Associate Professor Heather Ford[1]—written evidence (FEO0116)

Submission to the House of Lords Communications and Digital Committee inquiry into Freedom of Expression Online

How can content moderation systems be improved?
Are users of online platforms sufficiently able to appeal moderation decisions with which they disagree? What role should regulators play?

Today, a handful of platforms have become keepers of the public discourse, not only for online communities but for whole nations. Platforms adjudicate the truth by forbidding some types of speech, highlighting some and ignoring others. Moderation decisions are applied undemocratically: without transparency, consistency or justification. As a result, inaccurate claims fester and grow across platforms, fuelling polarisation, hate and distrust of public institutions.

Two solutions are generally proffered in response to the growing threat of misinformation and disinformation. The first is education. Users (or more appropriately, citizen users) need to be aware of how to spot fictitious claims. The second is improved moderation. Platforms need to become better at accurately targeting harmful content while enabling content that furthers public debate and enables artistic and intellectual freedom to flourish.

The problem is that these solutions reinforce platforms as the arbiters of truth, far removed from public deliberation. According to these views, platforms make the decisions (perhaps governments step in to provide an independent appeal process) but citizens are left on the other side of the debate, having to educate themselves about how platforms operate without the power to do anything to resolve inaccuracies and lacunae. If citizens notice misinformation on a platform, they can do little to affect its removal other than to participate in obscure reporting mechanisms with no feedback about what happened to their request. If their own content is removed for going against platform rules, they aren't provided adequate justification or the opportunity for appeal.

Instead of seeing users only as recipients of education, **we need to recognise citizen users as important agents in the moderation process**. Improving moderation is not about enhancing platforms' ability to accurately classify the quality of information. Moderation is not an end in itself – it needs to be seen as a vehicle for greater accountability. Public accountability offers the opportunity for platforms to think more creatively about how to develop moderation practices in the public interest.

Two key strategies are emerging by innovative organisations, companies and platforms involved in the development of the Internet as a public sphere. The first is enabling greater verifiability of claims made across digital platforms by user citizens. The second is enabling ensuring adequate justification for decisions made by platforms. Both of these strategies place the emphasis on the shared

[1] Acting on an individual basis. Affiliation: School of Communication, University of Technology Sydney.

problem of information quality and dampening the power of platforms to adjudicate the truth for billions of citizens around the world.

Verifiability is a quality of an information that points to its ability to be verified, confirmed or substantiated. Claims that are linked to a source that readers can check to confirm that they were authored by that source are more verifiable than those that are not. Merely existing in an external source, however, is not enough for readers to know that the claim is accurate. The claim may have been accurately cited but the original claim might be erroneous. Without being able to do the original research themselves, readers must be provided with information on which to judge whether the source is reliable or not.

Verifiability sets up a productive dynamic between readers and authors. On Wikipedia, verifiability is a core content policy and crucial to its relative resilience against misinformation. It is defined as the ability for "readers [to be] able to check that any of the information within Wikipedia articles is not just made up." For editors, verifiability means that "all material must be attributable to reliable, published sources."[2]

Verifiability is under threat as the Web becomes increasingly automated. Wikipedia researchers have found that factual data from Wikipedia that is surfacing as answers to user queries in digital assistants and smart search, notably by Google, does not cite Wikipedia as the source of that data[3]. Search engines and digital assistants are becoming authoritarian gatekeepers of factual knowledge as their answers are adjudicated by algorithms that often remove the source and provide no mechanism for people to appeal decisions made. The only way that changes seem to be made is via articles written in powerful media companies but then predominantly in the US. And even then, in some cases journalists have been told to try to get others to help them train the algorithm to remove erroneous content[4]. Platforms seem to have little control over the truths their algorithms are discharging.

At the least, verifiability should be ensured by citizen users' ability to check the source of information being proffered. But verifiability can go much further and evade some of the problems in defining universal rules for what constitutes a "reliable source". **Verifiability should ultimately be about the ability for citizen users to make determinations (individually and collectively) about the trustworthiness of information.**

Some work has started on verifying the authorship of images, for example, by surfacing metadata about its provenance. The Content Authenticity Initiative[5] (CAI), for example, is a partnership between Adobe, publishers such as the BBC and the New York Times, and platforms like Twitter, that enables citizen users to

[2] Wikipedia, s.v. "Wikipedia: Verifiability," last modified January 13, 2010, https://en.wikipedia.org/wiki/Wikipedia:Verifiability.

[3] McMahon, C., Johnson, I., & Hecht, B. (2017). The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14883

[4] https://www.nytimes.com/2017/12/16/business/google-thinks-im-dead.html

[5] https://contentauthenticity.org/

click through images they see in order to find out how those images were edited and the context of its source. In time, the CAI believes that people will be trained to look for data that helps them verify a source whenever they see startling information online, rather than to merely accept it at face value.

Tarleton Gillespie, in his book, "Custodians of the Internet[6]" (2018) suggests that "(p)latforms should make a radical commitment to turning the data they already have back to me in a legible and actionable form, everything they could tell me contextually about why a post is there and how I should assess it." (p199) Examples include flagging to users when their posts are getting a lot of responses from possible troll accounts (with no profile image and few posts) or labelling heavily flagged content or putting it behind a clickthrough warning. Gillespie writes that these could be taken even further, to what he calls "collective lenses". Users could categorise videos on YouTube as "sexual, violent, spammy, false, or obscene" and these tags would produce aggregate data by which users could filter their viewing experience (p199-200).

In my upcoming book about Wikipedia, I talk about platforms surfacing data that indicates the stability or instability of factual claims. Pandemics, protests, natural disasters and armed conflict are unexpected catalysts followed by a steep spike in information seeking while very little reliable information is available and consensus has not yet been developed. This rift between the demand and supply of reliable information has created the perfect storm for misinformation[7]. But rather than labelling facts as either true or false on the context of catalytic events, platforms can flag claims as stable or unstable. Instability is a quality of facts and their relation to breaking news events. Platforms have access to significant amounts of data that can signal instability: edit wars on Wikipedia and traffic spikes according to hashtags, search queries, keywords. Rather than marking claims as either true or false, platforms can educate citizen users about the instability of claims (what Professor Noortje Marres from the University of Warwick calls "experimental facts[8]") that are still subject to social contestation.

A handful of publishers and platforms are experimenting with flagging posts according to their stability. Wikipedia uses human moderators to flag articles subject to breaking news, warning users that information is subject to rapid change and alteration. But automated data indicating peaks in reading and editing would be a more accurate tool for indicating instability, and one that isn't subject to editorial politics. Platforms like Instagram automatically append tags to posts about vaccines with information from government health services[9]. Publishers like the Guardian flag articles that are more than a year old so that users recognise that the information is possibly out of date[10]. Factual claims in question and answer systems such as Google Knowledge Graph or Amazon's Alexa could indicate the instability of the answers that they select, and urge citizen users to find more information in reliable, institutional resources.

[6] Gillespie, Tarleton. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[7] https://datasociety.net/library/data-voids/

[8] Marres, Noortje. "Why we can't have our facts back." *Engaging Science, Technology, and Society* 4 (2018): 423-443.

[9] https://about.instagram.com/blog/announcements/continuing-to-keep-people-safe-and-informed-about-covid-19

[10] https://www.bbc.co.uk/news/technology-47799878?intlink_from_url=&

Platforms should be mandated by government to enable meaningful verifiability of content they host, giving users more control to make their own determinations. Verifiability is a critical principle for balancing the power of platforms to adjudicate the truth, but this doesn't solve the problem of platform accountability. Even if platforms surface information to help users better adjudicate content, they still make moderation decisions to block, highlight, filter or frame. The problem is not only that they make decisions that affect the health of nations, but that they do so obscurely, inconsistently and without having to justify their decisions adequately to users.

Critical to the principle of accountability is the right to justification, as I've argued[11] with my colleague, Dr Giles Moss from the University of Leeds. Decisions made by platforms need to be adequately justified to those affected by those decisions. The problem is that decisions made by platforms to moderate are obscure and not adequately justified. Facebook, for example, bans users without explaining the reasons[12]. Twitter flags tweets as "misleading" without explanation[13]. In addition to enhancing the verifiability of content, platforms must also adequately explain their decisions beyond merely flagging content or notifying users that they have been banned.

Platforms will have to experiment with how to provide adequate justifications at scale. They will have to uncover the principles underlying the algorithms that automatically make many of these decisions. And they will have to reveal that information in meaningful ways. **Governments can help by providing principles for platform justifications and enabling the independent review of a selection of decisions** – not those who have been successful in drawing popular support or media attention[14], but randomly selected decisions.

Platforms moderate and will continue to moderate. We can't prevent them making those decisions but we can improve the accountability by which they make those decisions.

14 May 2021

[11] https://www.elgaronline.com/view/edcoll/9781789903089/9781789903089.00019.xml

[12] See https://www.facebook.com/help/381336705253343

[13] E.g. see https://hotair.com/allahpundit/2021/04/18/why-did-twitter-flag-my-pro-vaccine-tweet-as-misinformation-n384095

[14] https://oversightboard.com/