Online Hate Prevention Institute

Online Hate Prevention Institute submission on Social Media and Australian Society

Online Hate Prevention Institute submission to the Joint Select Committee on Social Media and Australian Society.

About this submission

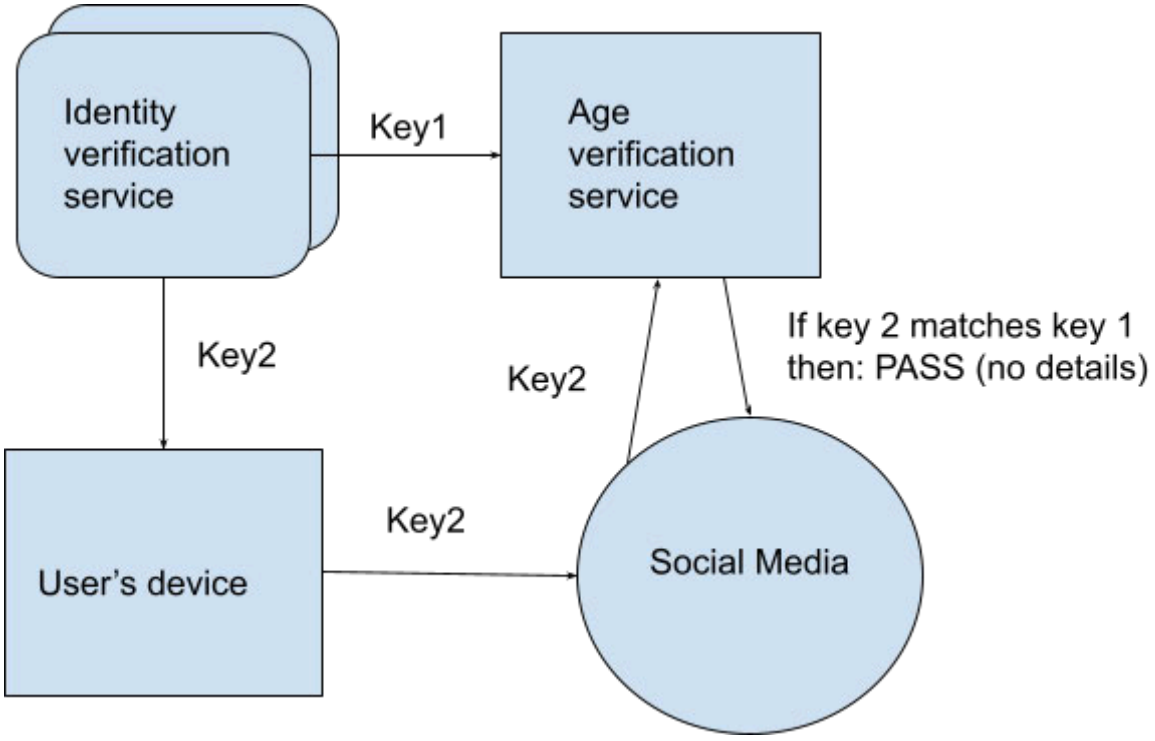
This submission is from the Online Hate Prevention Institute, an Australian Harm Prevention Charity established in January 2012 with a focus on online hate and extremism. This submission is prepared by our CEO, Dr Andre Oboler.

(a) The use of age verification to protect Australian children from social media

The Online Hate Prevention Institute believes the fundamental business model of most social media is harmful. Users are treated as the product, not the customer, and the systems are built to increase engagement which can often be best done through farming outrage, insecurity, a fear of missing out, and gamification that rewards regular and sustained use. Other business models are in fact possible.

When age verification is needed

Age verification for social media can protect Australian children from the current often toxic nature of social media. Where age verification is required, we recommend using a system that protects privacy by issuing a time limited token that verifies age without sharing a person's identity. As an example of such a system is shown below:



The system starts with an Identity Verification Service. This is an electronic system a user can securely log into, where their account has been strongly tied to their identity after verification against government ID. This service might be provided by government agencies, banks, schools, or universities.

Once a user is logged in to the Identity Verification Service, they can generate a short term set of two keys used to prove they are over a specific age. Key1 is sent to an Age Verification Service, and key2 is kept on their device and can be provided to any application (e.g. a social media platform) requiring age verification. The application provides the received key2 to the Age Verification Service which, if there is a matching key1, indicates the user is over the minimum age.

The Identity Verification System might give multiple key generation options, for example a 15+ key, an 18+ key, and an under-18 key. The keys would in one sense function in a similar manner to 2 factor authentication (and may in fact require 2-factor authentication to login to the Identity Verification System in order to generate the keys). In another sense they can operate in a similar manner to public key encryption for identification. The end result is that the application (e.g. social media platform) does not need to know the user's real identity, and neither does the age verification system. The Identity Verification System, which knows who the user is, does not have any information on what applications they are using.

When age verification is not needed

Age verification should be unnecessary for social media platforms built specifically for children (e.g. *YouTube Kids*) or for a general audience but with strong protections for users and business models that make the first priority user health and safety. Such systems will by necessity require strong limits on free expression making them inappropriate spaces for political debate, activism, or some journalism. Essentially they would be spaces where all content is required to be the equivalent of a "G" in the Australian Classification ratings.

Systems that are exempt from age limits and verification:

- Must take a more responsible approach and use business models that avoid incentivising harm.
- Be built using safety by designed principles to address known risks more completely than currently required by the Basic Online Safety Expectations. Advanced Online Safety Expectations would need to be created to provide the expected requirements for age limit exempt systems.
- Require regular certification that platforms still meet the requirements for exemption.

Advanced Online Safety Expectation would, among other requirements, need to:

- Outlines specific risks to be avoided or mitigated
- Prescribe certain mitigation mechanisms for some risks
- Proscribe tolerance levels for harms that cannot be entirely avoided and requirements to ensure this is measured, transparently reported, and independently verifiable. Environmental protection mechanisms provide a useful model for doing this.
- Require regular transparency reports that are specific to the Australian experience of using the platform, and specific in terms of the type of harm. For example, hate speech is not a specific enough category, but hate speech targeting First Nations

Australians is. The intersection of multiple categories of hate should also be reported when over a specified level, for example, hate speech targeting First Nations Australians that involved gendered hate.

(b) The decision of Meta to abandon deals under the News Media Bargaining Code

We believe the presence of news content from news media on social media platforms is important to democracy. The absence of news media leaves a void for commentary and fake news from sources that are not subject to media ethics or codes of conduct. We believe this poses a danger to democracy and public safety.

At the same time, our research shows that the comments on social media, made on news articles, can create toxicity on social media. Preventing this requires active moderation by social media companies, particularly on articles that are on controversial topics. The headline and summary of a news article can significantly incite hateful and abusive comments, and division in society, when they aim to generate high engagement.

Media companies should be financially rewarded for contributing content to social media, but financially penalised when their content creates toxicity that generates reports social media companies then need to address. User reports should first go to the media company, then to the social media company if not addressed, or if appealed by the reporter and then upheld by the social media company.

Social media companies should also be required to provide grants of free advertising to non-partisan organisations advancing the public good in Australia. This should apply to DGR Type-I organisations, excluding overseas aid charities.

(c) The important role of Australian journalism, news and public interest media in countering mis and disinformation on digital platforms

Australian journalism, news and public interest media play an important role in countering misinformation and disinformation, as discussed above, but so do non-partisan civil society organisations that have a public interest mandate. A wider focus is needed, noting that the media isn't the only trusted voice when it comes to the online space.

(d) The algorithms, recommender systems and corporate decision making of digital platforms in influencing what Australians see, and the impacts of this on mental health.

Current systems tend to create echo chambers. In some cases this can lead people down a rabbit hole of hate and conspiracy theories. It can also lead to more divisive content that

divides society and drives people apart as this leads to greater tension, stronger engagement, and therefore algorithmically more visibility which can directly, or indirectly, translate into more revenue for the content creator. This is harmful to our multicultural society and to rational discussion and debate. It harms democracy as well as individuals' mental health.

Conversely, content promoting facts and rational discussion only gains visibility on some platforms if it is boosted / paid for. This is a problem for public service content, e.g. from charities and more generally from experts. The media can't provide the volume of content to serve as a gateway. There needs to be some calibration to increase visibility of local content that informs and contributes to the public good.

(e) other issues in relation to harmful or illegal content disseminated over social media, including scams, age-restricted content, child sexual abuse and violent extremist material

We are concerned hate speech was not explicitly called out in this list. We deal primarily with hate speech and extremist content. We are responsible for removing multiple original copies of terrorist manifestos and videos of attacks. The role we play has been praised by both governments (including internationally) and social media companies. A strategic partnership between the Online Hate Prevention Institute and government is needed to leverage the 12 years of world class expertise developed in this space in Australia.

We take this opportunity to share details of our current work in 2024 relevant to this question. More detail is available on request. Our key message is that such work exists, and more could be done by the Online Hate Prevention Institute, and better use made of the existing material, if a strategic partnership was put in place.

Project 1: *Online Antisemitism in Australia*, a partnership with the Executive Council of Australian Jewry, has been running for 18 months and monitors 11 online platforms on an on-going basis. It has resulted in over 5,500 items of data being collected and three reports published. The data is classified into 27 categories, the prevalence of different categories and the effectiveness of platforms in removing each category varies. The project incorporates data from intensive monitoring in the *Moment Project* described below.

Project 2: *Building regional and national capacity in civil society to counter extremism* is a project funded under a grant from Home Affairs. The project supports the training and employment of three young people from remote communities to monitor and analyse harmful online content, create publications, provide training to local organisations, and support OHPI Exit's training for those supporting people vulnerable to extremism.

Project 3: *New Zealand Online Antisemitism Project* sees 4 young people in New Zealand being employed in New Zealand and seconded to OHPI for training and management to monitor and write articles about online antisemitism in New Zealand. 514 items of antisemitism have been collected and 6 articles produced in the last 2 months.

Project 4: The *Moment Project* documents both online antisemitism and online anti-Muslim hate between October 2023 and February 2024. The project involves a total of 320 hours of intensive monitoring on 10 different social media platforms, with equal time spent on each platform and that time evenly divided between work on antisemitism and anti-Muslim hate on the platfrom. The monitoring distinguishes between 27 types of antisemitism and 11 types of anti-Muslim hate. A resulting report *Online Antisemitism After October 7* shows a dramatic rise of online antisemitism, varying from 350% to 1000% by platform. The main type of antisemitism is traditional antisemitism, such as conspiracy theories, deicide, and blood libels. This traditional antisemitism is also used in relation to Israel and accounts for most of the Israel related antisemitism, e.g. accusing Israel of controlling the media or other national governments. Takedown rates vary by platform from 4% to 36%. A forthcoming report on anti-Muslim hate (which also covers anti-Palestinian racism and anti-Arab racism) notes this hate has also risen and is particularly prevelant on X (Twitter) and platforms used by the far right. The most common categories of anti-Muslim hate are demonising / dehumanising Muslims, presenting Muslims as a cultural threat, and presenting Muslims as a security risk. Take down rates vary from 2.9% to 38%, again, far from acceptable.

Project 5: Our *Referendum Project* identified 161 media articles that were shared in 528 social media posts during the campaign. Each post had 10+ comments, and we collected and analysed a total of 37,785 comments that were made across these posts. The posts were analysed for disinformation about The Voice, campaigns and the referendum process. We also looked for racism against First Nations People and others, cyberbullying and other kinds of hate. We found that 2249 of the total comments contained misinformation about Campaigns, and 1004 contained anti-Indigenous racism. These numbers are after removal by both platforms and media companies, so the initial numbers would likely have been significantly higher.