

Written Submission of Reddit, Inc. In Response to the Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023 Exposure Draft

Executive Summary

- Reddit has serious concerns about the draft Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023 (the Bill). We are concerned about the disproportionate effects of the Bill on small-to-medium sized platforms, given the significant burdens on platforms introduced by the Bill, under threat of heavy penalties if platforms do not comply. In addition, in our experience, a code development process led by industry tends to be weighted in the interests of larger platforms, at the expense of smaller ones or those with differentiated business models.
- Any approach to addressing misinformation and disinformation online should derive from a framework that protects users' freedom of expression, in recognition of the often highly subjective manner in which these terms are used and applied. Given the risk of abuse by those seeking to remove content they simply disagree with, including content of great public interest, any proposal should be cautious not to create incentives for platforms to err on the side of removing content or undermine the value of content governance models alternative to top-down, centralised systems.
- Due to the vague and open-ended powers the Bill grants to the ACMA, and the lack of guardrails on the exercise of those powers or protections to platforms, the Bill will likely have major ramifications on open discourse and debate on the Internet. We would strongly urge a serious reconsideration of the Bill, its workability, and whether it will achieve its original policy intent of preventing the spread of harmful misinformation and disinformation without unreasonably choking legitimate discourse. Collaboration with civil society and industry and in particular small- and medium-sized platforms will be essential to ensuring the success of any effort to address misinformation and disinformation, without impinging on users' freedom of expression.

I. Introduction

Reddit is pleased to provide its feedback on the exposure draft of the Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023 (the Bill). Reddit is a medium-sized, privately held company with approximately 2,000 employees worldwide. Headquartered in San Francisco, the company opened an Australian office in Sydney in July 2021. Reddit's platform serves more than 54 million daily active unique users worldwide, and Australia forms our fifth largest user base.

Reddit's mission is to bring community, belonging, and empowerment to everyone in the world. Accordingly, and unlike many other digital platform services, Reddit's communities (or "subreddits") form the basic unit of engagement on our platform, rather than the individual. Consistent with our core value of user empowerment, Reddit adopts a multi-layered, democratic approach to content moderation, where every user plays a vital role in the governance of our platform.

Given the community nature of Reddit as compared to other platforms, we wanted to use this opportunity to explain Reddit's approach to content moderation and how our system of community governance is uniquely suited to quelling the proliferation of harmful content online, before providing feedback on some key themes on the Bill.

II. Reddit's Community Moderation Approach is Highly Effective at Organically Suppressing the Proliferation of False and Harmful Information Online

Reddit provides a forum-like platform for people to create self-governing communities of shared interests, known as "subreddits". Every user (known as a "Redditor") can create and partake in discussions in their chosen subreddit, so long as they abide by Reddit's site-wide policies¹ as well as the rules of the individual subreddit. These subreddits can range from general interest (for example, for users located in the same geographic region, such as r/Australia² and r/Canberra³) to hobbies or interests (like r/matildas⁴ and r/ultralightAus⁵) to communities that share common traits or life experiences, such as r/ainbowOz.⁶ There are over 100,000 active subreddits on Reddit, all of which are maintained by our volunteer community moderators (or "mods"). Mods are not employees of Reddit, but impassioned users keen to play a role in creating and maintaining communities that share their interests (which includes setting the rules of their subreddits).

Reddit's community structure is uniquely effective at suppressing the proliferation of false and harmful content on the platform for a number of reasons. First, in enforcing community rules and norms, moderators are quick to identify content that does not fit within the culture of their community or abide by their community rules, including content that they believe to be inauthentic or harmful to their communities.

Second, due to the tailored nature of individual subreddits and the fact that each subreddit has particular rules, content that is relevant and of interest to one community is often not relevant or of interest to another, and therefore, not often shared across subreddits, limiting its potential virality.

Third, individual users have various options for voicing their views on harmful content and to organically suppress its proliferation. This includes both commenting, to explicitly voice disagreement or cite evidence, and downvoting, which reduces the visibility of the content in the community. In addition, users have a public reputation score known as "Karma", which is based on the number of upvotes and downvotes their posts and comments receive, and serves as both a reflection of the account's age, as well as a shorthand indicator of the potential quality of the account's contributions to the broader community. Reddit is, at its core, a community of forums for discussing different ideas, and we find that users are often quick to naturally debate, challenge, and downvote misleading or inaccurate content across the platform, effectively harnessing the wisdom of crowds.

Indeed, independent research has shown that Reddit's communities effectively identify and suppress inauthentic content at scale, and differentiate high quality news sources from low quality news

¹ See, for example, Reddit's Content Policy (https://www.redditinc.com/policies/content-policy)

² https://www.reddit.com/r/Australia

³ https://www.reddit.com/r/Canberra

⁴ https://www.reddit.com/r/matildas

⁵ https://www.reddit.com/r/ultralightAus

⁶ https://www.reddit.com/r/ainbowOz

sources.⁷ In a paper examining the dissemination of false stories on Reddit compared to other platforms, Bond and Garret concluded that on Reddit, not only do users tend to publicly fact-check each other's claims, but that information fact-checked as true received far more engagement than information fact-checked as false.⁸ This highlights the strength and effectiveness of allowing communities the autonomy to engage in contextual debate and meaningful discussions, rather than simply forcing platforms to move towards a top-down, one-size-fits-all approach to content governance. To do so would effectively disempower our users and reduce the important role that they play in our ecosystem.

Backstopping this system of community moderation are our internal safety teams, which enforce our sitewide Content Policy and undertake proactive work, in addition to responding to user reports, to keep our platform and users safe. Their work is particularly focused on more complex issues that have a sitewide nature or require specialised data or tools that aren't appropriate for volunteers, such as complex investigations involving malicious foreign state actors. Further detail about the work of our safety team and our content moderation system, including specific policies pertinent to the topic of mis- and disinformation, can be found at **Appendix A**.

Through this system of shared responsibilities, users, volunteer community moderators, and Reddit employees each play a vital role in governance on the platform, and in suppressing the proliferation of harmful content. We have found that this approach not only empowers users, but is more scalable and effective than a top-down, centralised approach to content moderation. It also provides scope for individuals to carefully consider the multitude of nuances and subtleties that come with determining whether content is appropriate for their communities, or whether it is harmful and should be removed.

A. Reddit's content moderation system in practice – Russian campaign case study

We thought it would be useful to provide a case study of how this system works in practice. In December 2019, we <u>posted</u>¹⁰ in <u>r/redditsecurity</u>¹¹ about a report released by Graphika, an organisation focused on social network analysis. That report studied the breadth of suspected Russian-connected Secondary Infektion disinformation campaigns spanning "six years, seven languages, and more than 300 platforms and web forums," to include Reddit, with 52 Reddit accounts linked to these disinformation campaigns. However, upon investigation, it was discovered that *all* of these accounts already either had their content removed by community moderators and/or were caught as part of Reddit's normal spam mitigation efforts.

Reddit's community design and multi-tiered moderation system makes it difficult for such content to be spread broadly. Many of the posts associated with these accounts were removed because they breached subreddit rules, or were downvoted into oblivion by users. Reddit's anti-spam and content manipulation safeguards, implemented by moderators in their communities and at scale by our internal safety teams, also played a key role in removing this content before it had a chance to

⁷ Cameron Martel, Jennifer Allen, Gordon Pennycook, David G. Rand, <u>Crowds Can Effectively Identify Misinformation at Scale</u>, 24 October 2022, PsyArXiv.

⁸ Robert M. Bond, and R. Kelly Garrett, <u>Engagement with fact-checked posts on Reddit</u>, 2023 *PNAS Nexus*, Vol 2 Issue 3.

⁹ Further detail about the work of our safety team and our content moderation system in general can be found in Appendix A.

¹⁰ https://www.reddit.com/r/redditsecurity/comments/ha885d/secondary_infektion_the_big_picture

¹¹ http://www.reddit.com/r/redditsecurity

proliferate. Finally, most of the accounts at issues had a zero or negligible Karma scores, which indicated their lack of traction or influence amongst users, resulting in their organic suppression.

III. Reddit's Feedback on Selected Themes of the Bill

Reddit thanks the Department for proactively engaging with industry and for the opportunity to comment on the exposure draft of the Bill. Reddit has serious concerns regarding the Bill's contents, how it will operate in practice and its impact on users' freedom of expression and open discourse online. We are also concerned about how the Bill conceptualises content moderation solely from the perspective of centralised platforms with top-down, one-size-fits-all systems of governance. The unintended consequence of the Bill would be that platforms with differentiated content moderation models are required to move towards these systems to ensure they are not breaking the law, enshrining into law the practices of the largest platforms, disempowering users, and disadvantaging smaller or alternatively structured platforms.

We want to stress that we are not disparaging the policy intent behind the Bill; misinformation and disinformation are a cause for concern, which is why we construct systems to combat such behaviour and, as demonstrated in our most recent Transparency Report, dedicate resources to detect and mitigate it.¹² Rather, we take issue with the starting position of the Bill, which is that platforms can, and should, under penalty from the government, determine the truthfulness of content created by users on their platforms, using the definitions proposed in the Bill and without sufficient oversight over the exercise of powers granted by the Bill. Platforms should not, through this legislation, be created as a cat's paw to do what the government itself is legally and ethically barred from doing.

To that end, we would strongly urge the Government to reconsider the Bill and its fundamental starting position. As drafted, the Bill will not achieve its original intent of preventing the harmful spread of misinformation and disinformation and instead, will require platforms to adhere to unworkable requirements under threat of heavy penalties, and divert resources away from addressing real harm. Collaboration with civil society and industry – and in particular small- and medium-sized platforms – will be essential to ensuring the success of any effort to address misinformation and disinformation.

A) The Bill Disadvantages Small- to Medium-Sized and Differentiated Platforms

We urge the Government to reconsider the effect of the Bill on small- to medium-sized platforms and platforms with differentiated business models, which may be unduly burdened by the obligations presented under the Bill. Digital platforms are not a homogenous group and there are digital platform service providers of varying sizes and resourcing capabilities, and with a broad range of product offerings and business models. Through the definitions under clause 2 of the Bill, many digital platform service providers will likely be captured, and the obligations required of a large, multinational corporation will be the same as a small platform with minimal presence in Australia.

As described above, we believe that a multi-layered, democratic approach to content moderation not only empowers users but is more scalable and effective than centralised, top-down systems. Our community moderators are fundamental to building the norms and nuances of each of the thousands of communities on Reddit, and our voting system allows users to have a say in what content is

¹² See "content manipulation, Chart 5", Reddit Transparency Report 2022, available from https://www.redditinc.com/policies/2022-transparency-report.

appropriate for their communities. But by requiring platforms to identify content as "misinformation", "disinformation" or "false, misleading or deceptive" as defined in the Bill, and take actions related to those terms, the Bill would effectively require community-structured platforms like Reddit to ignore and undermine the expertise of moderators in assessing the nuances and context that is so essential to online discourse.

The value of protecting community-focused structures cannot be understated. By way of example, the moderators of <u>r/transgenderau</u>, ¹³ a subreddit that "focuses on transgender rights and support here within Australia and trying to provide a space where trans folks can come and get help with accessing medical care", explain why Reddit's user-first content moderation model suits their needs:

We are a small and often misunderstood community and what is a transphobic dog whistle often gets missed by big moderation groups like that of Facebook. We need to be able to moderate ourselves to ensure that action is taken swiftly and appropriately. We need to ensure that our community is safe and we do get attacked from time to time, even in Australia.

In addition, we have reservations about the industry-led code development process. In our experience, such processes can be highly burdensome for small-to-medium sized platforms, which may not have the resources to dedicate exclusively to code-development or other industry processes. Moreover, because industry codes are often conceptualised with the business models of larger platforms in mind, the ability for larger companies to dedicate more time, money and effort to the code development process means that the outcomes of these processes are generally weighted more in their interests.

B) The Bill Threatens Open Discourse and Debate Online

At Reddit, we believe in empowering users and providing a space for authentic and meaningful discussion. Key to such discussions is the freedom for users to express their views, including unconventional or unpopular ones, so long as they adhere to sitewide and subreddit rules. The Bill threatens such free and open exchange of ideas. In particular, we are extremely concerned about the vague and open-ended nature of the proposed expansion of the ACMA's powers, as well as the fact that Bill would incentivise the removal of content, chilling important debates and silencing unpopular opinions.

i) Vague and Open-Ended Definitions Will Negatively Affect Speech Online

We are concerned that the definitions in the Bill are overly broad and vague, which could have potential ramifications for speech online, and how platforms moderate content. For example, both definitions of "misinformation" and "disinformation" refer to the provision of content that is "reasonably likely to cause or contribute to serious harm". "Harm", as defined in the Bill, encompasses a huge range of possibilities. For example, "harm to the Australian environment" could mean any number of things. Thus, the generality and lack of precision could encourage platforms to remove more content than they otherwise would, for fear of being penalised.

There have been important scholarly criticisms of using harm as the basis for regulating information online; for example, Dr Stephanie Alice Baker et al argue that "the concept of harm is not neutral and will vary from the standpoint of individuals and collectives". ¹⁴ Hinging the legislation on a poorly

_

¹³ https://www.reddit.com/r/transgenderau

¹⁴ Stephanie Alice Baker, Matthew Wade, Michael James Walsh, <u>The challenges of responding to</u>

defined and highly subjective threshold is likely to result in over-enforcement of policies and potential censorship of discussions, particularly in light of the serious penalties proposed by the Bill – this could even include discussions that may eventually be validated in the future. For example, early in the COVID-19 pandemic, some platforms considered it mis- or disinformation to suggest that the virus may have emerged from a laboratory. People discussing this possibility had their posts removed by those platforms, and in some cases their accounts banned, cutting them off from the conversation entirely just for expressing an opinion about a topic of global public interest. Now, the lab leak question is considered a serious hypothesis that is actively debated amongst experts. Premature suppression of debate on this topic had the unfortunate impact of politicising a matter of public health and entrenching people on all sides in their views, harming civic public discourse for years to come.

Following on this example, the Bill could have the effect of incentivising platforms to take down content even where the determination is highly subjective, or where a platform could not reasonably assess truth or falsity. As currently drafted, the Bill will empower the ACMA to require platforms to:

- arbitrate on the truthfulness of content on its platform and determine whether the provision of the content on its platform is reasonably likely to cause or contribute to serious harm (clause 7(1), under the misinformation limb), and
- determine whether the person disseminating, or causing the dissemination of, the content intends that the content deceive another person (clause 7(2), under the disinformation limb).

These rules will require a number of assessments on the part of a digital platform, which are often highly subjective in nature. The obligation to determine whether an actor *intended* to deceive another is troublesome, given the difficulty in deducing the intentions of others. As Jason Pielemeier of Global Network Initiative states, "*Determining a speaker's intent is notoriously difficult and can be doubly difficult in online contexts where nuance, jargon, and slang—not to mention the use of different languages—proliferate.*" As the prior example shows, misinformation and disinformation are notoriously difficult concepts to define and enforce against in a scaled and fair way, and we are concerned by the proposal to enact these terms into law.

ii) The Bill Has Insufficient Safeguards to Protect Platforms and Users

The Bill effectively empowers the ACMA to require platforms to perform assessments of possible misinformation and disinformation in any number of ways – through digital platform rules to require platforms to keep records (clause 14), requiring digital platform service providers to provide information and documents (clause 18), requests for industry to develop codes (clause 38), and/or through standards determined by the ACMA (clause 46). However, it is unclear what safeguards exist to prevent the ACMA from exercising their powers unnecessarily or in a way that creates undue burden for platforms and negatively impacts on users' experiences online.

For instance, clause 46 provides that the ACMA may determine a misinformation/disinformation standard if it refuses to register the requested industry code and determines instead that "it is

misinformation during a pandemic: content moderation and the limitations of the concept of harm, 2020 Media International Australia, Incorporating Culture & Policy, Vol 1 Issue 1, 103.

¹⁵ Cristiano Lima, "Facebook No Longer Treating 'Man-Made' Covid as a Crackpot Idea," Politico, 26 May 2021, available from https://www.politico.com/news/2021/05/26/facebook-ban-covid-man-made-491053.

Stolberg & Mueller, "Lab Leak of Not? How Politics Shaped the Battle over Covid's Origin," New York Times, 19
 March 2023, available from https://www.nytimes.com/2023/03/19/us/politics/covid-origins-lab-leak-politics.html.
 Jason Pielemeir, <u>Disentangling Disinformation: What Makes Regulating Disinformation So Difficult?</u>, 2020 ULR 917 (2020).

necessary or convenient" to draft the standard on its own. Given the severity of the penalties associated with breach of the standard that is set, it is concerning that the ACMA would only need to find that is it "convenient" to draft its own standard before platforms would be required to comply under threat of severe penalties. There is also no protection in this process for ensuring that the standard set does not create unreasonable expectations that harm free and open online discourse. There is no assurance that the ACMA would consult with stakeholders – and in particular small- to medium- sized platforms – in setting its own standard.

IV. Conclusion

A more nuanced approach to misinformation and disinformation must be taken than that presented in the Bill, to ensure that users' experience online and freedom of expression are protected. It is also critical to protecting the community moderation model, thereby ensuring that society more broadly can benefit from the wide-ranging benefits that this type of model offers, as well as ensuring that users have access to a diverse range of choices in terms of platform structures. To that end, greater consideration of the workability of the Bill and close consultation with civil society and industry are essential to achieving the Bill's initial objectives.

In closing, it may be instructive to recall the right to freedom of expression as enshrined in Article 19 of the International Covenant on Civil and Political Rights (ICCPR),¹⁸ to which Australia is a party. The Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda by the United Nations Special Rapporteur on Freedom of Opinion and Expression and others is unequivocal in highlighting that:

[T]he human right to impart information and ideas is not limited to "correct" statements... the right also protects information and ideas that may shock, offend and disturb, and... prohibitions on disinformation may violate international human rights standards. ¹⁹

_

¹⁸ Department of Foreign Affairs and Trade Canberra, International Covenant on Civil and Political Rights (New York, 16 December 1966), entry into force in Australia: 13 November 1980.

¹⁹ The United Nations Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information, Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda.

Appendix A

A. Upvotes, downvotes and karma

Content on Reddit is primarily ranked and curated by the votes of the users themselves, so that every user on Reddit participates in the content moderation process. Any Reddit user (known as a "Redditor") with a registered account can vote on each individual post and comment on the site. The content can be voted both "up" and "down" within the subreddit and the content consequently rises or falls in visibility on the page, based on these upvotes and downvotes.

The upvote/downvote system also provides an important signal about users themselves to other users, and to Reddit. Users who post or comment on Reddit have a public reputation score known as "karma", which is based on the number of upvotes and downvotes their posts and comments receive. These karma scores can serve as a shorthand indicator of the potential quality of the account's contributions to the broader community. It's also important to note that every Redditor's profile is public; the public nature of Reddit promotes transparency, and a user's karma score accordingly incentivises users to be constructive members of Reddit.

B. Volunteer community moderators

In addition to our users, our volunteer community moderators (or "mods") are essential to Reddit's content governance structure. Mods are not employees of Reddit, but impassioned users keen to play a role in creating and maintaining communities that share their interests.

Our users are empowered to create their own subreddits (as long as they abide by Reddit's sitewide policies²⁰), and mods can set and enforce subreddit-level rules for that community. In this sense, moderators play an important role in shaping their communities and providing a positive experience for users, by setting the rules and establishing the norms of their subreddit, actively moderating participation in accordance with those rules, and engaging with the community. Mods are expected to abide by Reddit's Mod Code of Conduct,²¹ which sets out Reddit's expectations of moderators and provides further information to help moderators develop subreddit rules and norms to create and nurture their communities. Users can also use the Code of Conduct to better understand standards of behaviour expected of mods, and flag to Reddit staff instances where moderators might be in violation of the Code.

Moderators are empowered to enforce their rules autonomously, such as by removing a post or comment, or banning individual users from posting within the community. They may do so manually, or using tools such as Automoderator, which is a simple automated content moderation tool that Reddit makes available to its moderators to configure as they see fit within their subreddit. These moderator actions happen without the involvement of Reddit, Inc., and form more than 89% of non-spam content moderation decisions on the platform, as noted in our biannual <u>Transparency</u> Report. ²²

We have a dedicated Community team at Reddit that proactively engages with moderators to ensure they are supported and equipped with the necessary resources to help them look after their

²⁰ See, for example, Reddit's Content Policy (https://www.redditinc.com/policies/content-policy)

²¹ https://www.redditinc.com/policies/moderator-code-of-conduct

²² https://www.redditinc.com/policies/transparency-report-2022/

communities. These include resources, such as the Mod Help Centre,²³ the Mod Education Centre,²⁴ which provides courses and certifications for mods, and Moderator Councils,²⁵ where employees engage moderators directly to hear their feedback on everything from safety tools to the Content Policy. The Community team also frequently engages with mods through a dedicated subreddit, r/ModSupport²⁶, which offers a space for open and transparent dialogue between Reddit and our moderators.

C. Reddit's sitewide policies and safety teams

Overarching these networks of subreddits and users is Reddit's Content Policy²⁷. The Content Policy is set by Reddit at the company level and applies across the entire site. In addition to forbidding unwelcome things such as hateful content, encouraging violence and sharing personal information or intimate imagery without consent, the Content Policy specifically prohibits users from posting inauthentic content and engaging in content manipulation.²⁸ Relevant to the topic of this submission, the Content Policy also includes an extensive rule against impersonation, which not only includes using a Reddit account to impersonate someone, but also encompasses things such as domains that mimic others, as well as deepfakes or other manipulated content presented to mislead, or falsely attributed to an individual or entity.²⁹ Health misinformation, namely falsifiable health information that encourages or poses a significant risk of physical harm, is also prohibited.³⁰

Reddit's internal safety and enforcement teams enforce the Content Policy in a variety of ways. We respond to user reports, which represent an important source of information for our teams. We are also constantly building and refining tools to proactively identify bad behaviour, including by monitoring data signals for evidence of sophisticated commercial spammers and building hashing tools to identify known instances of harmful and illegal content.

We publish statistics about actions taken by our safety teams in our biannual <u>Transparency Report</u>, ³¹ including content removed by Reddit employees, appeals to actions taken by Reddit and the consequent acceptance rate, and user account sanctions by Reddit. In addition, we recently established a <u>Transparency Centre</u>, ³² a one-stop shop for Reddit safety, security, and policy information. We also frequently provide updates to users on actions taken to ensure the safety and security of Reddit, including a quarterly safety and security report on the <u>r/redditsecurity</u> ³³ subreddit. By using a subreddit rather than a blog post, we are able to engage in an interactive dialogue with our users about issues they are interested in, rather than just talking at them.

²³ https://mods.reddithelp.com/hc/en-us

²⁴ https://modeducation.reddithelp.com

²⁵ https://mods.reddithelp.com/hc/en-us/articles/4415446939917-Reddit-Mod-Council-

²⁶ http://reddit.com/r/ModSupport

²⁷ https://www.redditinc.com/policies/content-policy

²⁸ See Rule 2 of Reddit's Content Policy: https://www.redditinc.com/policies/content-policy

²⁹ https://support.reddithelp.com/hc/en-us/articles/360043075032

³⁰ https://support.reddithelp.com/hc/en-us/articles/360043513151

³¹ https://www.redditinc.com/policies/transparency-report-2022

³² https://www.redditinc.com/transparency

³³ https://www.reddit.com/r/redditsecurity