## Abstract

At the time of writing (midMay 2020), mental health charities around the world have experienced an unprecedented surge in demand. At the same time, recordhigh 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 COVID19. 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 facetoface 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 COVID19 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 peoples mental health are no longer possible to enforce. This Social Media Society: 2K essay will address these two challenges in depth.

## Keywords

COVID19, 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 nationwide 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 Kingdoms lock down measures began in March 2020 (Chowdhury, 2020). At the same time, recordhigh numbers of people have turned to social media to maintain personal connections due to restric tions on physical movement (Newton, 2020a). But organiza tions 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 COVID19.

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 facetoface 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 (KellyLinden, 2020).

To develop and enact these changes, social media compa nies are reliant on their content moderation workforces, but the COVID19 pandemic has presented a number of unprec edented challenges to their ongoing efforts. Content modera tion is largely enforced by humans who spend their shifts reviewing user reports and soak[ing] up the worst of human ity in order to protect the rest of us (Chen, 2014, n.p.; see also Roberts, 2019). Social media companies also employ inhouse policy teams who are responsible for setting and enforcing the parameters of acceptable social media con duct (Gillespie, 2018), like developing the rulebooks moderators use to respond to user reports (Hopkins, 2017), and enforcing inplatform restrictions like limiting the search results for particular hashtags (Gerrard, 2018) or shadow banning users (Myers West, 2018).

In the context of the COVID19 pandemic, content mod eration workforces face two related conundrums, the effects of which are still likely to be felt when/if the pandemic sub sides: (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 peoples 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 com panies 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 nearinstant 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 Facebooks spam rules, which the platform vaguely attrib uted to a bug (Matsakis & Martineau, 2020, n.p.). While the stakes are high for content like mis/disinformation, plat forms 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 healthrelated content (Gerrard, 2018): a moment of transparency praised by the Electronic Frontier Foundations (EFF) York and McSherry (2020). But York and McSherry (2020) also warn platforms against rely ing 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 con tinue to work on suicide and selfinjury 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 nonwestern platforms tends to focus on the censor ship of coronavirusrelated content as opposed to changes to their content moderation workforces (BBC News, 2020a): what Newitz (2020) calls the most important job on the internet.

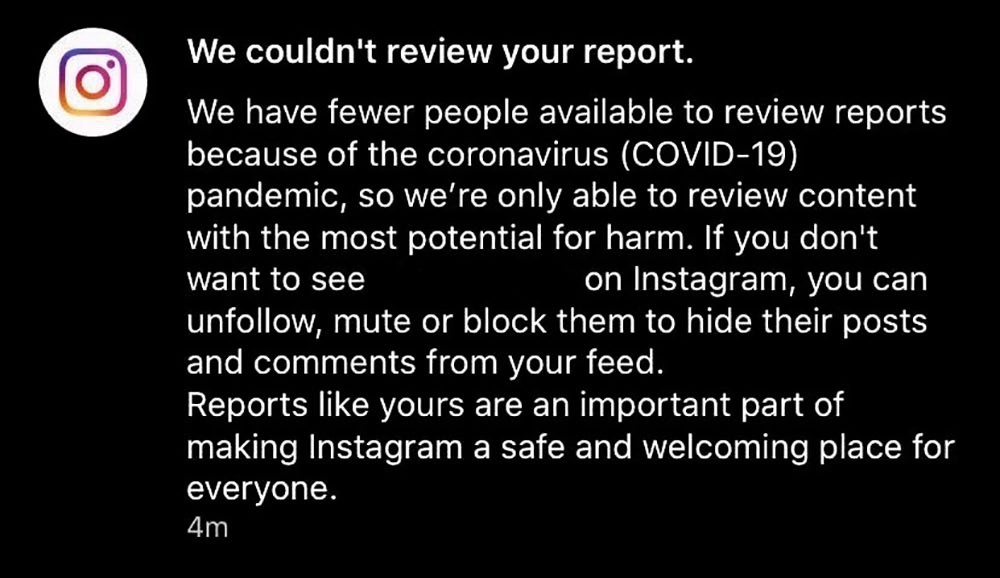


Figure 1. Screengrab of an automated response to a reported Instagram post.

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 coauthored 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 mod erators 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 theyre facing. If workforces have been reduced *and* we are in a mental health emergency (KellyLinden, 2020), its 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 con sequences of this could be dire.

The COVID19 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) decadelong research on content modera tion has revealed the prevalence of posttraumatic stress disorder (PTSD) among reviewers; in fact, in May 2020, Facebook agreed to pay a landmark US$52 million to cur rent 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 mod eration policies, some of which are now dated and others simply unfeasible. Before the pandemic, some globally dom inant 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),and Pinterest has teamed up with experts to design a set of wellbeing exercises for users who search for selfinjuryrelated 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 exam ple, 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 port manteau term to denote the promotion of anorexia) might become. 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 peoples mental health experiences. As more content moderation centers re open (BBC News, 2020b), it is important to acknowledge that some of the original actions moderators took will no longer work, and policymakers likely dont have enough information about the link between COVID19 and mental health to adapt accordingly (and quickly).

Suicide prevention efforts represent another change to moments of intervention. Since 2017, Facebook has con tacted first responders to conduct wellness checks on people who the platforms AI systems and human modera tors 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 psychiat ric evaluation (Thielking, 2019). But this practice rightly or wrongly implemented by the platform is simply unfea sible 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 selfharm scars be censored? What should happen to borderline con tent (posts that dont 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 dis order? But the goal posts have shifted. The work that goes into answering these questions including qualitative and quantitative information about peoples experiences of men tal health conditions is no longer entirely applicable.

# Social Media: A Psychiatrists Biggest Ally?

The word unprecedented is ubiquitous in the current cli mate, and for good reason. This is indeed an unprecedented situation and we dont yet know how it will affect peoples mental health. What we do know with certainly is that we will feel the repercussions of alterations to social medias content moderation workforce for years to come. For per haps the first time, the reduction to human content modera tion 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 COVID19 pandemic also renews debates about platforms parameters of responsibility: where does their responsibility for users mental health start and end? Who should be respon sible for overseeing their interventions?

Chaudhary and Vasan (2020) believe that technology and social media companies are uniquely suited to be a psychia trists 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 COVID19 pandemic, researchers, medical professionals, and tech company work ers 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, author ship, and/or publication of this article.