SM+S
social media + society

Social Media + Society
July-September 2020: 1–5
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2056305120948186
journals.sagepub.com/home/sms
SAGE

The COVID-19 Mental Health Content Moderation Conundrum

Ysabel Gerrard

Abstract

At the time of writing (mid-May 2020), mental health charities around the world have experienced an unprecedented surge in demand. At the same time, record-high numbers of people are turning to social media to maintain personal connections due to restrictions on physical movement. But organizations like the mental health charity Mind and even the UK Government have expressed concerns about the possible strain on mental health that may come from spending more time online during COVID-19. These concerns are unsurprising, as debates about the link between heavy social media use and mental illness raged long before the pandemic. But our newly heightened reliance on platforms to replace face-to-face communication has created even more pressure for social media companies to heighten their safety measures and protect their most vulnerable users. To develop and enact these changes, social media companies are reliant on their content moderation workforces, but the COVID-19 pandemic has presented them with two related conundrums: (1) recent changes to content moderation workforces means platforms are likely to be less safe than they were before the pandemic and (2) some of the policies designed to make social media platforms safer for people's mental health are no longer possible to enforce. This Social Media+Society: 2K essay will address these two challenges in depth.

Keywords

COVID-19, mental health, social media, content moderation

Mental health charities around the world have experienced an unprecedented surge in demand over the past few weeks and months. In the United Kingdom, for example, the Beat Eating Disorders (2020) charity saw a 50% increase in requests for its services since the nation-wide lockdown was first enforced, and calls to mental health charities like SANE and Anxiety UK were up by 200% at the start of May 2020 (Stephens, 2020). In the absence of access to professional support (Campbell, 2020), mental health apps have been downloaded more than 1 million times since the United Kingdom's lockdown measures began in March 2020 (Chowdhury, 2020). At the same time, record-high numbers of people have turned to social media to maintain personal connections due to restrictions on physical movement (Newton, 2020a). But organizations like Mind (2020) and even the UK Government (GOV.UK, 2020a) have expressed concerns about the possible strain on mental health that may come from spending more time online during COVID-19.

These concerns are unsurprising, as debates about the link between heavy social media use and mental illness raged long before the pandemic. But our newly heightened reliance on platforms to replace face-to-face communication has created even more pressure for social media companies to heighten their safety measures and protect their most vulnerable users. The pandemic has also widened the net of vulnerability: the increase in demand for mental health services could suggest that people without prior conditions are now struggling. We are indeed witnessing what the United Nations has called a "mental health emergency" (Kelly-Linden, 2020).

To develop and enact these changes, social media companies are reliant on their content moderation workforces, but the COVID-19 pandemic has presented a number of unprecedented challenges to their ongoing efforts. Content moderation is largely enforced by humans who spend their shifts reviewing user reports and "soak[ing] up the worst of humanity in order to protect the rest of us" (Chen, 2014, n.p.; see also Roberts, 2019). Social media companies also employ in-house policy teams who are responsible for setting and enforcing the parameters of "acceptable" social media conduct (Gillespie, 2018), like developing the rulebooks

The University of Sheffield, UK

Corresponding Author:
Ysabel Gerrard, Department of Sociological Studies, The University of Sheffield, Western Bank, Sheffield S10 2TU, UK.
Email: y.gerrard@sheffield.ac.uk

moderators use to respond to user reports (Hopkins, 2017), and enforcing in-platform restrictions like limiting the search results for particular hashtags (Gerrard, 2018) or "shadow-banning" users (Myers West, 2018).

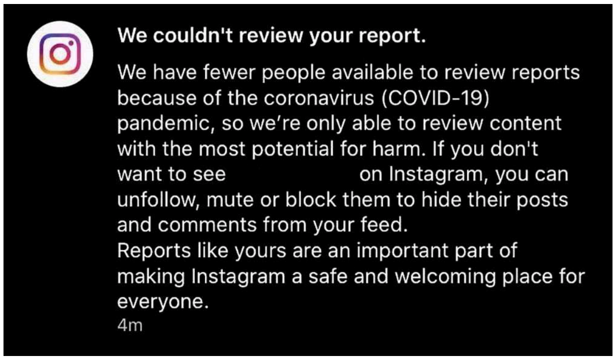
In the context of the COVID-19 pandemic, content moderation workforces face two related conundrums, the effects of which are still likely to be felt when/if the pandemic subsides: (1) recent changes to content moderation workforces means platforms are likely to be *less* safe than they were before the pandemic, and (2) some of the policies designed to make social media platforms safer for people's mental health are no longer possible to enforce. The remainder of this short paper will address these two challenges in depth.

Furloughing the Front Line of Social Media

In late March 2020, news broke that major social media companies like Facebook, Twitter, and YouTube had sent their human content moderators home "until further notice": a role that is "often difficult, if not impossible, to do from home" (Matsakis & Martineau, 2020, n.p.). Many platforms are now relying on artificial intelligence (AI) to take down problematic posts, but this was a near-instant problem (Roberts, 2017). For example, *WIRED* reported that links to articles from legitimate news outlets like *The Atlantic* and *BuzzFeed* had been wrongly removed for violating Facebook's spam rules, which the platform vaguely attributed to a "bug" (Matsakis & Martineau, 2020, n.p.). While the stakes are high for content like mis/disinformation, platforms also cannot afford to inadequately moderate mental health content at a time when so many of their users are at their most vulnerable.

Some social media companies are aware that AI is ill-equipped to moderate mental health-related content (Gerrard, 2018): a moment of transparency praised by the Electronic Frontier Foundation's (EFF) York and McSherry (2020). But York and McSherry (2020) also warn platforms against relying on AI when the world returns to some version of normal: in their words, "history suggests that protocols adopted in times of crisis often persist when the crisis is over" (n.p.). YouTube has warned users that AI might mistakenly remove videos (YouTube, 2020); Facebook says humans will continue to work on suicide and self-injury prevention (Zuckerberg, 2020, p. 12), and Instagram will still ask humans to review "content with the most potential for harm" (Figure 1).

Much less is known about how platforms like Weibo, WeChat, and VK are handling their content moderation workforces during the pandemic. Western press discourse about non-western platforms tends to focus on the censorship of coronavirus-related content as opposed to changes to their content moderation workforces (BBC News, 2020a): what Newitz (2020) calls "the most important job on the internet."



We couldn't review your report.

We have fewer people available to review reports because of the coronavirus (COVID-19) pandemic, so we're only able to review content with the most potential for harm. If you don't want to see [redacted] on Instagram, you can unfollow, mute or block them to hide their posts and comments from your feed.
Reports like yours are an important part of making Instagram a safe and welcoming place for everyone.
4m

Figure 1. Screengrab of an automated response to a reported Instagram post.

But what counts as "content with the most potential for harm"? For example, the eating disorder Anorexia has the highest death rate of any psychiatric condition (Quinn, 2020), but would posts about the promotion of eating disorders—rampant across some newer platforms like TikTok (Gerrard, 2020)—be prioritized according to these new rules? In an article I co-authored with McCosker (McCosker and Gerrard, 2020), we found that people who talk about depression on Instagram largely do so through pseudonymised, humorous meme accounts. Is AI alone capable of reading into these carefully coded contextual cues to detect the necessity of urgent intervention? Sadly, I suspect not.

One of my biggest concerns is that human content moderators struggled to find the time to deal with the onslaught of user reports *before* the pandemic (Roberts, 2017). The UK government has advised people who see "harmful content" on social media to "report it to the site" (GOV.UK, 2020a), and although this seems like the best advice on the surface, it glosses over the workload problems social media giants openly admit they're facing. If workforces have been reduced *and* we are in a "mental health emergency" (Kelly-Linden, 2020), it's incredibly unlikely that the remaining moderator workforces at any major social media company will have the time to deal with the current volume of user reports. The consequences of this could be dire.

The COVID-19 pandemic also has implications for the remaining moderators' own mental health. To ask someone to review the worst content the Internet has to offer during a pandemic is a terrifying, borderline unethical prospect. Roberts' (2019) decade-long research on content moderation has revealed the prevalence of post-traumatic stress disorder (PTSD) among reviewers; in fact, in May 2020, Facebook agreed to pay a landmark US$52 million "to current and former moderators to compensate them for mental health issues developed on the job" (Newton, 2020b). Although mental health content moderation was far from "complete" (which, I would argue, it never could be), its effectiveness has sadly declined at a time when it is most necessary.

Widening the Net of Vulnerability

The second, related conundrum social media platforms face is the undoing of their previous mental health content moderation policies, some of which are now dated and others simply unfeasible. Before the pandemic, some globally dominant platforms like Instagram, Pinterest, and TikTok expanded their safety efforts by teaming with independent experts. Instagram, for example, has a Suicide and Self-Injury (SSI) Advisory Board (Facebook, 2020),[1] and Pinterest has teamed up with experts to design a set of well-being exercises for users who search for self-injury-related terms: "When a pinner enters a related search term, the site will surface a prompt for these exercises" (Pardes, 2019, n.p.). Governments, activists, health professionals, journalists, academics, and other public figures are placing increased pressure on social media companies to minimize the risk of harm that might befall their most vulnerable users. Noteworthy examples from the United Kingdom include the Online Harms White Paper, which takes the first step in developing a new regulatory framework for online safety and "make clear companies' responsibilities to keep UK users, particularly children, safer online" (GOV.UK, 2020b, n.p.). The tragic suicide of British teenager Molly Russell in 2017 led to another wave of policy alterations at major platforms, mainly Instagram, and which included the introduction of "sensitivity screens" to warn users a post contains sensitive content (Hern, 2019).

While my argument is not that these efforts have been undermined, as the principle of minimizing the risk of online harms extends beyond the pandemic, my point is that the *moments of intervention* have temporarily changed. For example, policymakers across numerous platforms—Instagram, Pinterest, TikTok, Tumblr, to name a few—have long chased harmful hashtags and restricted users' access to them. But Chancellor et al. (2016) found that users develop code words to work around bans. As an example, the tag #proana (a portmanteau term to denote the promotion of anorexia) might become #proanaaa. A lot of work goes into identifying these terms and then chasing down related tags, but I worry that the code words will have changed in tandem with people's mental health experiences. As more content moderation centers reopen (BBC News, 2020b), it is important to acknowledge that some of the original actions moderators took will no longer work, and policymakers likely don't have enough information about the link between COVID-19 and mental health to adapt accordingly (and quickly).

Suicide prevention efforts represent another change to moments of intervention. Since 2017, Facebook has contacted first responders to conduct "wellness checks" on people who the platform's AI systems and human moderators identify as being at imminent risk of suicide (Zuckerberg, 2018). This particular practice has faced intense criticism, including concerns about the risk of false positives (people who are wrongly identified as being suicidal) and the consequences of such an error, which include Facebook users undergoing unnecessary psychiatric evaluation (Thielking, 2019). But this practice—rightly or wrongly implemented by the platform—is simply unfeasible in the current climate, as first responders in most places around the world are overwhelmed.

Social media companies already had a long way to go in their efforts to protect their users: should healed self-harm scars be censored? What should happen to "borderline" content (posts that don't quite break the rules but sound alarm bells anyway)? How long should a suicide note stay up for? How can moderators be sure a post "promotes" an eating disorder? But the goal posts have shifted. The work that goes into answering these questions—including qualitative and quantitative information about people's experiences of mental health conditions—is no longer entirely applicable.

Social Media: A Psychiatrist's Biggest Ally?

The word "unprecedented" is ubiquitous in the current climate, and for good reason. This is indeed an unprecedented situation and we don't yet know how it will affect people's mental health. What we do know with certainly is that we will feel the repercussions of alterations to social media's content moderation workforce for years to come. For perhaps the first time, the reduction to human content moderation has vividly brought to light "the traces, which are so often hidden, of human intervention" (cited in Matsakis & Martineau, 2020, n.p.). These traces include errors and blind spots, and once again remind us how impossibly traumatic this job is for the human content moderator workforce. The COVID-19 pandemic also renews debates about platforms' parameters of responsibility: where does their responsibility for users' mental health start and end? Who should be responsible for overseeing their interventions?

Chaudhary and Vasan (2020) believe that technology and social media companies are "uniquely suited to be a psychiatrist's biggest ally in our mission to improve mental health for the 2 billion people around the world struggling with brain and behavioral health disorders" (n.p.). I mostly agree, and in the throes and aftermath of the COVID-19 pandemic, researchers, medical professionals, and tech company workers need to commit to working together, sharing resources, and possessing a genuine, moral desire to help social media platforms' increasingly vulnerable global userbase.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Ysabel Gerrard https://orcid.org/0000-0003-1298-9365

Note

1. I have been a member of Facebook and Instagram's Suicide and Self-Injury (SSI) Advisory Board since March 2019. At present, this is an unpaid position and I predominantly advise on Facebook and Instagram's policies about eating disorder content.

References

Beat Eating Disorders. (2020). *Emergency appeal*. https://donate.beateatingdisorders.org.uk/page/56677/donate/

BBC News. (2020a, March 4).Coronavirus: Chinese app WeChat censored virus content since 1 Jan. *BBC News*. https://www.bbc.co.uk/news/world-asia-china-51732042

BBC News. (2020b, April 30). Coronavirus: Facebook reopens some moderation centres. *BBC News*. https://www.bbc.co.uk/news/technology-52491123

Campbell, D. (2020, May 7). Mental health patients in crisis because of coronavirus cutbacks. *The Guardian*. https://www.theguardian.com/society/2020/may/07/mental-health-patients-in-crisis-because-of-coronavirus-cutbacks

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016, 27 February–2 March). #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1201–1213). Association for Computing Machinery.

Chaudhary, N., & Vasan, N. (2020, February 12). 3 ways for big tech to protect teens from harm. *WIRED*. https://www.wired.com/story/opinion-3-ways-for-big-tech-to-protect-teens-from-harm/

Chen, A. (2014, October 23). The laborers who keep dick pics and beheadings out of your Facebook feed. *WIRED*. https://www.wired.com/2014/10/content-moderation/

Chowdhury, H. (2020, May 16). Mental health apps downloaded more than 1m times since start of virus outbreak. *The Guardian*. https://www.telegraph.co.uk/technology/2020/05/16/mental-health-apps-downloaded-1m-times-since-start-virus-outbreak/

Facebook. (2020). *Suicide prevention: Expert engagement*. https://www.facebook.com/safety/wellbeing/suicideprevention/expertengagement

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4511.

Gerrard, Y. (2020, March 9). TikTok has a pro-anorexia problem. *WIRED*. https://www.wired.com/story/opinion-tiktok-has-a-pro-anorexia-problem/

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

GOV.UK. (2020a, April 3). *Coronavirus (COVID-19): Staying safe online*. https://www.gov.uk/guidance/covid-19-staying-safe-online

GOV.UK. (2020b, February 12). *Online harms white paper*. https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

Hern, A. (2019, February 4). Instagram to launch "sensitivity screens" after Molly Russell's death. *The Guardian*. https://www.theguardian.com/technology/2019/feb/04/instagram-to-launch-sensitivity-screens-after-molly-russell-death

Hopkins, N. (2017, May 21). Revealed: Facebook's internal rulebook on sex, terrorism and violence. *The Guardian*. https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

Kelly-Linden, J. (2020, May 14). Coronavirus pandemic has triggered a mental health emergency, UN warns. *The Telegraph*. https://www.telegraph.co.uk/global-health/science-and-disease/coronavirus-pandemic-has-triggered-mental-health-emergency-un/

Matsakis, L., & Martineau, P. (2020, March 10). Coronavirus disrupts social media's first line of defense. *WIRED*. https://www.wired.com/story/coronavirus-social-media-automated-content-moderation/

McCosker, A., & Gerrard, Y. (2020). Hashtagging depression on Instagram: Towards a more inclusive mental health research methodology. *New Media & Society*. Advance online publication. https://doi.org/10.1177/1461444820921349.

Mind. (2020). *Online mental health*. https://www.mind.org.uk/information-support/tips-for-everyday-living/online-mental-health/safety-privacy/

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383.

Newitz, A. (2020, March 13). We forgot about the most important job on the internet. *The New York Times*. https://www.nytimes.com/2020/03/13/opinion/sunday/online-comment-moderation.html

Newton, C. (2020a, March 19). How Facebook is preparing for a surge in its depressed and anxious users. *The Verge*. https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

Newton, C. (2020b, May 12). Facebook will pay $52 million in settlement with moderators who developed PTSD on the job. *The Verge*. https://www.theverge.com/2020/3/19/21185204/facebook-coronavirus-depression-anxiety-content-moderation-mark-zuckerberg-interview

Pardes, A. (2019, November 14). Pinterest has a new plan to address self-harm. *WIRED*. https://www.wired.com/story/pinterest-self-harm-help/

Quinn, T. (2020, January 3). Anorexia is the deadliest mental illness: Why is the NHS still not taking it seriously? *The Guardian*. https://www.theguardian.com/commentisfree/2020/jan/03/anorexia-mental-illness-nhs-hospital-admissions

Roberts, S. T. (2017, March 8). Social media's silent filter. *The Atlantic*. https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Stephens, M. (2020, May 1). Risk of rise in agoraphobia due to lockdown as anxiety charities extend helpline hours. *The Telegraph*. https://www.telegraph.co.uk/news/2020/05/01/risk-rise-agoraphobia-due-lockdown-anxiety-charities-extend/

Thielking, M. (2019, February 11). ‘We don’t have any data’: Experts raise questions about Facebook’s suicide prevention tools. *STAT News*. https://www.statnews.com/2019/02/11/facebook-suicide-prevention-tools-ethics-privacy/

York, J. C., & McSherry, C. (2020, April 2). *Automated moderation must be temporary, transparent and easily appealable*. Electronic Frontier Foundation (EFF). https://www.eff.org/deeplinks/2020/04/automated-moderation-must-be-temporary-transparent-and-easily-appealable

YouTube. (2020, March 16).Protecting our extended workforce and the community [Blog post]. https://youtube-creators.googleblog.com/2020/03/protecting-our-extended-workforce-and.html

Zuckerberg, M. (2018, November 15). A blueprint for content governance and enforcement. *Facebook*. https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

Zuckerberg, M. (2020, March 18). *Transcript of Facebook March 18 press call*. https://about.fb.com/wp-content/uploads/2020/03/March-18-2020-Press-Call-Transcript.pdf

Author Biography

Ysabel Gerrard (PhD University of Leeds) is a lecturer in Digital Media and Society at the University of Sheffield. She has two current areas of research interest: social media content moderation and secret-telling social media apps. She is the Book Reviews Editor for *Convergence: The International Journal of Research into New Media Technologies* and the Vice Chair of ECREA’s Digital Culture and Communication section. She has published her work in journals like *New Media and Society* and *First Monday*, and her research and policy interventions have been cited in international venues like *BBC News, The Washington Post*, and *WIRED*.