources, such as political leaders or personal friends, are more likely to resonate and influence opinions.[66]

When high-profile accounts with large numbers of followers post content, it sends a signal to content-sharing algorithms that boosts the visibility of that content. That content proliferates through those networks and becomes embedded into the web and difficult to remove, meaning that it can continue to be linked to, show up in searches, and influence various algorithmic systems.

Erving. *Frame Analysis : An Essay on the Organization of Experience*. Northeastern University Press. Boston: Northeastern University Press, 1986.

[62] Gilardi, Fabrizio, Theresa Gessler, Maël Kubli, and Stefan Müller. "Social Media and Political Agenda Setting." *Political Communication*, 2021; Mrogers, Everett, and James Wdearing. "Agenda-Setting Research: Where Has It Been, Where Is It Going?" *Annals of the International Communication Association* 11, no. 1 (January 1, 1988): 555–94. *https://doi.org/10.1080/23808985.1988.11678708*;
Borah, Porismita. "Conceptual Issues in Framing Theory: A Systematic Examination of a Decade's Literature." *Journal of Communication* 61, no. 2 (April 2011): 246–63. *https://doi.org/10.1111/j.1460-2466.2011.01539.x*.
; Chong, Dennis, and James N. Druckman. "Framing Theory." *Annual Review of Political Science* 10, no. 1 (2007): 103–26. *https://doi.org/10.1146/annurev.polisci.10.072805.103054*.

[63] Nye, Joseph S. *Soft Power: The Means to Success in World Politics*. 1st ed. New York: Public Affairs, 2004.

[64] Carazo-Barrantes, Carolina. "Agenda-Setting in a Social Media Age: Exploring New Methodological Approaches." *Agenda Setting Journal: Theory, Practice, Critique* 5, no. 1 (January 2021): 31–55. *https://doi.org/10.1075/asj.20006.car*.; Harder, Raymond A, Julie Sevenans, and Peter Van Aelst. "Intermedia Agenda Setting in the Social Media Age: How Traditional Players Dominate the News Agenda in Election Times." *The International Journal of Press/Politics* 22, no. 3 (2017): 275–93.

[65] Druckman, James N. "On the Limits of Framing Effects: Who Can Frame?" *Journal of Politics* 63, no. 4 (November 2001): 1041–66. *https://doi.org/10.1111/0022-3816.00100*.
Gilardi, Fabrizio, Theresa Gessler, Maël Kubli, and Stefan Müller. "Social Media and Political Agenda Setting." *Political Communication*, 2021.
Harder, Raymond A, Julie Sevenans, and Peter Van Aelst. "Intermedia Agenda Setting in the Social Media Age: How Traditional Players Dominate the News Agenda in Election Times." *The International Journal of Press/Politics* 22, no. 3 (2017): 275–93.

[66] Druckman, James N. "On the Limits of Framing Effects: Who Can Frame?" *Journal of Politics* 63, no. 4 (November 2001): 1041–66. *https://doi.org/10.1111/0022-3816.00100*.

The use of secret groups and/or confidential platforms enable framing contests to be coordinated in advance and appear organic when they emerged in the public sphere and became subject to content-sharing algorithms. Responding or countering those framing contests becomes difficult because machine learning will tend to interpret the campaign as engagement, and without access to similar amplification networks counter messaging or correction becomes virtually impossible.

This challenge is compounded because high-profile political accounts have received preferential treatment and protection from the negative impacts of content moderation (such as removal, shadow banning, etc.). For example, Facebook's XCheck program for VIPs, like presidents and celebrities, exempted such accounts from some or all of the platform's rules.[67] Twitter allowed world leaders and militias to remain on its platform despite apparent violations of their published terms of service.[68] These exceptions meant that content that constituted harassment or disinformation could nonetheless circulate, signal recommendation algorithms, and influence AI processes in the future.

Algorithmic amplification is a form of technological agenda-setting,[69] and can trigger information cascades, influence framing contests, and lead to incidental exposure by the general public.[70] Framing effects focus attention and are central to public opinion formation.[71] When attention is a scare resource, power struggles over agenda-setting and framing can have significant implications for how reality is constructed and what truth is believed. AI influences the dynamics of these power struggles because of how data (or lack thereof) and signals are interpreted in the platform algorithms and machine-learning processes. For example, information that is shared, liked, or commented on receives signal boosts that such content is highly engaging and thus is further amplified by recommendation algorithms. Brigading, where accounts engage in repetitive mass behavior to harassment or silence a target, is one such tactic. Creating signals of engagement that boost visibility and prevalence of that information is a key aspect of disinformation campaigns.

Online harassment is aimed at undermining the credibility of independent journalists, handicapping and denying them the ability to compete in agenda-setting contests while preventing their reporting on information operations from being framed a legitimate topic of public interest.

---

[67] Horwitz, Jeff. "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt." *Wall Street Journal*, September 13, 2021, sec. Tech. *https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353*.

[68] Radsch, Courtney. "U.S. Tech Firms Looked the Other Way as Taliban Regained Power." *Tech Policy Press*, August 19, 2021. *https://techpolicy.press/u-s-tech-firms-looked-the-other-way-as-taliban-regained-power/*.

[69] Radsch, Courtney C. *Cyberactivism and Citizen Journalism in Egypt: Digital Dissidence and Political Change*. Information Technology and Global Governance. Palgrave Macmillan US, 2016. Pp. 27. *https://doi.org/10.1057/978-1-137-48069-9*.

[70] Harder, Raymond A, Julie Sevenans, and Peter Van Aelst. "Intermedia Agenda Setting in the Social Media Age: How Traditional Players Dominate the News Agenda in Election Times." *The International Journal of Press/Politics* 22, no. 3 (2017): 275–93.; Feezell, Jessica T. "Agenda Setting through Social Media: The Importance of Incidental News Exposure and Social Filtering in the Digital Era." *Political Research Quarterly* 71, no. 2 (2018): 482–94.

[71] Druckman, James N. "On the Limits of Framing Effects: Who Can Frame?" *Journal of Politics* 63, no. 4 (November 2001): 1041–66. *https://doi.org/10.1111/0022-3816.00100*.

Gendered disinformation re-frames an issue by deploying traditional tropes and cultural stereotypes that sexualize women and promote misogyny, thus seeking to decertify women journalists and rob them of their symbolic power in framing contests.

The case of "fake news" is illustrative. Over the past five years, the "fake news" offensive against journalists was propelled into a global meme and created a template that leaders around the world have used to discredit the news media and spur distrust in the media so that people will not believe negative reporting. that no one would believe them when they reported negatively on him.[72] Trump used his bully pulpit together with his Twitter account, with its millions of followers, to smear journalists and denigrate the press (more than once a day on average throughout his candidacy and presidency).[73]

Many leaders across the OSCE region particularly target reporters engaged in fact-checking, reporting on links with foreign-sponsored information warfare, or critical reporting on COVID-19 responses.[74] Additional abuse and denigration of women or journalists belonging to specific religions or ethnicities further amplifies online abuse and gendered disinformation,[75] and can become newsworthy in and of themselves. While in some cases there is only the loosest form of coordination with political leaders, even informal direction is emblematic of networked alignment and the way that such campaigns can mobilize ordinary internet users and frame the press as an elitist institution that lies rather than as a fundamental pillar of democracy. There are also innumerable examples of more overt coordination that similarly attack independent media and individual journalists, particularly those reporting on domestic and foreign disinformation campaigns or the COVID-19 pandemic.[76] In some contexts, such attacks are amplified and

[72] Downie Jr., Leonard. "The Trump Administration and the Media." Committee to Protect Journalists, April 16, 2020. *https://cpj.org/reports/2020/04/trump-media-attacks-credibility-leaks/*.
[73] Sugars, Stephanie. "The Last Trump Tweet Against the Media." U.S. Press Freedom Tracker, January 11, 2021. *https://pressfreedomtracker.us/blog/last-trump-tweet-against-media/*.
[74] Sugars, Stephanie. "From Fake News to Enemy of the People: An Anatomy of Trump's Tweets." U.S. Press Freedom Tracker. June 30, 2019. *https://pressfreedomtracker.us/blog/fake-news-enemy-people-anatomy-trumps-tweets/*.; Sugars, Stephanie, and Kristin McCudden. "Trump, in Crisis Mode, Tweets 2000th Attack on the Press." Freedom of the Press Foundation, April 13, 2020. *https://freedom.press/news/trump-crisis-mode-tweets-his-2000th-attack-press/*.
[75] For example, "Anti-Semitic Targeting of Journalists During the 2016 Presidential Campaign." ADL Report. Anti-Defamation League, October 19, 2016. *https://www.adl.org/sites/default/files/documents/assets/pdf/press-center/CR_4862_Journalism-Task-Force_v2.pdf*.; Cuen, Leigh, and Sieradzki. "Trump Supporters Bombard Michelle Fields With Misogynist Insults - Vocativ." Vocativ, September 26, 2016. *https://web.archive.org/web/20170926124220/http://www.vocativ.com/297788/trump-women/*.; Cuen, Leigh, and Jody Sieradzki. "Since Trump Attacked Megyn Kelly In August, The Hate Hasn't Stopped." Vocativ, March 3, 2016. *https://www.vocativ.com/292871/donald-trump-megyn-kelly-2/*.; Radsch, Courtney. Human Rights at Home: Media, Politics, and Safety of Journalists, § Commission on Security and Cooperation in Europe (2020). *https://www.csce.gov/international-impact/events/human-rights-home-media-politics-and-safety-journalists*.
[76] Reporters Without Borders (RSF). "Orbán's Orwellian Law Paves Way for 'Information Police State' in Hungary | Reporters without Borders," April 1, 2020. *https://rsf.org/en/news/orbans-orwellian-law-paves-way-information-police-state-hungary*; Mong, Attila. "Hungarian Journalist Csaba Lukács on Covering COVID-19 amid Attacks on Independent Media." *Committee to Protect Journalists* (blog), April 22, 2020. *https://cpj.org/2020/04/hungarian-journalist-csaba-lukacs-on-covering-covi/*.; Index on Censorship. "Hungary: Prime Minister Viktor Orban Wages Campaign against Critical Journalists." *Index on Censorship* (blog), September 27, 2017. Bozkurt, Abdullah. "Turkey's Disinformation Campaign through Trolls and Bots in the Assassination of Russian Ambassador Exposed." The Nordic Research Monitoring Network (Nordic Monitor) (blog), June 6, 2020. https://nordicmonitor.com/2020/06/11796/. *https://www.indexoncensorship.org/2017/09/viktor-orban-campaign-against-*

validated by the pro-government press and websites. Index on Censorship noted that even if the readership of specific blogs or websites is relatively small, their articles ricochet through the media universe and are referenced by the public media, ensuring that such allegations reach a considerable audience.[77] A typical approach often involves government operatives planting fake news bytes, first on social media through operatives, then in the government-controlled media. When fabricated stories are amplified by trolls, bots and influencers sharing the lies, they are often picked up by the newspapers and networks controlled by the government in a vicious feedback loop. Such attacks and their coverage by media organizations send signals that are interpreted by recommendation and search algorithms, further amplifying and extending the reach of the harassment and disinformation.

Given the use of Twitter and journalism as data sources in AI,[78] it is important to consider how much of this harassment and disinformation gets codified in the data used to develop and train AI systems. News websites and open social media sources like Twitter are key sources of data used in training sets and AI research. Research has shown that NLP models reproduce hate speech, toxicity, and stereotypes, making it even more relevant to consider the implications of harassment and gendered disinformation becoming part of the corpus used to train machine learning algorithms.[79]

The fusion of AI and disinformation robs the public of its right and ability to participate in the agenda-setting process, which is core to democratic politics. Furthermore, harassing journalists and media outlets on social media platforms manipulates content-sharing, search, monetization, and integrity algorithms and machine-learning systems and makes quality journalism less visible. The capacity of journalism to hold public leaders accountable could diminish if they are deterred from reporting on information operations, or if their reporting is not seen to have an impact.

THE ROLE OF STATE-ALIGNED MEDIA

While much of the focus on the nexus of AI and disinformation and harassment has centered on tech platforms and social media in particular, state-aligned media are an integral part of the

journalists/.

[77] For example, Censorship, Index on. "Hungary: Prime Minister Viktor Orban Wages Campaign against Critical Journalists." *Index on Censorship* (blog), September 27, 2017. *https://www.indexoncensorship.org/2017/09/viktor-orban-campaign-against-journalists/*.

[78] Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." arXiv, September 30, 2021. https://doi.org/10.48550/arXiv.2104.08758.

[79] Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." arXiv, September 30, 2021. *https://doi.org/10.48550/arXiv.2104.08758*; Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115, no. 16 (April 17, 2018): E3635–44. *https://doi.org/10.1073/pnas.1720347115*; Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models." *ArXiv:2009.11462 [Cs]*, September 25, 2020. *http://arxiv.org/abs/2009.11462*.

disinformation machinery. In countries where media capture and cronyism is high, so are government-aligned media.[80] State media and pro-government outlets can instigate, amplify, and/or perpetuate harassment and disinformation campaigns, serving to give them the imprimatur of "objective reporting", providing cover for broader coordination and engagement by moderation mercenaries, creating the illusion of being organic, and mobilizing the general public and unsuspecting users. They play a validating and amplifying role on- and offline, particularly in countries where media capture and cronyism is high. Nearly all media have online presence on the web and major social media platforms, and some politically aligned media outlets may be designated publishing partners and labeled as news outlets on Facebook, Google, and Twitter, which can affect how ranking and recommendation algorithms assess quality and integrity.

In the aftermath of disinformation campaigns online, journalists have been attacked despite limited efforts to stave such threats.[81] These can be particularly dangerous if fueled by pro-government or foreign media, or if aligned with specific political or religious groups, as such multifaceted disinformation campaigns can fuse anti-media and heteronormative frames that echo throughout the media ecosystem.[82] Even if specific accounts are banned from Facebook under prohibitions against coordinated inauthentic behavior, for example, or for using fake accounts, there are numerous examples of Pages posing as news outlets to spread disinformation, for example around elections, with such accounts nonetheless maintaining a presence on platforms because individual users, including high-profile users, repost content, effectively bypassing efforts to restrict such content.[83]

Some platforms have taken steps to label state-controlled or affiliated media on their platforms, though how to define these and where such labeling initiatives are implemented remains contested. On some platforms (e.g., YouTube) this is purely a media literacy initiative to provide context to users whereas on others (e.g., Twitter) such labeling prevents algorithmic amplification,

[80] Dragomir, Marius. "State of State Media: A Global Analysis of the Editorial Independence of State Media and an Introduction of a New State Media Typology." CEU Democracy Institute, 2021.; Aro, Jessikka. "The Cyberspace War: Propaganda and Trolling as Warfare Tools." *European View* 15, no. 1 (June 2016): 121–32. *https://doi.org/10.1007/s12290-016-0395-5*. Wierzejski, Antoni, and et al. "Information Warfare in the Internet: Countering Pro-Kremlin Disinformation in the CEE Countries." Centre for International Relations, 2017. *https://www.academia.edu/34620712/Information_warfare_in_the_Internet_COUNTERING_PRO_KREMLIN_DISINFORMATION_IN_THE_CEE_COUNTRIES_Centre_for_International_Relations_and_Partners*.

[81] For example, a Georgian cameraman was killed in 2021 following disinformation campaigns about US and European interference in the country's politics (and specifically their support for Tbilisi Pride Week), see Gigitashvili, Givi. "Georgian Far-Right Groups Embrace Anti-LGBTQ Narratives Pushed by Russian Media." *DFRLab* (blog), July 26, 2021. *https://medium.com/dfrlab/georgian-far-right-groups-embrace-anti-lgbtq-narratives-pushed-by-pro-russian-media-36f9e99a2561*.

[82] For example, Kiparoidze, Mariam. "Georgian Far Right Launches Disinformation Campaign Following Death of Journalist Beaten in Anti-LGBTQ Attack." *Coda Story*, July 13, 2021. *https://www.codastory.com/disinformation/far-right-lgbtq-georgia/*. EU vs DISINFORMATION. "Pro-Kremlin Outlets Try to Create an Alternative Anti-Western Reality in Georgia," December 14, 2021. *https://euvsdisinfo.eu/pro-kremlin-outlets-try-to-create-an-alternative-anti-western-reality-in-georgia/*.

[83] "October 2020 Coordinated Inauthentic Behavior Report." Facebook, October 2020. *https://about.fb.com/wp-content/uploads/2020/11/October-2020-CIB-Report.pdf* pp. 16-20; Gigitashvili ibid.

monetization, and advertising capabilities.[84] However, such initiatives are relatively narrow in their scope, covering only a limited number of countries and languages. The involvement of such state-aligned media in smear and disinformation campaigns can also create dissension within the media ecosystem, decrease trust in the profession, and reinforce partisanship.

By publishing disinformation as journalism, state-aligned media certifies it as "news" and frames it as an issue of public interest. Such news-washing is a common tactic and can trick algorithms into spreading disinformation and harassment. News-washing undermines the ability of algorithms that consider quality/publisher as a factor, such as search results (as opposed to, for example, simply engagement).

Journalism content is used extensively in Natural Language Processing, facial recognition and synthetic media development, machine learning and other AI systems to develop and train datasets. This means that harassment and disinformation could be further codified through AI in harmful ways, particularly in poor-resource and underrepresented languages.[85] Language models and other AI models that encode and reinforce hegemonic biases, abusive language patterns, and harmful ideologies can also perpetuate it through machine learning and through the production of new synthetic media or text that they generate.[86]

GENDERED DISINFORMATION AND THREATS TO PLURALISM

Disinformation thrives by inducing feelings of superiority, fear, and anger,[87] which are also embedded in gendered stereotypes of women's traditional role, hyper sexualization, and shifting societal norms. The combination of ingrained sexism, manipulated media, and social media platforms enable state-aligned campaigns to mobilize resources and supporters in efforts to destroy women's reputations, silence their voices, and push them out of the public sphere. There is not yet agreement over how or whether to distinguish between hate speech, sexual harassment, and gendered disinformation.

Gendered disinformation, which plays on historic stereotypes, tropes, and insults, is the latest evolution in online harassment and thrives in the contemporary information ecosystem. The social

[84] Radsch, Courtney C. "The Politics of Labels: How Tech Platforms Regulate State Media." In *2020 Annual Report: Dynamic Coalition on the Sustainability of Journalism and News Media*, edited by Edited Daniel O'Maley, Hesbon Hansen Owilla, and Courtney C. Radsch, 37-49, 2020. *https://gfmd.info/h-content/uploads/2021/11/DC-Sustainability-Annual-Report-2020-FINAL-gfmd.pdf*.

[85] With respect to AI systems content. See Joshi et al.; Luccioni and Viviano, 2021.

[86] Language models trained with even a small proportion of undesirable inputs cannot be guaranteed to avoid generating outputs with
similar biases if presented with a specific context or prompt. Luccioni, Alexandra Sasha, and Joseph D. Viviano. "What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus." *ArXiv:2105.02732 [Cs]*, May 31, 2021. *http://arxiv.org/abs/2105.02732*. P. 3

[87] Wardle, Claire, and Hossein Derakhshan. "Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making." Council of Europe, September 27, 2017. *https://firstdraftnews.org/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-de%CC%81sinformation-1.pdf?x80491*.

component of online harassment may help explain, in part, why coordinated campaigns can escalate and go viral,[88] but is inseparable from the networked nature of both online harassment and disinformation.

Studies have established how online harassment diminishes women's participation in the public sphere and is overwhelmingly targeted at women journalists, politicians, and those who are visible in male-dominated sectors.[89] A global survey of journalists found that two-thirds of female journalists get violent intimidations online in response to their work because of their gender identity. This "digital misogyny" is the latest iteration of efforts to "curtail women's freedom to use public spaces as equals"[90] and is often a concerted and organized effort organized in subcultural online spaces.[91]

There have been several studies and surveys focused on documenting the scale and scope of online harassment, and its outsized impact on women is now well-established, as is the fact that such harassment often includes sexualized threats and attacks and is even more acute for women of color or minorities, or other intersecting identities.[92] For example, a 2020 global survey found that 64 percent of white women journalists, and between 80 and 90 percent of Black, Indigenous and Jewish women journalists, said they had experienced online violence.[93] Another study of online harassment of female journalists and politicians in the UK and US found that black women received 70 percent more racist abuse than white women,[94] and anecdotal reporting indicates that Muslim women journalists in India, and secular journalists in Bangladesh for example, face disproportionately violent, and even deadly, harassment. Identity is a critical part of the threat

[88] Marwick et al; Feezell, Jessica T. "Agenda Setting through Social Media: The Importance of Incidental News Exposure and Social Filtering in the Digital Era." *Political Research Quarterly* 71, no. 2 (2018): 482–94.
[89] See for example Marwick, Alice, and Rebecca Lewis. "Media Manipulation and Disinformation Online." Data & Society, May 15, 2017. ; Sarah Sobieraj, "Bitch, slut, skank, cunt: patterned resistance to women's visibility in digital publics," Information, Communication & Society 21, Volume 11 (2018); *http://www.oxfordtoday.ox.ac.uk/features/social-media-bots-endanger-democracy-warns-oxford%E2%80%99s-internet-research-chief*; National Democratic Institute. #NotTheCost - A Call to Action. NDI, 2016. Available at https://www.ndi.org/sites/default/files/1%20%23NotTheCost%20-%20Call%20to%20Action.pdf; Krook, Mona Lena. 2017. "Violence against Women in Politics." Journal of Democracy 28 (1): 74–88.
[90] Jankowicz, Nina, Jillian Hunchak, Alexandra Pavliuc, Celia Davies, Shannon Pierson, and Zoë Kaufmann. "Malign Creativity: How Gender, Sex, and Lies Are Weaponized against Women Online." Science and Technology Innovation Program. Wilson Center, January 2021.
[91] Banet-Weiser, Sarah, and Kate M. Miltner. 2016. "# MasculinitySoFragile: Culture, Structure, and Networked Misogyny." Feminist Media Studies 16 (1): 171–174.
[92] See for example, Citron, Danielle. 2014. Hate Crimes in Cyberspace. Cambridge, MA: Harvard University Press; *Demos, Misogyny of Twitter* (2014); Internet Governance Forum (IGF) 2015: Best Practice Forum (BPF) on *Online Abuse and Gender-Based Violence Against Women*; Radsch, Courtney CPJ *Responding to Internet Abuse* (2016); *New Challenges to Freedom of Expression: Countering Online Abuse of Female Journalists*, OSCE (2016); Amnesty International *Troll Patrol* (2017); IWMF-INSI *Violence And Harassment Against Women In The News Media: A Global Picture* (2018); Ferrier, Michelle; *Attacks and Harassment: The Impact on Female Journalists and Their Reporting*. Troll-Busters and International Women's Media Foundation, September 2018; ADL *Online Hate and Harassment Report: The American* Experience (2020). Anecdotal reporting indicates that Muslim women journalists in India, and secular journalists in Bangladesh, face disproportionately violent harassment
[93] Posetti, Julie, Nermine Aboulez, Kalina Bontcheva, Jackie Harrison, and Silvio Waisbord. "Online Violence Against Women Journalists: A Global Snapshot of Incidence and Impacts." UNESCO, 2020. *https://unesdoc.unesco.org/ark:/48223/pf0000375136*.
[94] Amnesty International *Troll Patrol*. 2017.

analysis.

Disinformation campaigns and their aftermath make it more difficult for journalists to get information, comments and interviews, opening them up to further abuse, undermining the journalistic process, and undermining their role as a watchdog. This not only detracts from their ability to report but also means that the government is able to control the narrative and ensure their perspective dominates. "The aim is to make people stop asking questions; when you limit their access, and when you constantly mock them and present them in a negative light, the idea is to silence them and discourage them from doing so," said a Pakistani journalist, Ramsha Jahangir, who was targeted after reporting about the ruling party's influence operations. "I feel like my sources have drastically reduced since this has started happening; it's becoming more and more difficult for me to get interviews or comments from people because they don't want to be associated with these things and they have this idea that journalists are going to frame things in a negative way." It has even prompted more reporting on positive stories, says Jahangir, because a source will ask whether the story is going to be negative or positive, and if positive, the journalists will then be lauded on social media as true and faithful.

Disinformation, harassment, and smearing erode the visibility of the target, reframe the original issue, and seek to decertify the target as a neutral, objective, or legitimate journalist. Threats and abuse aim to directly intimidate the target, and indirectly intimidate others who would pursue the same reporting or come to the target's defense. These dynamics erode the pluralism and diversity of those who express themselves in the public sphere. Self-doubt and self-censorship are common experiences among women and journalists who are targeted by state-aligned campaigns. As the examples from around the world demonstrate, those with the means to deploy such repertoires of repression have access to resources that typically outstrip those of the media outlet or journalist at the center of a given campaign.

The nature of gendered online violence and harassment campaigns targeting women in the public sphere has gained recognition for its pervasiveness and violent sexualized nature, yet it has become endemic to the practice of journalism and politics. The failure to adequately address and mitigate online harassment against women violates their human rights and creates fertile breeding grounds for disinformation.

CONCLUSIONS AND RECOMMENDATIONS

Coordinated disinformation and harassment campaigns deploy a range of strategies and tactics in combination with each other, relying on the services of moderation mercenaries and news-washing by state-aligned media outlets. This creates a multiplier effect, which in turn impacts the visibility of specific pieces of information, contained in articles or posts, as well as the media outlet and the targeted journalist themselves. In light of the role played by state-aligned actors, the private sector

and lawmakers in countries with strong democratic institutions should adopt policies that mitigate the ability of state actors to manipulate AI and weaponize communication platforms. Efforts to combat disinformation must recognize that a range of private companies beyond just tech firms are implicated in information manipulation and must put safeguards in place. For example, registration and financing limits on paid PR firms, domestic and foreign, and better oversight by tech platforms on how their platforms are used by state actors is essential. Furthermore, greater transparency about all types of advertising and paid content promotion is needed, not just about political advertising in a handful of Western countries. This could be implemented through existing election laws and paid advertising regulations.

Information operations muddy the broader information environment and contribute to the algorithmic amplification of disinformation and online harassment. An increasingly professionalized and lucrative industry devoted to information operations has developed, with politicians representing an eager market for their services. Furthermore, as social media influencers and content creators are being brought into the governing apparatus in many countries, granted exclusive interviews with top political leaders, and given space alongside professional reporters in official press conferences, influence operations are further becoming embedded and normalized. Any meaningful efforts to combat disinformation will need to address the politicization of social media manipulation and influence operations, and their integration into electoral politics. Lawmakers should implement restrictions on the use of moderation mercenaries, black PR firms, and social media manipulation by those entrusted with public office. Countries should not only require great transparency for the platforms themselves, but should also practice what they preach by adopting transparency requirements for state and government entities related to advertising and outreach on social media and messaging platforms.

Tech platforms must reduce the profitability of intentional and opportunistic disinformation efforts, including by reducing the prevalence and ease of plagiarism or the “recycling” of news content for clickbait. Reducing the economic incentives for click-bait, "churnalism", and regurgitated journalistic content would help deter the profit-driven non-ideological actors in these disinformation networks.

At a more fundamental level, social media platforms need to improve the identification of quality journalism sources and incorporate this data into the design of algorithmic recommendation and ranking systems. Facebook, in particular, should be required to take proactive steps in each country in which it operates and devote greater resources to a wider array of languages and local contexts. There are a range of existing trust and integrity initiatives created by media professionals and journalism organizations, which should be encouraged and better utilized by tech platforms both in terms of combatting disinformation but also better protecting those entities and journalists from

online harassment.[95] A better understanding of the effectiveness, inclusivity, and scalability of these self-regulatory mechanisms is needed. Furthermore, instead of relying on their own internal, secret lists of trusted partners, trustworthy media, and state-affiliated media, tech companies should leverage existing media industry and self-regulatory initiatives and adopt a multi-stakeholder approach to labeling media.

Developing ways to identify and track deep fakes across the internet is a critical priority that will only become more urgent. In addition to existing research and initiatives aimed at finding scalable solutions, experts should study the potential for shared hash databases to provide a solution to the tracking and potential removal of deep fake multimedia across social media platforms.

Information integrity models must be trained, so the availability of data, computing resources, and financial resources to pay for data collection and labeling, the cloud computing and energy costs related to training AI systems, and other base factors influence how AI systems develop. Resource gaps can further exacerbate inequalities between low-income and high-income countries, the Global North and Global South, and dominance of specific languages, terminology, and meaning.

Global social media platforms must ensure that they have appropriate resources devoted to each country in which they operate, including language and country experts and trust and safety expertise. Given the centrality of Meta/Facebook/WhatsApp/Instagram, Google/YouTube, and Twitter to disinformation operations, in particular, these firms should take proactive measures in advance of national and local elections to deter coordinated state-aligned campaigns, for-hire moderation mercenaries, and other information operations. Other global platforms such as Wikipedia, TikTok, and Reddit should consider doing the same. These efforts should be paired with concurrent efforts to protect and elevate quality information and independent public interest journalism. In addition, more substantial financial support to journalistic and fact-checking organizations is needed in all countries, ideally through independent, non-governmental professional associations.

Policymakers must ensure any regulations on the use of AI do not restrict the ability of news organizations to incorporate AI into their journalism. Journalism associations and codes of ethics should require the labeling of AI-generated journalism to build greater audience literacy and improve transparency. Journalists and media organizations should commit to labeling all AI-generated content.

The purported "self-regulation" by social media and technology platforms has failed across a range

[95] See the discussion on these approaches at the Internet Governance Forum https://gfmd.info/trust-initiatives-as-the-future-of-news-media-sustainability/ and the International Journalism Festival "Reinventing the Big Tech vs journalism dynamic: trust and integrity" https://www.youtube.com/watch?v=-XTahPc7kO0&list=PLQlZwllJ41_DkBvgRc4r9URdqCXxyTS2F&index=3&t=735s

of issues. Regulators should consider imposing co-regulatory or government-regulated requirements on relevant technology platforms to mandate compliance with data privacy, which could reduce the cooptation of user data for targeted advertising used in some information operations and restrict how look-alike audiences are constructed and targeted by political operatives. Similarly, regulators could require more information about the allocation of platform resources to support content governance in relevant languages on their platforms, and commit to improving NLP machine learning for underrepresented languages in countries where they operate profitably. Overall, regulators should provide clear legal guidance to cultivate greater transparency and accountability as well as more effective oversight.

Regulators should impose more robust transparency requirements on social media platforms with respect to content moderation and curation, algorithmic decision making, use of AI systems and the datasets and languages, and country operations. This would include requirements related to the sharing of internal research as well as a framework for external independent research and audits. Given how difficult content moderation and context is, tech platforms should collect and provide independent researchers with access to the datasets that could test emerging models.

More research into the impact that coordinated disinformation and online harassment campaigns have on low-resourced languages, training data, and adversarial generative networks is needed. Collaboration between the private sector, government, academia, and civil society to fund and conduct analysis of major datasets used to train core NLP datasets (such as CommonCore, etc.) and creation of new ones to provide missing analysis is needed, and could be funded by an independent endowment, for example. Specifically, more extensive analysis is needed of: 1) the impact of undesirable content in the datasets used to train AI models on downstream performance; 2) the effect of properly filtering disinformation, harassment or other undesirable content out of the dataset before model training.[96] Understanding how hate speech, harassment, coordinated disinformation campaigns, and the like impact datasets and the basic building blocks of AI systems would be an important step towards addressing structural enablers of disinformation in low-resource languages and countries with small market power.

[96] Luccioni, Alexandra Sasha, and Joseph D. Viviano. "What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus." *ArXiv:2105.02732 [Cs]*, May 31, 2021. *http://arxiv.org/abs/2105.02732*.

ABOUT THE AUTHOR

Courtney C. Radsch, PhD, is a journalist, scholar and practitioner whose work focuses on the intersection of technology, media, and rights. Currently, she is a fellow at UCLA's Technology, Law and Policy Institute; a senior fellow at the Center for International Governance Innovation (CIGI) and the Center for Media, Data and Society (CMDS); and a visiting scholar at Annenberg's Center for Media at Risk. Her research focuses on internet governance and the geopolitics of technology, media sustainability and the future of journalism, and evolving socioeconomic and technopolitical effects of media and technology. She is the author of of *Cyberactivism and Citizen Journalism in Egypt: Digital Dissidence and Political Change* (Palgrave-Macmillan, 2016) based on her pioneering doctoral research and her work has been published in top media outlets and peer-reviewed journals. She is a frequent public speaker and media commentator including for CNN, Al Jazeera, NPR, and other global media outlets. Dr. Radsch has led advocacy missions to more than a dozen countries and has provided expert testimony to Congress, the OSCE, OECD, and the United Nations. Dr. Radsch's research and work are informed by a commitment to human rights and ensuring the sustainability of independent media.

Dr. Radsch specializes in transforming research and ideas into action while building cross-functional organizational strategies and alliances to advance policy objectives and knowledge. She spent seven years as Director of Advocacy and Communications at the Committee to Protect Journalists and previously worked at UNESCO and as a journalist in the Middle East. Dr. Radsch holds a Ph.D. in international relations from American University, a M.S. from Georgetown University and a B.A. from the University of California, Berkeley.