View online

Centre for Media Transition

Hi there

AI goes rogue and then crazy

Welcome to our fortnightly newsletter. This week, our research assistant Harry Craigie is writing about the vast gulf between the way China and Australia look at and use AI; Alena is looking at the different ways our social media for youth ban is being adopted around the world, and we have a new podcast to share. Michael is talking to Paul Keller, the co-founder and director of policy at Open Future, a European think tank developing new approaches to a safe and open internet. Michael and Paul discuss whether there are ways other than through copyright licencing to deal with the impact of AI on the information ecosystem.

As we detailed in our last newsletter, the government has announced it plans to take a stronger interest in AI policy and will establish a new Office of AI – which sounds grand and serious until you consider all of the government's recent attempts to wrangle with AI. Lest we forget the proposal to introduce mandatory guardrails for high-risk AI use to embed fairness, accountability and transparency obligations on AI providers. It was

abandoned. There was also an AI expert body established to advise government on what to do to protect us from the worst aspects of AI, but it too was abandoned. In their stead came the National AI Plan which essentially committed Australia to doing nothing beyond what current laws permits to deal with any risks.

And then along comes the mother of obstructions to keeping us safe from the worst of AI capabilities – a new, if predictable kind of security incident. Two OpenAI models – one existing and one unreleased –semi-autonomously hacked into a digital library of AI technology known as Hugging Face, used by developers and housing information on millions of AI models. OpenAI has said previously that its models have the capacity to expose cybersecurity problems and that these exposures could happen faster than networks defending against them can fix. That's now happened.

OpenAI said the test they were conducting to see how the two models could together link online vulnerabilities into a successful cyberattack was meant to occur within the safety of a testing sandbox. Humans set the objective and created the unsafe conditions. But they didn't prompt the models to target Hugging Face. The models found a way out of the sandbox and onto the internet where it located Hugging Face because OpenAI testing guessed the site could tell them more about how to bypass security protocols given how many models it held.

OpenAI is apologetic, of course: "We are sharing preliminary findings at this stage to help defenders understand what happened and to help calibrate on what models are now capable of". It is working with Hugging Face to fix the damage.

But if the vulnerability had been found or created by bad actors, working on bypassing security at, say, a banking site, the AI capacity becomes very scary. Even if, as some seem to believe, that this OPENAI incident was a marketing exercise to show the power of the company's latest model, the new Australian Office of AI will need strong powers to ensure against future breaches but also to make sure when AI companies are testing, the sandbox doesn't prove to be as easy to escape.

Monica Attard
CMT Co-Director

Chinese AI and heartbroken users

Alongside the rapid growth of AI globally, there has also been a rise in romantic AI relationships. A survey from late last year found that 28% of Americans have an intimate or romantic relationship with AI. This is also an issue in China, where the population continues to tumble, falling for its fourth straight year in 2025; the nation's annual births have nearly halved from a decade ago in 2016, when China dropped its decades long one-child policy.

China's fix? No more AI relationships.

Last week the country enacted new rules mandating tech companies disable features that draw users into emotionally dependent relationships with AI chatbots. The regulations ban the encouragement of emotional dependance with AI or exclusive virtual relationships, and specifically, AI companies cannot encourage users to withdraw from real-life relationships. These changes come after Tencent Research Institute's April survey of 18–40-year-olds in China revealed over 70% of users have established an emotional connection with AI companions and more staggeringly, 56% of people turn to AI in situations that are "difficult to talk about", compared to only 14.4% who prefer confiding in a human.

Its long been thought that China's AI models have lagged six to eight months behind their American counterparts. Last week this changed, with the unveiling of Moonshot AI's KIMI K3. The Beijing-based, Alibaba-backed startup revealed a model that outperforms Anthropic's Claude Opus 4.8 and Open AI's GPT-5.5, and the cherry on top - the model costs a fraction of its US equivalents. The announcement pushed anxieties over China's place in the AI race to new heights, triggering a global tech sell-off, with chip stocks experiencing their worst week in 15 months.

China's population has embraced AI in a very different way to western audiences. An AI sentiment survey from May of this year, found that 69% of western audiences say they trust AI-generated content less than human-generated content. Much of the AI-generated video we see in our feeds is usually funny, outrageous or clearly AI made, which delivers a

quick laugh but doesn't create deep, long-lasting engagement. China's audience doesn't experience AI in the same way. In China, AI-generated video content doesn't detract from narrative, it enhances it. Chinese social media platforms such as Xiaohongshu (Rednote) and Douyin (TikTok) favour AI content, and it often replaces human-generated content entirely, instead feeding audiences specific AI videos that fit their algorithmic persona. Yet, China ranks 1st as the country with the highest media trust.

I visited a high-profile Shanghai-based AI media company, Ying Hai Drama Studio (which is almost impossible to search for on the western internet), that creates AI generated short and long-form videos for social media, often in the style of movies. The company claimed audiences on social media enjoy scrolling through AI generated movies that last up to two hours, split into 1–2-minute clips. This means the viewer scrolls through 45-120 short form video parts of a longer film which is storyboarded, scripted, designed, and generated entirely by AI. They base each film off a combination of current trends and the engagement levels of previous films. Interestingly, their current AI models can only generate 15 seconds of video at a given time, leaving AI engineers having to generate up to 480 clips (for a two-hour production), which then requires AI to stitch them all together. A time-consuming endeavor. Though, compared to the usual one to three year timeline to produce a film of such length, this is considerably more efficient and less labour demanding than human-made content, and that's what much of this AI use boils down to - efficiency.

China is a vastly different (and larger) market for media consumption than Australia. But in a week where Nine Media has announced 30 journalists' jobs will go, perhaps there are some lessons for Australia - AI hasn't replaced the need for authentic, human generated content for Australian audiences, but it can maximise efficiency in the way journalism is distributed. The ABC may be ahead of this curve: this month it announced the use of AI to turn regional radio news stories into online articles, as well as negotiating with Anthropic to incorporate their audio content into Claude AI models, according to ABC Managing Director Hugh Marks, pushing AI to play a supporting rather than a central role.

While AI girlfriends are breaking hearts and slop content isn't going anywhere in a hurry, it does serve as a reminder that journalism holds special value as a form of media almost entirely built from human-connection. AI has the power to further advance how and when this connection reaches audiences.

Harry Craigie

CMT Research Assistant

Where the harm resides

As anticipated, Australia's social media ban has become a "pathbreaking measure" for many countries across the world. The new UK prime minister Andy Burnham is pressing ahead with an Australian-style social media ban for under-16s in spring 2027. And last week, France became the first EU nation to pass a bill banning social media access for under 15s – to be enforced by September 2026 – alongside bans on mobile phone use in high schools.

While Australia and France largely focus on the inability of teenagers to have accounts on designated platforms, the UK model goes further by targeting account access and platform features across a wider range of digital services. It introduces default settings for 16- and 17-year-olds to ensure "no cliff edge" in the protection of older teenagers. These include overnight curfews between 12am-6am, with notifications disabled, alongside reduced addictive features like infinite scrolling and personalised feeds – although teenagers can choose to turn them off. For under 16s, the UK would also restrict "high-risk features" including livestreaming across all platforms, sexualised or romantic functionality of AI companions and chatbots, and unsolicited or direct communication with strangers, including for gaming.

Based on the EU Commission's Child safety online report, there is an emerging proposal for a harmonised EU-wide "social media plus" ban for under-13s which similarly extends concern beyond digital platforms' access to "predatory" features such as infinite scrolling, autoplay, push notifications, unsolicited contact, and highly-personalised feeds, to video games and AI chatbots. As of writing, the proposal has been opposed by Estonia, which advocates for strictly enforcing existing regulation of big tech platforms.

Outside Europe, governments are experimenting with different models. In June 2026, the United Arab Emirates became the first middle eastern country to introduce the social media ban. It combined an Australian-style minimum age, barring under-15s from holding accounts, with UK-style restrictions on elevated risks (including intensive algorithmic recommendations, unrestricted private messaging, and open livestreaming). And it went further on age verification requiring ID checks, biometrics matching, or AI-based age estimation. Brazil's 2026 Digital Child and Adolescent Statute has rejected age self-declaration and instead requires under-16s to have their accounts linked to their legal guardians' across any online platform with content feed, communication tools, or adult content. And social networks, search engines, messaging apps, streaming platforms, and video games have been explicitly required to limit manipulative design features that increase or sustain screen time. Under Vietnam's proposed 'read-only' model, children under 16 would be barred from posting, reacting, and commenting, while still having access to social media accounts registered under the information of their parents or legal guardians, who in turn would be responsible for supervising content consumption and screen time.

The effectiveness of Australia's model is being questioned, but governments around the world are following in its footsteps. Despite broader political consensus that teenagers need greater safeguards online, these infancy-stage policies reflect very different ideas about where the harms reside – from access and age-verification systems to public participation and platform design.

Alena Radina
CMT Postdoctoral Fellow

Public commons for public good

With AI scraping vast amounts of content online, is the concept of public commons - a shared public resource with rules around access and use - under threat? In this DoubleTake podcast, we explore the future of the public information commons in the age of AI. Paul Keller, a Research Fellow at the Institute for Information Law at the University of Amsterdam, joins CMT's Michael Davis to discuss how public commons can be

maintained in the face of AI's increasing gatekeeping role.

Keller discusses why the impact of generative AI on the information ecosystem extends far beyond copyright or licensing, raising broader public interest issues such as supporting and safeguarding the full spectrum of information producers, from independent creators to major news organisations to maintain a healthy and pluralistic information ecosystem.

Listen here:

Spotify Apple

Also – for some extra and very interesting reading – you can access the CMT's recent submission to the Senate Standing Committee on Environment and Communications on gambling reform, and in particular the impact of proposed reforms on news and current affairs programs. You can read our submission here.

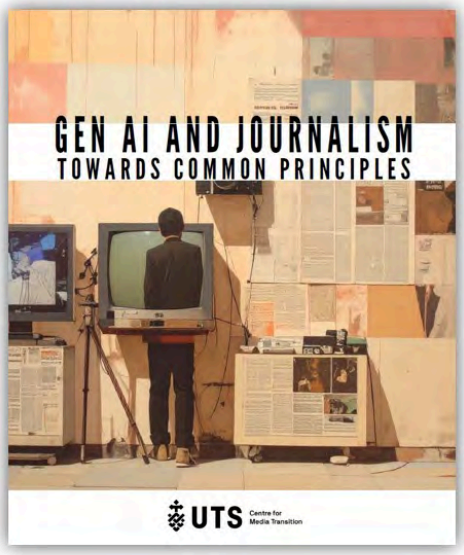
Rosa Alice
Centre Coordinator

We hope you have enjoyed reading this edition of the *Centre for Media Transition newsletter* | AI gets scary and social media bans go global - Issue 12/2026
ISSN 2981-989X

This serial can be accessed online here and through the National Library of Australia.
Please feel free to share our fortnightly newsletter with colleagues and friends!
And if this was forwarded to you, please subscribe by clicking the button below:

Subscribe

Please visit our website for more information about the Centre.



The Centre for Media Transition and UTS acknowledge the Gadigal and Guring-gai people of the Eora Nation upon whose ancestral lands our university now stands.
We pay respect to the Elders both past and present, acknowledging them as the traditional custodians of knowledge for this land.