Dr Ysabel Gerrard – written evidence (DAD0093)

This submission is based on research I have conducted since 2017 on: (1) the moderation of eating disorder communities across globally popular platforms like Instagram, Pinterest, Reddit and Tumblr, and (2) young people's experiences of anonymous 'secret-telling' social media apps. My expertise also derives from my role on the FACEBOOK company's Suicide and Self-Injury (SSI) Advisory Board, where I am a named contributor to Instagram's policies on, for example, advertisements for cosmetic surgery and weight loss products.

1. How do the content moderation policies of major technology platforms shape democratic discussion online?

Content moderation policies shape online democratic discussion in various ways, not limited to the ones discussed here, but one of the most important factors to consider is the influence of platforms' policies on the *cultures of acceptability* around the kinds of speech that are and are not acceptable.[i] This can vary wildly between platforms. For example, the founders of Facebook and 4chan have opposing views on the value of using 'real names' online: policies which directly feed into the nature of users' democratic discussions.[ii] A further influence on democratic discussion is the vagueness of rationales used to support the removal or retention of particular posts, and the inconsistency with which these rules are applied. For example, *ProPublica's* 2018 investigation revealed that Facebook permitted moderators to ignore posts 'display[ing] hate symbols for political messaging', like swastikas.[iii] While it is likely Facebook overhauled this particular rule after the investigation was released, it is equally likely that there remain other inconsistences and subjectivities in their moderator guidelines. Social media companies' Community Guidelines – the public-facing documents explaining a site's rules to its users – are also infamously vague.[iv] Related to this point, there are disproportionate levels of power that particular people – often public figures – have to bend the rules and remain active on platforms despite their clear violations of Terms of Service and/or Community Guidelines.

Content moderation policies are not the only factor shaping democratic discussion online: the algorithmic amplification of particular posts and viewpoints can undermine even the best intentions. As Siva Vaidhyanathan notes, one of Facebook (and other platforms') defining features is 'Its algorithmic design that amplifies content judged to attract attention and interaction (clicks, shares, likes, comments) [and] favors extremism and powerful emotions over rational and measured expression'.[v] Social media platforms are designed to show users

what they likely want to see – their fundamental, perhaps defining feature – but this is a real concern for democratic debate and, in the case of my research, users' mental health.[vi]

Although this might seem unrelated, I also want to note that the definition of 'major technology platforms' can be fairly unstable because some platforms with small workforces can become extremely popular and therefore 'major' in a stunningly short space of time. Saudi Arabian secret-telling app Sarahah, for example, started out as an app for workers to give anonymous feedback about their employer, but young people across the world reappropriated the app and used it for online harms like bullying.[vii] Small content moderation teams behind apps like Sarahah often cannot infrastructurally deal with the waves of abuse that flow through the platforms, sometimes leading to their removal from app stores. My point here is that the 'major' technology companies have unequal resources and sometimes radically different rules, so the terrain for online democratic discussion is changing all the time and it is important to be aware of this.

2. Is a lack of transparency a problem in online moderation? If it is, how could this be improved?

A lack of transparency is absolutely a problem in content moderation and there are four different aspects of the process that are arguably the least transparent and therefore need greater oversight and intervention:

a) *The rules:* we need to know far more about who influences and – crucially – agrees to implement social media platforms' rules. In particular, I would like to know more about the top-level decision-making that goes on *after* companies like Facebook consult experts for guidance on their policies, and I would also like to know more about which platforms are and are not working with independent experts to co-create policies;
b) *Labour conditions:* despite extensive research by scholars like Sarah T. Roberts, we know too little about the conditions under which human content moderators work at different companies, and how they experience their mental health both during and after their roles. It is also important for us to know how long each type of content moderator (in-house, outsourced, and so on) has to examine each category of post. For example, are moderators given longer to deal with particular kinds of content, like suicide or self-harm?;
c) *Moderator guidelines: The Guardian* leaked Facebook's internal guidelines for human content moderators in 2017, but in my view these documents should be far more transparent across all companies.[viii] These documents

should also be co-produced by external independent experts (academics, charity workers, practitioners, government representatives, activists, and so on) to ensure greater accuracy;

d) *In-platforms restrictions:* we need to know more about why particular kinds of content (for example, searches for hashtags) are restricted in particular ways, and why this changes so frequently.[ix] I would also like to know more about the criteria for 'shadowbanning' – the process of partially blocking a person from a platform, or restricting their content from view – and would like social media companies to be far more transparent about the fact that they do this and the reasons why they do it to particular people. [x]

3. What role could civil society play in improving content moderation? Would the social media companies not actually benefit from greater transparency?

One of the many roles civil society could play – and which would help researchers to produce more knowledge about content moderation – is to raise funds to launch faster-moving research grants to match the speed with which social media and digital technologies grow. As an example, I am currently applying for a grant worth between £100,000-£300,000 and the application process takes at least a year. I will then wait around six months to hear the outcome and have a fairly slim chance of actually receiving the grant. It is therefore difficult to expect UK-based researchers to produce knowledge about, and contribute to the regulation of, social media companies without the means to do so.

Social media companies would absolutely benefit from greater transparency. Being more transparent about different aspects of the content moderation process would generate public discussion and feedback to ultimately *improve* those systems. Greater transparency would also provide policy researchers with more information on which to base their research. Given increasing restrictions to social media platforms' Application Programming Interfaces (APIs), researching platforms can be a frustrating experience for those whose job it is to produce knowledge for and about society.

4. What should best practices in moderation look like?[xi] Do moderators have enough time and contextual information to make reliable decisions? Do the labour conditions that moderators face impact the decisions that are made? Can it be right that

'moderators' are at times required to sign NDA's, the effect of which is to encourage suspicion of malpractice?

Best practices for content moderation ultimately depend on the platform in question, its core userbase, and its dominant social 'ills'. But one of the moves all platforms should arguably aim to make is to moderate at a *local, not global* level. While I understand the push to globalise policies, 'both to increase transparency for users and to operationalize their enforcement for employees', this is not always the best approach.[xii] I sit on Facebook's Suicide and Self-Injury (SSI) Advisory Board and witness the huge tensions and trade-offs of trying to globalise something as complex and individualised as mental health. For example, should social media companies contact law enforcement when they receive credible suicide threats if they know the user is based in a country where suicide is criminalised? Globally dominant platforms like Facebook, YouTube and Twitter are dealing with huge tensions between consistency and locality.

Further best practices include eliminating for-profit rule-making (rules implemented predominantly to please advertisers); implementing oversight boards at all social media companies for major violations that require debate and publishing the results to the public; and crowdsourcing policy contributions (to an extent), as platforms with the populations of entire countries should surely allow democratic participation.

Even if they had all the time and data in the world, it is very difficult for a content moderator to ever have *full* contextual information about a post. The global spread of moderator labour means contextual cues are often missed and moderators' own subjectivities and biases are bound to leak into their decision-making. As an example, I research eating disorder communities and will admit that even I struggle to decide how I would moderate particular posts, and few people have the expertise on *both* content moderation and eating disorders to make an accurate decision. It is therefore highly doubtful that moderators will always make the right call. That said, increasing the amount of time moderators have, coupled with developing specialist teams of moderators who are knowledgeable in specific areas and assigned to only those posts might help to increase accuracy.

Researchers have compiled enough evidence to argue that the labour conditions for content moderators are often abhorrent, and the absolute priority should be that moderators receive health benefits and aftercare, regardless of where they are based and the type of role the occupy (e.g., in-house or outsourced).[xiii] I can understand asking moderators to sign Non-Disclosure Agreements (NDAs) to protect the identities of individual *users* whose posts they see, but moderators should not be barred from speaking about their working conditions: the hours

they work, how long they are permitted to spend on each post, rough descriptions of the kinds of post they see, and so on.

5. If Government could do one thing to regulate content moderation what should that be?

Require all globally dominant social media companies to recruit teams of independent experts to work with them to develop their policies. There is absolutely no way this can be done in-house at *any* social media company and teams of non-experts should not be left unattended to write the rules of platforms the size of entire countries.

---

[i] See for example Phillips, W. (2016). *This is why we can't have nice things: mapping the relationship between online trolling and mainstream culture*. Cambridge, MA: MIT Press.
[ii] van der Nagel, E. and Frith. J. (2015). Anonymity, pseudonymity, and the agency of online identity: examining the social practices of r/Gonewild. *First Monday*. *20*(3). Available at: https://firstmonday.org/ojs/index.php/fm/article/view/5615.
[iii] Angwin, J. (2018, June 28). Facebook's secret censorship riles protect white men from hate speech but not Black children. *ProPublica*. Available at: https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.
[iv] Gillespie, T. (2018). *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.
[v] Vaidhyanathan, S. (2019, October 18). Mark Zuckerberg doesn't understand free speech in the 21st century. *The Guardian.* Available at: https://www.theguardian.com/commentisfree/2019/oct/18/mark-zuckerberg-free-speech-21st-century.
[vi] Gerrard, Y. (2018). Beyond the hashtag: circumventing content moderation on social media. *New Media and Society*. *20*(12): 4492-4511.
[vii] Lunden, I. (2019, Jan 31). After bans from Apple and Google, Sarahah debuts Enoff, an iOS app for anonymous feedback at work. *Tech Crunch.* Available at: https://techcrunch.com/2019/01/31/after-bans-from-apple-and-google-sarahah-debuts-enoff-for-anonymous-feedback-at-work/.
[viii] Hopkins, N. (2017, May 21). Revealed: Facebook's internal rulebook on sex, terrorism and violence. *The Guardian*. Available at: https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence.
[ix] Suzor, N. (2016, Sep 18). How does Instagram censor hashtags? *Medium: Digital Social Contract*. Available at: https://digitalsocialcontract.net/how-does-instagram-censor-hashtags-c7f38872d1fd.
[x] Myers West, S. (2018). Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media and Society*. *20*(11): 4366-4383.
[xi] This is a question Jillian C. York – Director of International Freedom of Expression at the Electronic Frontier Foundation (EFF) – asks frequently, and would be a great contributor to this discussion if she isn't already involved.
[xii] Caplan, R. (2018). Content or context moderation? Artisanal, community-reliant, and

industrial approaches. *Data and Society Research Institute, New York*. Available at: https://datasociety.net/output/content-or-context-moderation/. p.1.

[xiii] See Roberts, S.T. (2019). *Behind the screen: content moderation in the shadows of social media*. New Haven: Yale University Press.