June 14, 2024

The Online Harms Bill – Part 2 – Private Messaging

By: Sanjampreet Singh and Emily Laidlaw

Matter Commented On: Online Harms Bill C-63

This is the second in a series of posts about the Online Harms Bill C-63, proposed federal legislation the stated aims of which are to reduce harmful content online, hold social media platforms accountable, promote safety and child protection, empower users and victims, and increase transparency.

This post examines the social media services that would be regulated by the proposed *Online Harms Act* (*Act*) and potentially investigated by the Digital Safety Commission. More specifically, this post focuses on what is *excluded* from this Bill – private messaging – a "wicked problem" in online harms where one is damned if you do or damned if you don't include it. We propose a middle path.

The decision about who to regulate with this type of legislation is as much a practical question as it is a substantive one. The Bill proposes the creation of a new regulatory body in the form of a Digital Safety Commission (Commission), Digital Safety Office, and Ombudsperson. Creating a new administrative agency is a complex task, and there is only so much practically that a new body can oversee.

Background

As background, the Bill would regulate large social media services – including live streaming and adult services – based on the number of users, to be determined by regulations (at s 3, 2(2)). This aligns with aspects of the European Union's (EU) *Digital Services Act* (EC, *Digital Services Act*, [2022] OJ, L 277/1) (DSA), which impose the greatest obligations on very large online platforms (at s 5). Unlike the DSA, which regulates other platforms, hosting, and intermediary services, Bill C-63 leaves most of these services out of scope. Rather, the Bill enables the Governor in Council to designate a service as regulated where there is a "significant risk that harmful content would be accessible on that service" (at s 3(3)). Thus, a smaller social media platform that is a hotbed for the spread of harmful content, such as extremist or child sexual exploitation material, may be brought into the *Act*'s scope. This is a good solution to enable smaller social media services with outsized harmful impacts to be targeted without overwhelming both the Commission and smaller services with compliance obligations.

The definition of social media service is one whose primary purpose is to facilitate "online communication among users of the website or application by enabling them to access and share content" (at s 2). While not the focus here, it is worth noting the definition should be fine-tuned

when studied at Committee. The definition has been criticized as too broad, potentially sweeping in services that should not be included, such as Wikimedia (see comments of Florian Martin-Bariteau here).

Private Messaging

The issue examined in this post is the decision to exclude private messaging from the Bill. There are two aspects to this issue. First, should private messaging be excluded? Second, if the Bill passes as is, what are some of the practical challenges with navigating the public/private split?

Under s 5, if a social media service "does not enable a user to communicate content to the public", meaning the service does not enable communication to "a potentially unlimited number of users not determined by the user", the service is not social media and therefore not regulated by the *Act*. What this means is that a private messaging service (one that does not allow any communication to the world at large, such as Signal) would be excluded. It would not be social media for the purposes of the *Act*.

What about services that offer public and private spaces, which is the case for most messaging services, such as Telegram, SnapChat and WhatsApp, and most social media, such as Instagram, TikTok and Reddit? Under s 6, the duties in the *Act* do not apply to "any private messaging feature of the regulated service" (s 6(1)) (emphasis added). A private messaging feature is one that:

> 6(2)(a) enables a user to communicate content to a limited number of users determined by the user; and
> (b) does not enable a user to communicate content to a potentially unlimited number of users not determined by the user.

If a feature, like Instagram messaging, could be communicated to an unlimited number of people, even if the user chooses to only message two friends, then it would be in scope. As it stands, Instagram limits direct messaging groups to 250 people, and therefore this feature would be excluded from regulation. The result is that for a single social media service, some of its features would be regulated and some not. As will be discussed below, sometimes this division is obvious, but there are several grey areas that will make it more challenging than it may appear at first glance to identify what is in scope.

Should Private Messaging be Excluded?

Excluding private messaging from the scope of the Bill is defensible and practical for two primary reasons. Firstly, from a rights-based perspective, it protects the privacy and freedom of expression of all users. Private spaces online can be used for covert criminal activity, certainly, but they are equally important spaces to explore who we are, what we think, and our identities and feelings far from the public gaze. This is especially crucial for vulnerable groups, including children, women, the 2SLGBTQ+ community and people living in oppressive regimes and war-torn areas. The Office of the United Nations High Commissioner for Human Rights (OHCHR) reported that downloads of the secure messaging application Signal in Ukraine increased by 1000 percent within two months after the beginning of hostilities by Russia. Secure messaging is an important tool for

dissent as it allows journalists, lawyers, and civil society to operate without the fear of being surveilled.

Simply put, private messaging provides a platform to share the most intimate aspects of our lives with the people we trust, and state intrusion into our private sphere threatens a general system of surveillance.

Secondly, from a practical standpoint, the exclusion would preserve the security and integrity of private communications. Many messaging services are end-to-end encrypted, and solutions proposed to regulate private messaging tend to entail undermining encryption in scanning for harmful content. A myriad of cybersecurity risks follow as these technical measures create security vulnerabilities for all users. Undermining encryption also opens up avenues for the government to look into the communications of all users indiscriminately. This makes content moderation an exceptional challenge in this sphere. This is a developing area, and many technical measures to moderate content in encrypted communications are being explored. These include message franking (used by Facebook Messenger), client-side scanning (proposed by Apple but plans shelved later), traceability (proposed by Brazil and India) and homomorphic encryption. While the former three compromise encryption by creating security vulnerabilities, homomorphic encryption is different, but adoption at scale is not yet practically feasible given the high computing power required to implement it (one of the authors, Sanjampreet Singh, is examining encrypted private messaging in his LLM, so stay tuned).

If the Bill is passed as is, we understand and can support the decision to go narrow. However, this is a looming challenge and private messaging will eventually need to be tackled, whether now or down the road when the Commission is more established.

The Bill aims first and foremost to protect children. Most cases of child sexual exploitation or bullying now take place via private messaging. Cybertip reports that 79% of cases of online sextortion of youth take place on Instagram and Snapchat, most often starting on Instagram and then moving to Snapchat. The consequences for victims of sextortion are severe, and in worst case scenarios have led to youth taking their own lives. Most alarmingly, the time to intervene and help a child can be a matter of a few hours. In the case of a Manitoba teen, Daniel Lints, he was friended on Snapchat by purportedly a female teen, convinced to send an explicit image, and then blackmailed. The window between being blackmailed and the teen taking his life was three hours. This is not an uncommon situation or timeframe, made all the more concerning with the sharp increase in sextortion rates, up 150% in the last 6 months.

There is a middle path forward for Bill C-63. There are many safety measures that private messaging services can implement that do not entail scanning content and undermining encryption. For example, Snapchat has introduced in-app warnings to teen users when they receive a friend request from someone with no mutual connections. WhatsApp puts a cap on forwarding to limit viral messages and labels forwarded messages as *'forwarded'* and *'forwarded many times'*. WhatsApp also allows access to unencrypted metadata whose analysis can help detect harmful patterns. Other safety features can include:

- Appropriate and accessible complaints mechanisms, including clear avenues to report and flag content;
- Alerts when friend requests are suspicious, such as from a distant location and/or unrelated to friends;
- Access to a resource person or other resources to help in times of crisis;
- Removal of network expansion prompts (a functionality that recommends connecting with other users based on similarity of interests, location etc to expand networks) for children (proposed by the UK's Office of Communications (OFCOM));
- Behavioural analysis through available metadata;
- Disabling accounts.

While some private messaging services have voluntarily implemented safety measures, like all social media, there is a need for oversight, accountability, and transparency. This is a role for the Digital Safety Commission. In essence, private messaging services should have a digital safety plan. It would simply be a different type of plan given the stakes for privacy, freedom of expression, and cybersecurity, among other rights. Further, such an approach would align more closely with the EU, United Kingdom (UK), and Australia, which all include, in different ways, private messaging, albeit controversially.

This change should not be implemented by simply adding private messaging to the scope of the Bill. Some safety duties for private messaging features should be prescribed, albeit in a more limited manner. Currently, under Bill C-63, social media platforms have a duty to act responsibly, which entails mitigating the risk that users will be "exposed" to harmful content (s 55(2)). Regulated social media services also have a special duty to protect children, including designing their service safely (s 64) and a duty to take down child sexual victimization content and non-consensually shared intimate images once identified (s 67). There are other provisions in the Bill that are more specific, such as user empowerment tools to block and flag content, labelling harmful content, and researcher access, to name a few (ss 57-60, 73). These provisions do not translate seamlessly to private messaging without incentivizing undermining encryption. This is despite section 7, which states that there is no duty to proactively search for harmful content.

The middle path forward would be to require that, for private messaging services and features, operators must file a digital safety plan that demonstrates how they mitigate the risks of harm and protect children. The Bill could explicitly state that the duty (and remit of the Commission) would not extend to content communicated using private messaging features, nor mandate use of technology that would undermine privacy or security in private spaces. Instead, the safety duties outlined for private messaging services could include introduction of user empowerment tools like blocking and reporting other users, display of warnings to child-users before connecting with distant unknown users, and usage of metadata analysis to detect harmful behavioural patterns (for example, if an adult user constantly interacts with or tries to connect with child-users). These duties can be fulfilled without regulating the content shared in private spaces. A digital safety plan centred around user empowerment and metadata analysis could help mitigate some of the risks presented by private messaging applications while ruling out state intrusion. In this way, safety is built around private communications to empower user safety and privacy.

While we do not recommend wholesale adoption of the UK approach, which prompted threats from Signal and WhatsApp to pull out of the UK, some lessons can be learned from their attempt to differentiate between content communicated publicly and privately in terms of obligations (see s 136(6) and Schedule 4, para 13) and their commitment to safeguard encryption by not enforcing the "spy clause" until appropriate technical measures are available.

What's In and Out Under the Current Bill

If the Bill is passed as is, it is important to explain how this would work in practice. As discussed above, the definition of private messaging features under section 6 of the Bill means that some features of a regulated social media service would be exempted and some not. Before continuing with this section, it is important to remember that a 'feature' would be regarded as a 'private messaging feature' if it allows a user to determine and limit the recipients of their content and does not allow communication to a potentially unlimited number of users. This poses a few practical challenges, as many features have characteristics that constitute grey areas from a regulatory perspective.

For example, private servers (including community servers) on Discord are invite-only, however, the administrator may make the server discoverable by the public at large at any time and increase the membership limit by contacting Discord support. Discord primarily works through roles and permissions which allows administrators to determine recipients but also allows the potential to communicate content indiscriminately. Similarly, public channels on Zello have no membership limits. However, at any given time, no more than 10,000 users can be a part of an ongoing conversation on a channel. Some features, like private profiles on Instagram and protected accounts on X, allow users to readily switch between privately communicated and public-facing content. Some platforms, like OnlyFans, allow sharing content with paid subscribers which effectively puts a limit on recipients but also maintains the theoretical potential of sharing content to an unlimited audience.

The following table summarizes some of the features that would be regulated or excluded under the Bill. A more detailed table is on file with the authors.

Regulated Features	Public and Private Profiles on Facebook, Instagram, TikTok and X; Facebook pages, Instagram Broadcast Channels, Twitter Spaces and communities, WhatsApp Channels and Telegram Public Channels, Telegram Group Chats, SnapChat Public MyStories and spotlight, and Twitch Livestreams.
Excluded Features	One-on-one and group chats/calls on WhatsApp, Signal, Facebook Messenger, Instagram, TikTok, Discord, SnapChat and X; Telegram private chats, WhatsApp Status and Signal Stories, SnapChat MyStory for friends and private snaps, Emails, Twitch Whispers, Zello Walkie-Talkie, messaging and groups.

Grey Areas	Discord private servers, Telegram private channels, Zello Public Channels, OnlyFans mass messaging, Facebook Groups and Community Chats, LinkedIn posts

If the Bill passes without change, the first step will be for the Governor in Council to pass regulations delineating the criteria to qualify as a regulated social media service under the *Act*. The question of what features of these regulated services are private will quickly follow and be a matter for the Digital Safety Commission. Importantly, it will require regular monitoring as a change in design or services offered will impact what is regulated. There is no avoiding this type of iterative oversight as social media products and features are constantly changing. There is an opportunity to explore expanding the scope of regulated services to private messaging when this is studied at Committee. We do not argue that the content of private messaging should be targeted. Indeed, we would strongly argue against such an approach. However, there is a way to thread the needle of privacy and cybersecurity concerns while improving safety on private messaging by targeting the package that surrounds private messaging – complaints mechanisms, user empowerment, behavioural patterns, transparency – not the content itself.

This post may be cited as: Singh & Laidlaw, "The Online Harms Bill – Part 2 – Private Messaging" (14 June 2024), online: ABlawg, http://ablawg.ca/wp-content/uploads/2024/06/Blog_SSEL_Online_Harms_Bill2.pdf