STATUTORY REVIEW OF THE ONLINE SAFETY ACT

ADM+S SUBMISSION

Contributing authors: Kath Albury, Zahra Stardust, Kimberlee Weatherall, Haiqing Yu.

ARC Centre of Excellence for Automated Decision-Making and Society

21 June 2024

About ADM+S

The ADM+S is pleased to have this opportunity to engage with this review of the *Online Safety Act*. The ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) is a cross-disciplinary, national research centre established and supported by the Australian Research Council to create the knowledge and strategies necessary for responsible, ethical, and inclusive automated decision-making (ADM). This submission draws on research from across ADM+S.

This submission

This submission is the product of a collaborative process involving direct contributions from researchers from ADM+S, seeking to highlight ADM+S research relevant to certain questions raised in the review. ADM+S researchers come from many different institutions, disciplines and perspectives. It should not be assumed that every contributing author, or every member of the Centre subscribes to every comment or recommendation in this submission. The submission highlights research from ADM+S relevant to two points in the review: the question of **age assurance**, and the protection of persons who have experienced online harms, with specific reference to the question of **bystander** reporting. The text for this submission is drawn from a larger set of material prepared also for the Commonwealth Joint Select Committee on Social Media and Australian Society.

We also highlight **research projects and programs** relevant to the work of the eSafety Commissioner. We would be happy to connect the Review or the eSafety Commissioner with relevant researchers from the Centre's projects.

Age assurance and its limitations

The government is about to trial age assurance technologies to restrict access to pornography, and is considering restricting social media to young adults 16 and over. While the government refers to these tools as "age assurance", many of them are more accurately called "age estimation". Published in *Big Data and Society*, our new study into one common facial age estimation tool shows that such technologies are unreliable, and have a racial and gender bias.[1]

Civil society groups have cited privacy and feasibility concerns about age estimation technology. These include: accessibility issues for people without identity documents; the potential burden on public interest projects such as Wikipedia, and small, low-income websites; queries about what data could be collected, sold or exploited; and the likelihood of circumvention.

Age estimation is already a fraught task when done by humans, who regularly misjudge age. It is no better when done by machines. Age estimation software that uses facial recognition relies on stereotypical indicators of age, such as hair, wrinkles and jawlines. These are highly variable – for example, wrinkles can be altered by cosmetics or injectables. Studies also indicate that facial recognition software often has a significant racial and gender bias.

In our research, we used an accepted industry-leading convolutional neural network technology to analyse a dataset of 10,139 facial images. We found that the model was most accurate in estimating age in the "Caucasian" category and least accurate in the "African" category. Boys were more likely to be misclassified than girls, especially in the 0–12 age bracket. People aged 26 and over were generally misclassified as younger, sometimes by as much as 40 years.

Another study published by ADM+S researchers in the journal *Information, Communication & Society* looked at the implementation of age estimation video surveillance set up on the physical premises of a large Australian gambling chain. When the developers of the age estimation tool were interviewed they admitted that it was of limited efficacy in detecting underage subjects. The tool was set to "err on the side of caution", but this necessitated cumbersome real-life double checking of the system's alerts. Ultimately the study concluded that the age-estimation tool was largely "performative in nature", with humans still required to do the actual work of age-verification.[2]

[1] Stardust, Z., Obeid, A., McKee, A., & Angus, D. (2024). Mandatory age verification for pornography access: Why it can't and won't 'save the children'. Big Data & Society, 11(2). https://doi.org/10.1177/20539517241252129

[2] O'Neill, C., Selwyn, N., Smith, G., Andrejevic, M., & Gu, X. (2022). The two faces of the child in facial recognition industry discourse: biometric capture between innocence and recalcitrance. *Information, Communication & Society*, *25*(6), 752–767. https://doi.org/10.1080/1369118X.2022.2044501

In the eSafety Commissioner's own research young people were concerned age assurance is of limited efficacy, and comes with privacy and security issues.[3]

It is also worth questioning whether proposed age assurance mechanisms or limits on access to social media will address societal concerns about the activities of minors online. A counterpoint is offered by developments in China. In 2019, Chinese authorities restricted minors to playing 90 minutes a day on weekdays and banned them from playing between 10 p.m. and 8 a.m. Harsher restrictions followed in 2021: minor gamers can only play for an hour a day and only on Fridays, weekends and public holidays. Then in December 2023 draft legislation was publicised that would limit how much people (not just minors) could spend on games.

These regulations respond to public concerns over gaming addiction among the youth. Gaming addiction and mobile phone addiction are called the new "spiritual opium" by the media, moralists, conservative parents and educationalists in China. They are part of the same growing pains with digital technology that all countries and peoples are experiencing. No one is an island in the age of digital connectivity. We all share anxieties about online bullying, AI generated content in misinformation and opinion warfare, gaming and social media additions, to name just a few. Chinese regulators are setting important, notably different precedents in internet governance, social media governance, and AI governance that are a counterpoint to developments in Europe and the west more generally.

Notably too, China's regulators, like regulators elsewhere, have to balance protecting the minors with protecting the market interests of the game industry, particularly during economic downturn. The regulations described above hurt the gaming industry, causing Chinese gaming stocks to plunge and the market to shrink. As a result, Chinese gaming companies like Tencent have expanded their video game consumer market overseas. More recently, the December draft restriction has been withdrawn, and of the head of the regulatory body removed. At the time of writing, there is still the time restriction on minor gamers in China.

Protecting those who have experienced or encountered online harms, and the question of bystanders

In relation to protections for those who are at highest risk of online harms, and the question of whether the Act should empower 'bystanders' to report illegal or harmful material, care should be taken to recognise and protect positive rights to self-expression in digital environments.[4]

[3] eSafety Commissioner, *Questions, doubts and hopes: young people's attitudes towards age assurance and the age-based restriction of access to online pornography*, Report, September 2023.

[4] Article 19, International Covenant on Civil and Political Rights (ICCPR), General Comment No. 34, Human Rights Committee, 2011; Communication 488/1992; Resolution 32/2 Human Rights Council, 2016; UN Special Rapporteur on freedom of opinion and expression, reports: A/HRC/38/35 (2018) and A/74/486 (2019)

Populations known to be highly vulnerable to online harassment and abuse (including Aboriginal and Torres Strait Islander people and LGBTIQ+ people) may feel *less* safe in digital environments where increased surveillance and/or policing are framed as safety mechanisms.[5]

As noted by the United Nations, LGBTIQ+ communities globally are increasingly targeted by discriminatory 'wedge' campaigns that falsely frame gender-diverse people as threats to the rights and safety of women and children - and these campaigns are often waged in digital environments.[6] Any increased promotion of bystander reporting should be designed cautiously, with an understanding that it may inadvertently enable these organised forms of harassment.

Other relevant research projects

ADM+S has an active research program and projects in areas relevant to the work of the eSafety Commissioner, including:

- ADM+S PhD candidate Joanna Williams' thesis exploring why sexual health organisations do not produce social media content that aligns with the digital and sexual cultures of young Australians; work demonstrating that arbitrary and *ad hoc* content moderation practices of social media significantly constrain the content that sexual health organisations produce;

- Louisa Bartolo's PhD work on socially responsible recommendation on Amazon Bookstore and Twitch;

- Lucinda Nelson's PhD research on subtle, 'everyday' manifestations of online misogyny on social media platforms;

- The Ad Observatory Project, using novel citizen science approaches through national data donation and analytics to examine ephemeral and personalised advertising (or 'dark' advertising);

Insights from this research are summarised in our report, *AI and automated decision-making in news and media* (December 2023).[7] We would be happy to provide further information on this research or the Centre's broader research program around the impacts of generative AI.

[5] Stardust, Z., Gillett, R. and Albury, K., 2023. Surveillance does not equal safety: Police, data and consent on dating apps. *Crime, Media, Culture*, 19(2), pp.274-295; Albury K, Byron P, McCosker A, et al. (2019) *Safety, Risk and Wellbeing on Dating Apps. Final Report.* Swinburne University of Technology.

[6] UN Women (2024) *LGBTIQ+ Communities and the Anti-Rights Pushback: 5 Things to Know*. 24 May.

[7] Nguyen, Meese, Burgess and Thomas, *AI and automated decision-making in news and media* (Report, December 2023) doi: 10.60836/qnz4-kw43.