The Golden Circle: Creating Socio-technical Alignment in Content Moderation

Abhishek Gupta ¹  Iga Kozlowska ²  Nga Than ³

Abstract

This paper outlines a conceptual framework titled The Golden Circle that describes the roles of actors at individual, organizational, and societal levels, and their dynamics in the content moderation ecosystem. Centering “harm-reduction” and “context-moderation,” it argues that the ML community must attend to multi-modal content moderation solutions, align their work with their organizations’ goals and values, and pay attention to the ever changing social contexts in which their sociotechnical systems are embedded. This is done by accounting for the why, how, and what of content moderation from a sociological and technical lens.

1. Introduction

Content moderation has become an increasingly challenging sociotechnical issue as sociolinguistic norms change at an ever-increasing rate and more people than engage online. While social scientists have focused on social issues of content moderation like poor labor conditions (Roberts, 2019) (Sablosky, 2021), technologists have mainly focused on developing and perfecting AI solutions to remove harmful content from the information ecosystem (Vijayaraghavan et al., 2019). Moving beyond the dichotomy of the technical vs. the social, we develop the Golden Circle Framework (GCF), an interdisciplinary approach to tackle the content moderation problem, which takes into account both individual ML researchers, developers’ positions, and the collective action of organizations and other society actors.

2. The Golden Circle

A growing body of literature has pointed out that one should move beyond content moderation to “context-moderation” and “harm-reduction” (Caplan, 2018). The GCF, an adaptation of the Golden Circle by Simon Sinek (2009), provides a road map for evaluating content moderation harms and organizing responses to mitigate them on an individual, organizational, and societal level keeping in mind both the technical and sociological interventions. The central tenet of the GCF is that these vectors of effort (having both direction and magnitude as per the notion in Physics) need to be aligned to make content moderation efforts effective.

(a) Radialization: the availability of radical, extremist, violent and otherwise objectionable content online may lead to the radicalization of political or cultural views, whether through organized radicalization efforts such as ISIS recruitment networks or less organized methods like the spread of misinformation and disinformation (e.g. QAnon conspiracy theories) (Zuckerman & Gessen, 2019; Lewis, 2018). Radicalization and inci-
attention to hatred or violence result from harmful content not being detected before it is posted or not being taken down after it has been posted.

(b) **Polarization:** the “echo chamber” effect results in users seeing more and more content that they are already more prone to engage with, over time resulting in highly homogeneous content consumption (Bail et al., 2018). Polarized political worldviews result in not only high levels of mistrust of the “other” but also different versions of the truth, or different epistemologies (Freelon et al., 2020). No longer able to agree on basic facts and a common reality, the public suffers polarization that makes democratic participation, deliberation, and decision-making difficult. Polarization results not necessarily from harmful content not being removed, but the patterns of information that are systematically served to some and not others (Liao & Fu, 2014).

(c) **Online harassment:** online bullying, harassment, and discrimination, often of marginalized groups or individuals based on race, gender, sexual orientation, disability, religion, or ethnicity, whether targeted at specific individuals or entire groups such as Rohingya Muslims (Siddiquee, 2020), results in exclusion, marginalization, and in the most extreme cases physical violence, and even genocide. This virtual (Patton et al., 2014) and physical violence results from harmful content not being detected before it is posted or not being taken down after it has been posted.

(d) **Attention control:** online content has been shown to be addictive by design. Internet “influencers” produce content that could go “viral” to get more followers; companies post “click-bait” to increase ad revenue (Wu, 2017). Ubiquitous social media content production and consumption result in increased feelings of isolation, decreasing ability to focus, feelings of a sense of loss of self-control, and the opportunity cost of not engaging in more healthy or productive activities (Sujarwoto et al., 2019). Attention deficit results from the sheer volume of content production and the commercial incentives to generate attention-grabbing and low-quality content (volume over quality).

(e) **Labor:** moderating online content incurs emotional and psychological costs on low-paid wage workers, who often reside in the developing world, to screen objectionable content (Roberts, 2019).

(f) **Environmental:** Automated content moderation systems which utilize large language models require a sprawling infrastructure that uses up natural resources and vast amounts of energy to operate (Bender et al., 2021). They incur environmental costs that must be taken into account when deciding how to regulate content production and moderation.

Why have these well documented harms not been adequately addressed? Our framework shows that even though everyone may agree on the “Why” (healthy information ecosystem), the “How” and the “What” are not aligned. In other words, though few would disagree with the need to maintain social trust and a healthy information ecosystem, current solutions to content moderation are not aligned with how individuals and organizations operate, especially when it comes to their incentives.

Our framework provides a road map for realigning content moderation solutions to the motivating “Why” while respecting the fundamental “How” of individuals and organizations that are largely immutable. The “How” for ML developers is to seek job rewards like promotions and benefits. Yet organizations don’t incentivize harm reduction through improve content moderation, so sufficient technical improvements are not innovated. There is a feedback loop from organization to individual. Similarly, the “How” of organizations is to maximize profit while avoiding negative public opinion. Yet society, through civic and legislative action, has not demanded sufficient organizational change. There is a feedback loop from society to organization. Each stakeholder level feedback enables the other and must be aligned. Then, and only then, can technical and sociological solutions - the “What” - be successfully implemented.

We propose this framework not as a single, one-time “solution” to content moderation but rather as a way to better align the vectors to drive toward outcomes that reduce harm, acknowledging that eliminating all harm completely is unrealistic. Let’s turn now to the “What” - proposed solutions to reduce content moderation harms.

3. **Individual**

What can those who build algorithms that power social media experiences do about it? ML developers are uniquely positioned to offer technical solutions to sociotechnical problems, so their improvements are only a piece of the puzzle, but let us consider some proposals.

To reduce the amount of harmful content circulating in the information ecosystem, we need to improve the accuracy of models in identifying extremist, violent, or otherwise objectionable content, whether text, image, video or more and more some combination of each. The difficulties lie in algorithms understanding cultural references and symbols that are unspecified or implicit. To understand this nuance, models need to be better able to capture the cultural context within which meaning is socially produced (Caplan, 2018). For example, Gao & Huang (2017) improved logis-
The Golden Circle: Creating Sociotechnical Alignment in Content Moderation

| Why                                                                 | Who                                    | How                                                      | What                                                                 |
|----------------------------------------------------------------------|----------------------------------------|----------------------------------------------------------|----------------------------------------------------------------------|
| Preserve social cohesion and social trust sufficient to sustain a robust and fair democracy | Individual: ML developer designing content moderation algorithms | Seek job rewards like promotion and benefits | Technical: Improve algorithms used in content moderation. Sociological: ML developers should recognize their own implicit biases as well as technical limitations in solving what is a complex social problem |
| Organization: social media platform                                  | Maximize profit while avoiding negative public opinion | Sociological: • Increase resourcing for ethical AI work and diversify AI workforce • Build interdisciplinary teams • Restructure ML developer incentives to reward human-centered design • Develop business models that more fairly distributes platform benefits • Foster a culture of transparency, integrity, and accountability |
| Society                                                             | Sustain a healthy information ecosystem through better content moderation | Sociological: Hold social media platforms account-able through regulatory and policy instruments NGOs and academics must continue to scrutinize content moderation practices Sustain robust public dialogue on the role of social media in society |

Table 1. Description of the role of each actor within the Golden Circle Framework

tic regression and neural net hate speech detection models by introducing contextual features like screen name, full comment thread, and the news article associated with each thread. They enlarged the object of analysis by including culturally-relevant content as features in the model. Improved model accuracy will also have the spillover effect of minimizing as much as possible the necessity for human review of objectionable content. The unfair toll on human reviewers has been well documented (Roberts, 2019).

Other machine learning innovations can have positive indirect effects on content moderation like attempts at reducing the echo chamber effect. Recommendation algorithms need to be redesigned to optimize for a healthy plurality of content (Sheth et al., 2011) rather than the spiral of homogeneous content that is the product of algorithms that optimize chiefly for views and clicks. Similarly, these algorithms need to consider features that can help reduce the addictiveness of content, focusing less on virality more on quality content (Del Olmo & Gaudioso, 2008). This can also help reduce the incentives to serve customers more and more radical content instead privileging more balanced and higher quality content. Finally, the environmental impacts of running large language models used in content production and moderation have been well documented. Algorithmic efficiency up and down the whole production chain must be improved or new solutions introduced to reduce environmental burden (Lacoste et al., 2019).

Recognizing that ML developers have a bias for technical solutions, next we need to expand the context within which models get built. This requires examining the organizational structures of social media platforms.

4. Organization

How can more research like what discussed above and additional innovations be encouraged and incentivized within social platform organizations? Currently there is a misalignment between organizational goals and culture, incentivizing ML teams to build bigger and better language models (Chelba et al., 2012) that are better at targeting content to users, not necessarily reducing harmful content or otherwise improving content quality. Unfortunately, there is limited available research on the kinds of organizational processes, tools or cultures that are most practicable and conducive to more ethical algorithmic solutions (Baucus & Beck-Dudley, 2005). Some innovative social scientific research, however, is moving in that direction. Moss & Metcalf (2020) published a report on “ethics owners,” those in tech orga-
nizations assigned to address ethical concerns, champion ethical causes, designs, development, and deployment of technology from within the tech industry. Central to our argument, they find that the “personal ethics of the ethics owner do not always align with those of the corporation” and that organization practices and business models “have implications for how ethics owners approach their work.”

Our framework proposes several possibilities by which social media platforms can align content moderation practices to more ethical outcomes while still maintaining profitability. These include increasing resources for ethical AI work and diversifying the AI workforce. An extensive body of research shows that more diverse teams build better products, increase sales revenue, leading to better organizational outcomes (Page, 2008; 2019; Muller et al., 2019; Harrington, 2010; Jehn et al., 1999; Herring, 2009). Investing in these areas should not cut into the bottom line. Second, organizations need to build interdisciplinary teams composed of technical and non-technical roles so that social implications of poor content moderation can be brought to the fore. This does not require hiring external personnel, but rather expanding the kind of experts that have a seat at the table. Third, organizations must restructure ML developer incentives to reward the development of algorithms that put human well-being at the center. This can include alternative ways of measuring employee progress or impact by de-emphasizing speed and quantity and instead focusing on harm reduction, alignment with the latest social scientific research, and customer satisfaction. Fourth, organizations should explore business models that more fairly distributes platform benefits. This could include compensation for content generators or data subjects while incentivizing quality content (Wohn, 2019). Technology companies routinely update, tweak, or “pivot,” and sometimes completely invent new business models (Shestakofsksy, 2017; Ravenelle, 2019; Griesbach et al., 2019). Just because the current advertising-based business model of social media platforms is the most widespread and profitable does not mean it is the only viable model possible. Investing in business model innovation will create incentives for model developers to think more creatively about how machine learning can be employed for different purposes than it currently is. Finally, organizations must foster a culture of transparency, integrity, and accountability where critical research and criticism of company policy is not only tolerated but encouraged. ML researchers and developers cannot propose innovative solutions without first being able to point to what is not working.

6. Conclusion

To conclude, efforts like ours and many others (Sloane & Moss, 2019; Lindgren & Holmström, 2020) at bridging the gap between machine learning experts and social science researchers (and the broader public) strive to better align the practices of all stakeholders (ML developers, organizations and society) for the purpose of generating improved social outcomes. Our framework points out current day misalignments so that we can address these gaps in a more organized manner. This is needed because the harms of an unhealthy information ecosystem are real, as evidenced by recent events like the attack on the U.S. Capitol. Our framework uses the case study of content moderation as one field within which ML solutions are applied. But it has significance for other sociotechnical AI-based problems as ML becomes applied to more and more aspects of our social lives.

5. Society

Broadening our lens even further, and to force the kind of realignments mentioned above, here we propose some actions for legislators, NGOs and academe.
References

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., and Volfovsky, A. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

Baucus, M. S. and Beck-Dudley, C. L. Designing ethical organizations: Avoiding the long-term negative effects of rewards and punishments. *Journal of Business Ethics*, 56 (4):355–370, 2005.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT*, 2021.

Caplan, R. Content or context moderation? artisanal, community-reliant, and industrial approaches. *Data & Society*. https://datasociety.net/library/content-or-context-moderation, 2018.

Chelba, C., Bikel, D., Shugrina, M., Nguyen, P., and Kumar, S. Large scale language modeling in automatic speech recognition. *arXiv preprint arXiv:1210.8440*, 2012.

Del Olmo, F. H. and Gaudioso, E. Evaluation of recommender systems: A new approach. *Expert Systems with Applications*, 35(3):790–804, 2008.

Freelon, D., Marwick, A., and Kreiss, D. False equivalencies: Online activism from left to right. *Science*, 369 (6508):1197–1201, 2020.

Gao, L. and Huang, R. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*, 2017.

Gorwa, R., Binns, R., and Katzenbach, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, 2020.

Griesbach, K., Reich, A., Elliott-Negri, L., and Milkman, R. Algorithmic control in platform food delivery work. *Socius*, 5:2378023119870041, 2019.

Hameleers, M. and van der Meer, T. G. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, 47(2):227–250, 2020.

Harrington, B. *Pop finance: Investment clubs and the new investor populism*. Princeton University Press, 2010.

Herring, C. Does diversity pay?: Race, gender, and the business case for diversity. *American sociological review*, 74(2):208–224, 2009.

Jehn, K. A., Northcraft, G. B., and Neale, M. A. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly*, 44(4):741–763, 1999.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Langvardt, K. Regulating online content moderation. *Geo. LI*, 106:1353, 2017.

Lewis, R. Alternative influence: Broadcasting the reactionary right on youtube. *Data & Society*, 18, 2018.

Liao, Q. V. and Fu, W.-T. Can you hear me now? mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 184–196, 2014.

Lindgren, S. and Holmström, J. A social science perspective on artificial intelligence: Building blocks for a research agenda. *Journal of Digital Social Research*, 2(3):1–15, 2020.

Moss, E. and Metcalf, J. Ethics owners: A new model of organizational responsibility in data-driven technology companies. 2020.

Muller, M., Fussell, S. R., Gao, G., Hinds, P. J., Oliveira, N., Reinecke, K., Robert Jr, L., Siangliulue, K., Wulf, V., and Yuan, C.-W. Learning from team and group diversity: Nurturing and benefiting from our heterogeneity. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pp. 498–505, 2019.

Page, S. E. *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press, 2008.

Page, S. E. *The diversity bonus: How great teams pay off in the knowledge economy*. Princeton University Press, 2019.

Patton, D. U., Hong, J. S., Ranney, M., Patel, S., Kelley, C., Eschmann, R., and Washington, T. Social media as a vector for youth violence: A review of the literature. *Computers in Human Behavior*, 35:548–553, 2014.

Ravenelle, A. J. *Hustle and gig: Struggling and surviving in the sharing economy*. Univ of California Press, 2019.
The Golden Circle: Creating Sociotechnical Alignment in Content Moderation

Sablosky, J. Dangerous organizations: Facebook’s content moderation decisions and ethnic visibility in myanmar. *Media, Culture & Society*, pp. 0163443720987751, 2021.

Schradie, J. *The revolution that wasn’t: How digital activism favors conservatives*. Harvard University Press, 2019.

Shestakofsky, B. Working algorithms: Software automation and the future of work. *Work and Occupations*, 44(4): 376–423, 2017.

Sheth, S. K., Bell, J. S., Arora, N., and Kaiser, G. E. Towards diversity in recommendations using social networks. 2011.

Siddiquee, M. A. The portrayal of the rohingya genocide and refugee crisis in the age of post-truth politics. *Asian Journal of Comparative Politics*, 5(2):89–103, 2020.

Sinek, S. *Start with why: How great leaders inspire everyone to take action*. Penguin, 2009.

Sloane, M. and Moss, E. Ai’s social sciences deficit. *Nature Machine Intelligence*, 1(8):330–331, 2019.

Sujarwoto, S., Tampubolon, G., and Pierewan, A. C. A tool to help or harm? online social media use and adult mental health in indonesia. *International journal of mental health and addiction*, 17(4):1076–1093, 2019.

Sylvia Chou, W.-Y., Gaysynsky, A., and Cappella, J. N. Where we go from here: health misinformation on social media, 2020.

Vijayaraghavan, P., Larochelle, H., and Roy, D. Interpretable multi-modal hate speech detection. In *AI for Social Good Workshop at the International Conference on Machine Learning*, 2019.

Wohn, D. Y. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–13, 2019.

Wu, T. *The attention merchants: The epic scramble to get inside our heads*. Vintage, 2017.

Zuckerman, E. and Gessen, M. Unreality and social corrosion: Masha gessen and ethan zuckerman in conversation. *Journal of Design and Science*, (6), 2019.