Facebook response to Australian regulatory proposals for age-appropriate experiences online

SEPTEMBER 2021

FACEBOOK

Executive summary

Facebook welcomes the opportunity to respond to the consultation processes being run by the Australian eSafety Commissioner in relation to the *Age verification roadmap* and *Draft Restricted Access System declaration*.

Protecting our users - particularly young people - is of paramount importance to Facebook. We recognise the role that proportionate and risk-based age assurance regulation (in addition to other safety and privacy safeguards) can play in helping to ensure that young people have an age-appropriate experience online. This submission contains a range of constructive suggestions to assist eSafety in establishing an overall age assurance framework (part of which is the Restricted Access System) that could be an innovative model of regulation.

Within Facebook, we have a range of technologies, products and tools to ensure age-appropriate experiences for young people on our services. In particular, we have recently announced a package of new measures[1] for young people that involves:

- Defaulting new teen accounts on Instagram to private.
- Encouraging existing teen accounts on Instagram to be private.
- Making it more difficult for adults to find, follow and message unconnected teens.
- Limiting advertisers' ability to reach young people.

These measures are underpinned by a significant investment in age assurance technologies. We've developed technology that allows us to estimate people's ages - such as whether someone is below or above 18 - and this is currently in place in Australia. Combined with other measures, such as our age screen when a user signs up for an account, we have a comprehensive suite of privacy-protective tools for understanding a user's real age.

However, despite these measures, there remains a complex, industry-wide challenge: how to really know the age of someone on the internet. This is not as simple as it sounds. There is no foolproof way to stop people from misrepresenting their age. It also not possible to guarantee that a young person will never see sensitive or objectionable content online - just as the current methods for restricting young people from viewing harmful films are not perfect.

Regulation has a role to play in addressing the challenges of age-appropriate experiences online. Facebook has supported the Government's enhancement of online safety laws via the Online Safety Act. And we commend work on the UK's Age Appropriate Design Code as a good starting point for regulators in Australia considering new requirements for protecting the data of young people.

Australia is assembling a series of regulations that will comprise an overall framework of regulating for age-appropriate experiences online. The Restricted Access System declaration and any regulations resulting from the Age Verification roadmap will be

[1] Facebook, 'Giving young people a safer, more private experience on Instagram', *Facebook Newsroom*, 27 July 2021, https://about.instagram.com/blog/announcements/giving-young-people-a-safer-more-private-experience

core pieces, along with the Basic Online Safety Expectations and various industry regulatory codes that sit under the Online Safety Act. This submission provides comment on the age-related aspects of all of them, which we refer to as the regulatory framework for age assurance (noting that some components of the framework are not yet known or finalised). The various, interacting components should be viewed together holistically, to fully understand the cumulative impact of the proposals.

We believe the overall regulatory framework for age assurance could potentially be (if well designed) a novel and world-leading approach to age-appropriate content regulation - although it would require some amendments from the current course.

Many current pieces of content regulation around the world rely on "notice and takedown" schemes (including the cyber-abuse scheme, cyber-bullying scheme and image-based abuse scheme under the Online Safety Act) for harmful content. However, the age assurance regulatory framework has the foundation and makings of a "notice and age gating" scheme for objectionable or sensitive content. Under the legislation, eSafety is able to advise a digital platform of a piece of content that is inappropriate for young people and that platform can either take steps to age-gate it (ie. keep up the content but restrict the ability for users under 18 to view it) or remove it.

This model is an innovative approach to regulation. It is more protective of free expression than simple takedown schemes. It allows takedown schemes to be reserved for the most harmful material, while still protecting the safety of young people online. And it provides an innovative way of dealing with borderline content that holds significant public interest value but should not necessarily be viewed by young people (for example, graphic footage of a war zone).

However, we are concerned that the Restricted Access System declaration - or other forthcoming regulations under the online safety law - could deviate from the "notice and takedown" or "notice and age-gate" scheme contemplated by the legislation. The regulations could impose requirements in relation to age assurance and restricting access which are disproportionate or based on unrealistic expectations about the capabilities of technology.

Specifically, our submission raises the following concerns (and proposes constructive suggestions to address them):

- The definition of "class 2 material" (which informs the definition of content captured by the Restricted Access System) is very broad, ill-suited for user-generated content and subject to change. We recommend that the eSafety Commissioner prepare detailed guidance for platforms that host user-generated content to clarify how the definition should be applied for the purposes of age assurance.

- The framework for age assurance should not compel companies to proactively search and restrict access to this content. Given the ambiguity of the class 2 definition, we do not believe it would be fundamentally possible to build

technology that could effectively meet this requirement at the scale required for user-generated content.

- The measures that can be taken to ensure age-appropriate experiences in private messaging should be different to those taken for public social media. Private messaging is fundamentally different to social media, and is more akin to SMS or email. We very strongly contest any suggestion that digital platforms should be scanning the content of private messages to detect class 2 material. Our concern is even greater at the suggestion that we should scan for this material on end-to-end encrypted services, where the provider cannot see the content of private messages. While the Australian community does expect technology platforms to take measures to prevent the sharing of illegal and seriously harmful content, we do not believe Australians would expect or want Facebook to monitor their private messages for nudity or swearing.

- The regulatory framework should recognise that technologies and processes for age-appropriate experiences online are still evolving and are not foolproof. Although there is significant progress and innovation in this space, the technologies and processes for age assurance and restricting access are imperfect.

- The regulatory framework should not prescribe particular technologies to be used. We recommend instead that regulation should allow companies to evolve and to deploy measures that are right for their services. It could provide a suggested but non-exclusive list of technologies that companies may deploy. A non-exclusive list would accommodate players from across industry (including phone operators, platforms, and apps) working together towards comprehensive outcomes, rather than interpreting age assurance as a singular responsibility of each company. In particular, the Restricted Access System declaration should allow for flexibility in the technology and measures that companies use for age assurance or restricting access.

- The timeframe for implementing the Restricted Access System declaration is very short. If the Restricted Access System declaration takes effect in January 2022 (as contemplated in the Online Safety Act), the eSafety Commissioner could issue guidance to providers that it will not be enforced until six months after commencement. Much of the impact of the Restricted Access System will depend on *how* the Office of the eSafety Commissioner chooses to enforce it.

We would welcome the opportunity to work closely with eSafety and the Government on how to shape the Restricted Access System declaration (and any regulations stemming from the Age Verification roadmap) into a best practice model or template for other countries considering content regulation. Given the shared objective of ensuring young people are not exposed to inappropriate content, there is significant value to the community of government and industry collaborating on these issues.

Table of contents

EXECUTIVE SUMMARY 2

TABLE OF CONTENTS 5

FACEBOOK'S WORK IN PROVIDING AGE-APPROPRIATE EXPERIENCES 6

Protecting all users from harmful, objectionable or sensitive content 6

Age-appropriate experiences 7
- Consulting with young people 8
- Understanding a user's age 9
- Age-appropriate controls 10
- Underage users 13

A POSSIBLE GLOBAL BEST-PRACTICE MODEL 15

Principles 15

Designing an effective Restricted Access System 16
- The concept 16
- The details 17

KEY CONCERNS 17

Facebook's work in providing age-appropriate experiences

Protecting our users - particularly young people - is of paramount importance to Facebook. Many of the goals that the Australian Government's consultation is working towards are well-aligned with work that we already have underway.

There are a number of steps that we take that are relevant in this area, and this section outlines:

- Our efforts to protect *all* users of Facebook and Instagram from harmful, objectionable or sensitive content, including adult sexual material, nudity or graphic violence (the focus of the Restricted Access System declaration and the Age Verification roadmap).
- Our efforts to provide an age-appropriate experience and controls for teens (between 13 and 18) who use our services. This also includes our work to find and remove accounts belonging to people under the age of 13.

The Age Verification roadmap specifically consults on technology to limit young people's access to adult sexual material, and the Restricted Access System declaration will apply to a broader category of content that also includes nudity and graphic violence.

Facebook and Instagram are working hard to proactively offer products, tools and controls that give young people an age-appropriate experience. But we also believe that regulation has a role to play.

Facebook has supported the Government's enhancement of online safety laws via the Online Safety Act. And we commend work on the UK's Age Appropriate Design Code as a good starting point for regulators in Australia considering new requirements for protecting the data of young people.

There is significant discussion and debate around the world on the right role for regulation to play in encouraging age-appropriate online experiences. It's a complex question.

Protecting all users from harmful, objectionable or sensitive content

Facebook and Instagram already have strong policies about what material is and is not allowed on our services, which is outlined under our Community Standards.[2]

We do not allow nudity or sexual activity content on our platform, because we know some people in our community may be sensitive to this type of content (with some exceptions, such as for educational or medical reasons, or for the purposes of a protest). In the last quarter, we removed 32.8 million pieces of content for adult nudity or sexual content, 98.9 per cent of which we detected proactively before it was reported by a user.

[2] Facebook, *Community Standards*, https://www.facebook.com/communitystandards/

Content actioned for adult nudity or sexual content

We also do not allow content that glorifies or depicts violence (although we allow graphic content with some limitations to help people raise awareness about issues, such as human rights abuses or acts of terrorism). In the last quarter, we removed 30.1 million pieces of violent and graphic content, 99.6 per cent of which we detected proactively before it was reported by a user[3].

Age-appropriate experiences

We have a number of safety and privacy features that we make available for *all* users - as well as young people. These include:

- **Privacy settings.** Users on Facebook and Instagram can choose their audience settings and who sees their content.
- **Safety tools to block, unfriend, mute, or restrict unwanted interactions.** Facebook and Instagram have easy-to-use in-app tools that allow users to "unfriend", "unfollow", or block accounts they do not want to interact with.
- **Messaging privacy controls.** We provide a number of controls to protect users on Messenger and Instagram Direct.
- **In-app tools to report violating content.** Facebook and Instagram have in-app reporting tools so that users can report accounts, post, comments, and direct messages that they feel are inappropriate, and go against our Community Standards.
- **Extensive support for potentially vulnerable users.** Facebook also provides anonymous reporting tools for content such as self-injury posts, so that Facebook can connect the account reported to organisations that offer help, as well as anonymous reporting for live videos to report at-risk behaviour during a live broadcast, so the person reported receives a message offering help, support, and resources.

[3] Facebook, *Community Standards Enforcement Report Q2 2021 - Violent and Graphic Content*, https://transparency.fb.com/data/community-standards-enforcement/graphic-violence/facebook/

- **Anti-bullying tools and resources.** We have developed a dedicated Bullying Prevention Hub on Facebook and Anti-Bullying Centre on Instagram.
- **Ads Restrictions, and Advertising and Branded Content Policies.** Facebook applies numerous safeguards and restrictions around advertising.

In addition to these general service-wide privacy and safety features, we have a number of dedicated initiatives specific to young people, including:

- consultation directly with young people
- measures to understand a user's real age
- age-appropriate controls relating to objectionable or sensitive content, privacy settings, and the ability of advertisers to reach young people
- steps to discourage underage users from using our platform.

These are outlined in more detail below.

Consulting with young people

Creating an experience on Facebook and Instagram that's safe and private for young people, but also fun, comes with competing challenges.

In order to make sure we are striking the right balance and providing an appropriate and safe experience for young people, we engage closely with experts in this space - and with young people themselves.

To help us develop new products and features for young people, in 2017, we convened a group of experts in the fields of online safety, child development and children's media to share their expertise, research and guidance.[4] This group, known as the Youth Advisors, helps shape our work by providing feedback on the development of new products and policies for young people. We meet regularly with the group, which includes the Family Online Safety Institute, Digital Wellness Lab, MediaSmarts, the Cyberbullying Research Center and a number of new experts we recently added in privacy, youth development, psychology, parenting and youth media. Importantly, Facebook's global Youth Advisors Group includes PROJECT ROCKIT, an Australian-based NGO who also delivers the Facebook-funded Digital Ambassadors program, a youth-led, peer-based anti-bullying initiative. This year, the Digital Ambassador's program will reach more than 16,000 Year 7 students across Australia, equipping them with tools and training to create safe, supportive online communities[5].

We have undertaken Australia-specific consultation to inform our global work as well. In 2019, we partnered with a number of organisations - including the eSafety Commissioner - in July 2019 to deliver a Safety by Design Jam. The Safety by Design Jam was a workshop designed specifically for young people, which sought to gather insights and feedback from the people best placed to talk about youth safety online - young people themselves. It was the first of five age-appropriate Design Jams around the world, aiming to bring together policymakers, academics, safety and privacy

[4] Facebook, 'Hard questions: so your kids are online, but will they be alright?', *Facebook Newsroom*, 4 December 2017, https://about.fb.com/news/2017/12/hard-questions-kids-online/

[5] PROJECT ROCKIT, *Launching Digital Ambassadors*, https://www.projectrockit.com.au/digitalambassadors/

experts, and, of course, young people, to share new ideas and perspectives about how we can build age-appropriate experiences on our products that meet the needs and expectations of our young community.

We synthesised the outcomes of this consultation with young people (as well as consultation with the US Federal Trade Commission, UK Information Commissioner's Office, academics, civil rights groups, and industry) into a Youth Design Guide.[6] The Guide was co-developed with Trust Transparency and Control Labs and provides advice for product designers and developers. It suggests any products used by young people should follow the principles of (1) designing for different levels of maturity; (2) empowering young people with meaningful transparency and control; and (3) undertaking data education for young people.

We draw from all of the feedback provided (as well as other best practice resources) as part of our design of all products. For example, we have recently launched our internal Youth Knowledge Library, which includes best practices for product teams throughout the company. The Library is modelled on external guidance and frameworks developed by organisations like the UNCRC, OECD and children's rights groups - as well as our own consultation with third party experts, young people, parents and guardians. This resource ensures we have a company-wide understanding of how to apply general principles to specific experiences in our products, and will provide product teams with guidelines on how to design youth products in an age-appropriate way.

We have also assembled a cross-company research group across the Facebook family of apps to coordinate and share insights and to provide best practices and guardrails, to enable all product teams to effectively and appropriately engage with young people.

Understanding a user's age

It is a complex and industry-wide challenge to understand the age of users on the internet. Verifying someone's age is not as easy as it sounds and relying on identification documentation can raise privacy concerns and may not be truly effective to achieve the intended policy goal.

For this reason, we take a multi-layered approach to understanding a user's age on Facebook or Instagram.

We require users to provide their date of birth when they register new accounts, a tool called an age screen. Those who are underage (under 13) are not allowed to sign up. The age screen is age-neutral (ie. does not assume that someone is old enough to use our service), and we restrict people who repeatedly try to enter different birthdays into the age screen.

---

[6] Trust, Transparency and Control Labs, 'How to design with trust, transparency and control for young people, *TTC Toolkit,* March 2021, https://www.ttclabs.net/research/how-to-design-with-trust-transparency-and-control-for-young-people

But we also recognise that some people will lie about their age online. For that reason, we have been investing and developing artificial intelligence tools to help us understand someone's real age. We've developed technology that allows us to estimate people's ages, like if someone is below or above 18, and this is currently in place in Australia. We train the technology using multiple signals. We look at things like people wishing you a happy birthday and the age written in those posts: for example, "Happy 21st Birthday!". We also look at the age users have shared across apps: for example, if a user has shared their birthday on Facebook, we'll use the same for linked accounts on Instagram.

We're focused on using existing data to inform our artificial intelligence technology. Where we do feel we need more information, we're developing a menu of options for someone to prove their age. This is a work in progress.

We're also in discussions with the wider technology industry on how we can work together to share information in privacy-preserving ways that helps apps establish whether people are over a specific age. One area we believe has real promise is working with operating system (OS) providers, internet browsers and other providers so they can share information to help apps establish whether someone is of an appropriate age.

This has the dual benefit of helping developers keep underage people off their apps while removing the need to go through differing and potentially cumbersome age verification processes across multiple apps and services. While it's ultimately up to individual apps and websites to enforce their age policies and comply with their legal obligations, collaboration with OS providers, internet browsers and others would be a helpful addition to those efforts.

Technology like this is new, evolving and it isn't perfect. It also may not always be the most appropriate measure for all use cases. Inaccurate AI predictions could undermine people's ability to use services, for example, by incorrectly blocking them from an app or feature based on false information.

However, we think a range of tools are needed to understand the ages of users on our services, in order to provide controls and tools that enable an age-appropriate experience.

Age-appropriate controls

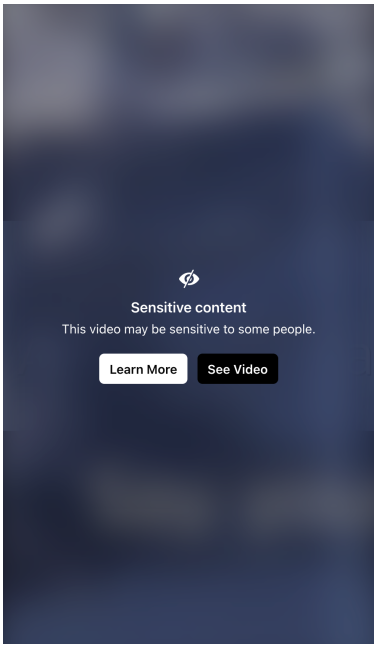
For those users that we know or suspect are between the ages of 13 and 18, we take a number of steps to ensure they have an age-appropriate experience on Facebook and Instagram:

- **Warning label for sensitive content:** There are various categories of content that we may allow on our platform for public interest, newsworthiness or free expression value, that we recognise may be disturbing or sensitive for some users. This may include:

- Violent or graphic content that meets our list of exceptions (for example, it provides evidence of human rights abuses or an act of terrorism).
- Adult sexual activity or nudity that meets our list of exceptions (for example, culturally significant fictional videos that depict non-consensual sexual touching).
- Suicide or self-injury content that is deemed to be newsworthy.
- Imagery of non-sexual child abuse, where law enforcement or child protection stakeholders ask us to keep the video visible for the purposes of finding the child.

Once a piece of content is identified as 'disturbing' or 'sensitive', we apply a warning label that limits users from being unable to see the content unless they click through. The content will not appear, or present the option of viewing it, for users who are under the age of 18.

Example of a piece of content that is "marked as sensitive" on Facebook



- **Stricter controls for sensitive content:** We recently announced a Sensitive Content Control on Instagram. We recognise that people have different preferences when it comes to sensitive content which does not break our rules but could be potentially upsetting (such as sexually suggestive or violent content). This new tool allows users the control to adjust their preference to either see more or less sensitive content, in line with their preferences. We default users who are under the age of 18 on Facebook and Instagram into being unable to see sensitive content; however, this tool also allows them and their parents the option of removing even more potentially sensitive content, i.e. for an even safer viewing experience.[7]

[7] Facebook, 'Introducing sensitive content control', *Facebook Newsroom*, 20 July 2021, https://about.instagram.com/blog/announcements/introducing-sensitive-content-control

- **Defaulting new teen accounts to private:** Wherever we can, we want to empower young people to make the right choices about what information they share online. We believe private accounts are the best way to do this. In line with this, we now default all new Instagram users who are under the age of 16 in Australia onto a private account.

- **Default account limitations:** there are a range of other default limits that are placed on a minor's account on Facebook. For example, minor profiles cannot be found on Facebook or search engines off our platform; Post and Story audiences are defaulted to Friends (rather than public); and Location is defaulted off.

- **Encouraging existing teen accounts to be private:** For young people who already have a public account on Instagram, we have recently announced we will show them a notification highlighting the benefits of a private account and explaining how to change their privacy settings. We'll still give young people the choice to switch to a private account or keep their current account public if they wish.

- **Restricting adults from privately messaging young people:** Since 2020, we have sent safety notices to users in Messenger, and subsequently Instagram, if we believe an adult could be pursuing a potentially inappropriate private interaction with a teen. These notices are designed to discourage inappropriate interactions with children and to limit the potential for grooming to occur via Messenger and Instagram.[8] These are over and above restrictions in place on Messenger and Instagram preventing an adult from privately messaging an unconnected young person.

- **Making it more difficult for adults to find and follow teens:** We've developed new technology that will allow us to find accounts that have shown potentially suspicious behaviour and stop those accounts from interacting with young people's accounts. By "potentially suspicious behaviour", we mean accounts belonging to adults that may have recently been blocked or reported by a young person for example.

  Using this technology, we won't show young people's accounts in Explore, Reels, 'People You May Know' or 'Accounts Suggested For You' to these adults. If they find young people's accounts by searching for their usernames, they won't be able to follow them. They also won't be able to see comments from young people on other people's posts, nor will they be able to leave comments on young people's posts. We'll continue to look for additional places where we can apply this technology.

  These changes are being rolled out in Australia and a small number of other countries initially, and will expand to include other countries soon.

[8] J Sullivan, 'Preventing unwanted contacts and scams in Messenger', *Messenger News*, 21 May 2020, https://messengernews.fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger/.

- **Keep Instagram as a fun, low pressure place for teens to express themselves**. We have tested on Instagram the ability to hide like counts, to see if it helps depressurise the experience for young people and reduce social comparison. We found that some users found it beneficial, and some users found it annoying. So we've now given all users - including young people - the choice on whether they would like to hide like counts on Instagram.[9]

- **Limiting advertisers' ability to reach young people:** We've also recently announced changes to how advertisers can reach young people with ads. We now only allow advertisers to target ads to people under 18 (or older in certain countries) based on their age, gender and location. This means that previously available targeting options, like those based on interests or on their activity on other apps and websites, will no longer be available to advertisers.

This is in addition to age-gating controls made available for those advertisers who publish age-sensitive ads or content (such as related to gambling).

We already give people ways to tell us that they would rather not see ads based on their interests or on their activities on other websites and apps, such as through controls within our ad settings. But we've heard from youth advocates that young people may not be well equipped to make these decisions. For this reason, we are taking a more precautionary approach in how advertisers can reach young people with ads.

Underage users

Facebook and Instagram are designed for users aged 13 and above.

We allow anyone to report suspected underage user accounts on Instagram and Facebook. Our content reviewers are also trained to flag reported accounts that appear to be used by people who are underage. If these people are unable to prove they meet our minimum age requirements, we delete their accounts.

However, the technology industry faces a significant challenge in ensuring that users below the minimum age do not gain access to services that are not designed for them. The reality is that they're already online, and there is no foolproof way to stop people from misrepresenting their age.

In order to reduce the incentive to misrepresent their age, we are also working on providing products and experiences designed specifically for users under 13, managed by parents and guardians.

In 2020, in response to the COVID-19 pandemic, we accelerated the launch in Australia of a product called Messenger Kids. This is a new messaging product for users who are not yet 13, and provides them with much greater privacy and security

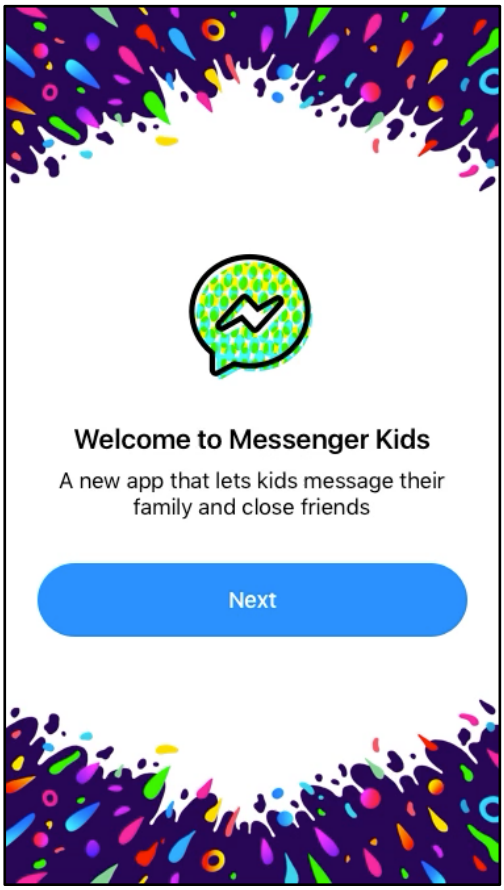
[9] Facebook, 'Giving people more control on Instagram and Facebook', *Facebook Newsroom*, 26 May 2021, https://about.instagram.com/blog/announcements/giving-people-more-control

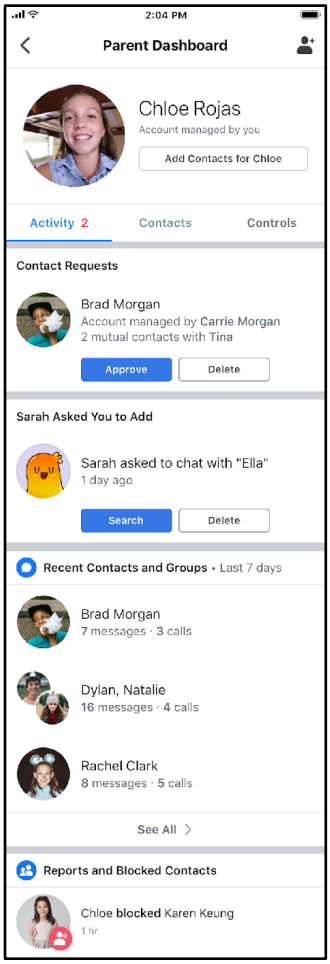
controls than regular Messenger. It's a fun way for younger users to connect with their friends, especially while in lockdown or isolation during the pandemic.

Parental control is at the heart of Messenger Kids. Parents manage who their child interacts with and can monitor their child's activity in the app through the Parent Dashboard, where they can also download their child's information at any time.

The design of Messenger Kids, and the control measures, have been developed after extensive consultation with a team of experts in online safety, child development and media, as well as parents.

Screenshots from Messenger Kids, including the Parent Dashboard (right)





We are also working on a new Instagram experience for tweens. We believe that encouraging them to use an experience that is age-appropriate and managed by parents is the right path. But we know how important it is to get this right, and so we are working closely with experts around the globe on how we could design a product that is compelling enough for tweens to want to use it, while providing a best-in-class safety experience. We are developing these experiences in consultation with experts in child development, child safety and mental health, and privacy advocates.

Providing fun and compelling products for young people in this age group helps to significantly reduce the incentive for them to join Facebook and Instagram, which are not built for users who are under 13.

We continue to prioritise tools and controls that give age-appropriate experiences to young people.

A possible global best-practice model

Facebook and Instagram are working hard to proactively offer products, tools and controls that give young people an age-appropriate experience. But we also believe that regulation has a role to play in ensuring all players in industry are committed to the challenge.

Facebook has supported the Government's enhancement of online safety laws via the Online Safety Act. And we commend work on the UK's Age Appropriate Design Code as a good starting point for regulators in Australia considering new requirements for protecting the data of young people.

There is significant discussion and debate around the world on the right role for regulation to play in encouraging age-appropriate online experiences. It's a complex question.

But we can see potential for the overall regulatory framework for age assurance under the Online Safety Act, including the Restricted Access System declaration, if well-designed, to be a novel and world-leading approach to age-appropriate content regulation. This section suggests how such regulation could be shaped in this way, by providing suggested principles and design of an effective scheme.

Principles

We believe effective regulation in this space will be underpinned by a number of principles:

- **Youth empowerment.** Teens come to Instagram and Facebook to express themselves, to keep up with their families and best friends, to find new passions and interests, and to follow the creators and celebrities they admire. These services provide opportunities for teens to organise around things they care about, support underrepresented voices and push for societal change. They are the place where teens explore their identities and find people they identify with. We strongly believe in the importance of responsibly empowering young people, and ensuring regulation does not remove their choice or agency.

- **Age-appropriate safeguards.** Younger users require additional safeguards for their safety, privacy, and wellbeing. Depending on the nature of the online service, some of these safeguards should be in place by default.

- **Innovation.** Industry is moving quickly and there are a lot of developments in the area of age-appropriate experiences internationally. Good regulation should encourage this innovation by ensuring not to prescribe particular technologies or processes (which may quickly become outdated) and allow the flexibility for new approaches to age assurance and restricting access to objectionable or sensitive content.

- **Privacy-preserving.** Regulation should respect the data protection principle of data minimisation, and should not be dependent on excessive collection of sensitive data.

In particular, good regulation will not depend on the collection of young people's ID. There are significant limitations to relying on ID collection: many young people don't have an ID, ID collection isn't a fair or equitable solution, nor is it foolproof. Lack of ID access disproportionately impacts underserved communities around the world. Even if they did have an ID, some young people may be uncomfortable sharing it. For example, perhaps they're a young member of the LGBTQIA+ community and they worry about having their identity attached to a pseudonymous account.

Designing an effective Restricted Access System

A large number of different regulations will comprise an overall framework of regulating for age-appropriate experiences online. Some of these regulations are not yet known. But it appears that the framework will impose two different age-related sets of obligations on digital platforms: (1) to restrict access to objectionable or sensitive content by minors; and (2) to verify the age of users to ensure the access restrictions are effective. The Restricted Access System cuts across both, as the previous declaration included both measures to assure the age of users but also steps taken to restrict underage access to certain content.

The concept

The age assurance regulatory framework has the foundation and makings of a "notice and age gating" scheme for certain types of user generated content. The legislation specifies that eSafety is able to advise a digital platform of a piece of content that is inappropriate for young people and the platform can either take steps to age-gate it (ie. keep up the content but restrict the ability for users under 18 to view it) or remove it. The requirements for age gating the content will be set out in the Restricted Access System declaration.

This model is an innovative approach to regulation, and it potentially has many benefits:

- It is similar and comparable to how notice and takedown schemes work, meaning it can draw from existing processes within both regulators and companies.

- It could potentially be more protective of free expression than takedown schemes, allowing takedown schemes to be reserved for the most harmful material, while still protecting the safety of young people online. For example, there are categories of Restricted Material[10] that Facebook believes we should not be taking down: for example, graphic footage of human rights abuses in a war zone. Allowing these to be subject to age-gating (once a notice is received) would mean the substantial public value of this content can be preserved.

- It is entirely consistent with the legislative requirements around the Restricted Access System.

However, the interaction between the Restricted Access System declaration and the other regulatory components of the age assurance framework will be critical. If the overall framework deviates from the "notice and takedown" or "notice and age-gate" scheme in the legislation, it could become disproportionate and problematic to implement.

The details

We recommend that the Restricted Access System declaration should allow for flexibility in the technology and measures that companies use for age assurance or restricting access.

Companies subject to the previous Restricted Access System declaration relied heavily on tick-boxes for users to verify that they are over 18. A framework built to rely on a single line of defence, rather than multiple measures working in concert, is less likely to be effective. The application of the Restricted Access System Declaration to social media companies provides opportunities, particularly as we anticipate there will be future innovation from social media companies in future.

The earlier sections outline the way that Facebook is developing our thinking about appropriate measures for assuring age or restricting access to content. Although not all of these will necessarily be appropriate for codification in regulation, the principles that underlie these efforts may be helpful in designing more specific parts of the Restricted Access System declaration.

We would welcome the opportunity to work closely with eSafety and the Government further on the details.

Key concerns

Notwithstanding the potential for the regulatory framework under the Online Safety Act to be a global exemplar of regulation, there are some indications about the direction of the regulatory framework that are cause for concern.

[10] This is the term used in eSafety consultation paper to refer to content caught by the Restricted Access System, namely certain categories of class 2 material: material that has been, or would likely be, classified as R 18+ or Category 1 restricted under Australian classification legislation.

We have some concerns that elements of the regulations would establish a scheme that is disproportionate, based on unrealistic expectations about the capabilities of technology, or not entirely privacy-protective. Primarily our concerns are less about the concept of a Restricted Access System itself, and more about how the interrelated regulations could build on the requirements in the declaration. We offer our concerns below - noting that it is very challenging for industry to fully comprehend how the full range of interrelated age assurance regulations will interact at this point (especially when some components of the framework are not yet known).

Specifically, we have the following concerns:

- **The definition of "class 2 material" to be captured by the Restricted Access System is very broad, ambiguous and subject to change.** Although the public commentary around this content has focussed on pornography, the definition of Restricted Material to be caught by the Restricted Access System is ambiguous and is much broader than pornography alone. It could cover examples as broad as: a written piece of text relating to drug use; music or audio files that depict high severity violence; or public interest footage from a war zone.

  While the definitions are well-established for films, publications and computer games, the challenge arises in applying these definitions to user-generated content. Platforms are required to make complex assessments about how online content "is likely to" be classified. The classification scheme was designed and is suitable for distributors of films and games, who have full control over the types of content that they distribute. It poses a number of major implementation difficulties for digital platforms who host a variety of user-generated content. For example, multiple copies or edits of essentially the same piece of content could result in dramatically different classification ratings, meaning a platform cannot assume that - just because a previous copy of that content is likely to receive a certain classification - all copies would be classified the same.

  The definition is also subject to change, as the Department of Communications is currently reviewing classification legislation.

  Given the breadth of the definition of Restricted Material, there are large gaps of content that could be Restricted Material that Facebook would not consider to violate our Community Standards, or to fall under our view of content that is objectionable or sensitive. The two primary types of content that fall under this category are (1) fictional depictions of some forms of violence, drug use, or sexual activity; and (2) music, audio or text that would be classified as class 2 (for example, a song with high-impact swearing that might be used in a video clip).

- **There should not be a requirement on companies to proactively search and restrict access to class 2 material**.

The Online Safety Act requires that platforms either remove or age-gate a piece of Restricted Material, once they have received a remedial notice from the eSafety Commissioner. This is consistent with how the other aspects of the Online Safety Act work, and provides a workable basis for a regulatory scheme to age-gate content on a social media platform.

The regulatory framework for age assurance should not go further than this by setting a requirement for digital platforms to proactively search for class 2 material and remove it or apply age-gates to it.

Facebook and Instagram already proactively search for content on our platform that violates our Community Standards (such as adult sexual material, nudity and graphic or violent content) and material that could be objectionable or sensitive - much of which would likely fall within the definition of class 2 material. However, the very broad and ambiguous nature of the content which may fall within class 2 material means existing technologies would not be able to detect everything that falls within the legislation.

Given the ambiguity of the class 2 definition, we do not believe it would be fundamentally possible to build technology that could effectively meet this requirement at the scale required for user-generated content. While we understand some film distributors have developed some technology that relates to the Australian classification regime, user-generated content is fundamentally different and much more challenging and complex to assess.

A model that relies on receiving a notice from the eSafety Commissioner's Office would be more effective, proportionate and comparable to existing notice and takedown schemes.

- **Measures for ensuring age-appropriate experiences for private messaging should be different to public social media.** Private messaging is fundamentally different to public social media, and is more akin to SMS or email. Users do not come across content incidentally: someone (usually someone who they are already connected to) needs to have intentionally sent a piece of content to them. There is no broad-ranging search function, private messaging services do not suggest or promote content to users, nor is there an aggregated feed of content.

  And the greatest challenge for age assurance arises from the fact that private messaging services collect (and have access to) very limited categories of user information, especially when compared to social media. Far less detailed personal data is required to use certain services such as these.

  For example, Facebook builds our messaging services with privacy and data minimisation at heart, and WhatsApp deliberately collects limited categories of user information. As a result, the accuracy of modelling and machine learning related to user interactions is consequently very limited.

Instituting the same age assurance requirements for private messaging as social media would necessitate *increased* data collection by private messaging services, which comes into tension with data protection principles such as data minimisation, purpose limitation, storage limitation, and security. We believe this is not what Australians expect. Privacy is clearly the expectation of consumers: for example, on WhatsApp, 90 per cent of chats are between two people, and the average group size is fewer than ten.

Our concern is even greater at any suggestion that we should scan for the categories of content in the Restricted Access System (Restricted Material) on end-to-end encrypted services, where the provider cannot see the content of private messages.

We use end-to-end encryption for WhatsApp and are working towards applying it to Messenger and Instagram Direct. The use of end-to-end encryption means it is not possible for anyone other than the sender or the receiver to view the content of a message. End-to-end encryption protects privacy and helps keep people, including young people, safe from hacking, fraud and other forms of cybercrime. Encryption is therefore an essential safety tool.

There are ways of encouraging age-appropriate use of private messaging that do not rely on seeing the content. For example, WhatsApp takes a number of steps around age assurance. These include:

- Providing clear guidance about the minimum age a user must be (i.e. 13 or 16 years old) in order to use WhatsApp.
- Providing a channel for people to report an underage user.
- Promptly disabling WhatsApp accounts where it is reasonably verifiable that the account belongs to an underage user.
- Making it easy for parents to delete a child's account should they wish to do so.

However, it is not possible to restrict content (messages) if we are not able to see the content. We have built a number of safety features into private messaging services (especially to combat the sharing of child sexual abuse material), but these are not dependent on seeing the content.

We very strongly contest the suggestion that we should be proactively scanning the content of private messages to detect Restricted Material (the definition of content caught under the Restricted Access System) or class 2 material generally. We do not believe Australians would expect or want Facebook to monitor their private messages for nudity or swearing. Our concern is even greater at the suggestion that we should scan for this material on end-to-end encrypted services, where the provider cannot see the content of private messages.

We would welcome the opportunity to work with the eSafety Commissioner on the types of age assurance measures that would be suitable for private messaging.

- **The regulatory framework around age-appropriate experiences should not expect perfection.**

  The legislative framework already sets very high expectations about the efficacy of digital platforms' measures to age-gate content. If a platform fails to proactively detect Restricted Material on just *two* occasions, they could be potentially subject to a service provider notification from the eSafety Commissioner.

  The components of the overall age assurance framework that are not yet known or finalised (such as the Basic Online Safety Expectations or other industry regulations) hold the risk of going further. The current draft of the Basic Online Safety Expectations requires companies to take reasonable steps to prevent access to class 2 material by children. It is entirely possible the eSafety Commissioner could find a provider has not met basic expectations because the inherent limitations of current technology means they have not detected (and age-gated) a very small number of pieces of content. The eSafety Commissioner has full discretion over what constitutes reasonable steps.

  The forthcoming regulations (especially the draft Basic Online Safety Expectations) should recognise that technologies and processes for age-appropriate experiences online are still evolving and are not foolproof. Although there is significant progress and innovation in this space, the technologies and processes for detecting relevant content and applying age-gates are imperfect.

  Sophisticated content regulation around the world recognises that enforcement is imperfect.[11] Just as we recognise that it is not possible for companies to proactively detect and remove every piece of harmful content, it is also not possible to detect and age-gate every piece of objectionable or sensitive content.

- **The age assurance framework should not prescribe particular age assurance technologies.** There is significant innovation underway and Facebook has added new methods and tools in this regard in the last year alone. We are still assessing the effectiveness of these and other technologies, and regulation should allow companies to evolve and to deploy measures that are right for their services.

  A framework built to rely on a single line of defence, rather than multiple measures working in concert, is also less likely to be effective. Similarly,

[11] M Bickert, 'Charting a way forward - online content regulation White Paper', *facebook.com*, February 2020, https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf ; E Douek, 'Governing online speech: from 'posts-as-Trumps' to proportionality and probability, *Columbia Law Review*, 27 August 2020, vol. 121, no. 3, https://columbialawreview.org/content/governing-online-speech-from-posts-as-trumps-to-proportionality-and-probability/

companies need to balance data protection principles - such as data minimisation - in the deployment of these technologies.

Age assurance should also take account of evolving technological solutions. We are continuously improving our approach to age assurance and working on new ways to verify age. This requires innovative, privacy-preserving techniques to help prevent young people who should not be on our services from opening an account and, where a user is younger, ensure they get age-appropriate experiences and important safeguards when online.

Similarly, while we currently mark content as sensitive or disturbing (which renders it unable to be viewed for any users we know to be underage), there may be further innovation in how to deploy technical measures to restrict access of this content to young users.

If the regulatory framework around age assurance is overly prescriptive and fails to allow room for innovation, there is a real risk that service providers will be forced to abandon it and will simply remove the content for *all* users - even if it has high levels of public interest value.

We accept there is more to do. And so we support closer industry collaboration to develop effective measures to ensure young people consistently receive age-appropriate experiences across the online ecosystem.

On both age verification and restricting access, there are opportunities for players from across industry to work together (including phone operators, platforms, and apps), towards comprehensive outcomes rather than interpreting age assurance as a singular responsibility of each company. Each part of the industry could play different roles, and so the underpinning regulations should not prescribe the same measures for all types of players.

We believe it is ultimately for industry to continue to innovate in this area, rather than by trying to prescribe innovation by regulation.

- **The timeframe for implementing the Restricted Access System declaration is very short**. Given the potentially burdensome and onerous requirements associated with implementing particular technologies or processes, it is unlikely to be feasible for companies to put these measures in place by January 2022 - a mere three months away.

To address these concerns, we make the following constructive recommendations:

- The definition of class 2 material should be clarified, especially for the purposes of Restricted Material caught by the Restricted Access System declaration.

  As the definition is critical to both the application of the Restricted Access System and to certain expectations in the draft Basic Online Safety Expectations, we suggest that the eSafety Commissioner issue detailed

guidance about types of content to be included, to adapt the definition to be more workable for platforms that host user-generated content.

- We recommend that the regulatory framework for restricting access should continue to preserve the notice and takedown/age-gating model contained in the legislation. That is, it should rely on the provision of notice from the eSafety Commissioner (as is contemplated by the Online Safety Act), and should not set a proactive requirement to detect class 2 material.

- The age assurance regulatory framework as a whole should not be prescriptive in setting what technologies or processes digital companies use to verify the age of users or age-gate content.

  The regulations should be flexible enough to allow service providers to employ a whole range of technologies (such as use of age screens or artificial intelligence) and should not be exhaustive.

- If the Restricted Access System declaration takes effect in January 2022 (as contemplated in the Online Safety Act), the eSafety Commissioner could issue guidance to providers that it will not be enforced until six months after commencement.

Our subsequent submissions on the Basic Online Safety Expectations and other safety regulations will provide further suggestions on how to address the concerns raised in this submission.