QUESTION 1:

- In your *Digital Forensics and Incident Response Year in Review Report for 2023*, you revealed that multi-factor authentication "isn't stopping business email compromises". We continue to see social media accounts hacked and overtaken, coming at real financial and future-earnings cost to Australian businesses.
    - **What should social media companies be doing to support Australian businesses, sole traders, and users to combat these kinds of cyber attacks?**

RESPONSE 1:

Business email compromise attacks and incidents involving the take-over of social media accounts typically use different methodologies, and so there is minimal overlap in that sense. However, users can make themselves more exposed to each type of threat due to similar lapses in account security. Multifactor authentication (MFA) is one of the most effective ways to protect your accounts, and CyberCX has advises in our report the adoption of the strongest MFA possible, such as those MFA solutions aligned with phishing resistant authentication standards namely FIDO2 (Fast IDentity Online 2). Social media account takeovers most often happen due to users not having MFA in the first place either because they have not activated this option or because it has not been provided by social media companies. MFA is better than single factor authentication (only password authentication) while FIDO2 aligned solutions are phishing resistant and better than standard MFA. Social media companies can enforce MFA for business users while encouraging all users to adopt more secure practices. But ultimately this is a shared responsibility and users, especially business users need to understand the risks of losing access to their accounts, the potential for misuse and repercussions, and adopt appropriate security measures.

QUESTION 2:

- Much has been said about the impact of AI image generation, particularly in relation to child sexual abuse material, sexual content, and deepfakes. But, as you pointed out in December of last year, voice cloning opens an enormous Pandora's box – particularly when it comes to voice-based scams or "Vishing".
    - **Are social media companies doing enough to identify, prevent, and remove this kind of vocal deepfake material?**
    - **What can be done to counter this emerging threat?**

RESPONSE 2:

Preventing the proliferation of AI-generated content on social media will be practically impossible. What should be the focus then is social media platforms clearly identifying or 'tagging' content as AI-altered or generated. This can be supported by allowing users to report AI-generated content that they have encountered (or created), but platforms should

also be investing in research and innovating to mobilise AI tools to identify relevant content at scale. Use of such tools will be most important when it comes to identifying explicit, abhorrent or otherwise illegal AI-generated material that needs to be removed. For example, as software becomes more sophisticated the community will have an increasing right to expect that social media platforms use AI tools to rapidly identify child sexual abuse material and, at minimum, remove it from the platform while simultaneously notifying law enforcement.

CyberCX accepts that AI-augmented communications, especially live communications over a voice or video call will be especially difficult for social media platforms to police, particularly where they have chosen to facilitate encrypted communications as Meta has done. For these kinds of 'live' instances of AI-assisted deception, the greater opportunity for detection and prevention likely lies with the manufacturers of digital devices as well as third party software developers to provide detection software at the point-of-receipt.

Government efforts should therefore focus on supporting research that will make such detection tools more mature and more widely available. Meanwhile regulation that embeds secure-by-design principles into AI tools available to consumers should contribute to preventing users from generating illegal material in the first place.

QUESTION 3:

- Earlier this year, the eSafety Commissioner expressed concerns about violent extremists weaponising technologies like live streaming, algorithms and recommender systems to promote harmful material.
    - **Based on your research and experience, do you agree with this assertion?**

RESPONSE 3:

Yes, violent groups have long leveraged online means to assist with propagating harassment and violent messaging; spreading terror; as well as radicalising and enlisting potential recruits. We should expect these groups, including terrorist organisations and other maliciously motivated actors, to continue to be early adopters of new technologies as means to broadcast and carry out their activities. Indeed, the Director-General of Security Mike Burgess, recently notes that the Australian Security Intelligence Organisation assessed that AI was expediting the radicalization process of extremist and violent groups. Governments will therefore need to be as agile and innovative as possible in finding new ways to regulate digital devices and the online environment to curtail the use of social media and other online spaces for this purpose. Close consultation with technology industry firms,

through groupings like the Executive Cyber Council, will help the government identify where opportunities for new interventions may exist.

QUESTION 4.1:

- In August of this year, CyberCX Intelligence discovered a network of at least 5,000 inauthentic X accounts which were being used to fuel divisive arguments in Australia and in likeminded western democracies. You identified they were controlled by a Chinese language AI model, linked to a Chinese AI company and the prestigious Tsinghua University. This kind of Russian-style of polarisation is a shift from just pushing CCP propaganda.
    - **Was this just the tip of the iceberg when it comes to the Chinese Government using social media to sow social discord?**

RESPONSE 4.1:

Agencies of the People's Republic of China (PRC) and their affiliates have long sought to use social media as a means to undertake covert information campaigns against foreign states and communities, both domestic and foreign. That said, CyberCX did not publicly attribute the Green Cicada Network as being a state entity or otherwise state-controlled, even though its activities appeared to be broadly aligned with previously state-sanctioned information operations. Two details were especially noteworthy about the particular instances observed by CyberCX. One was the usage of AI to recreate, distort, and amplify divisive content on the X platform on a scale and speed rarely observed. The Green Cicada Network is one of the largest networks of inauthentic activity publicly exposed to date (measured by number of accounts) and may be the first significant China-related information operation to use generative AI as a core element of operations. Second was the way in which this capability was designed to sow discord by amplifying division about hot-button political issues. While China-nexus propaganda has often focused on amplifying narratives favourable to the CCP (or silencing narratives unfavourable to the CCP's interests) this network took an approach more closely aligned to strategies utilised by Russian state or state-backed information operations. Researchers have observed this type of strategy from PRC-nexus information operations increasing since around 2019.

QUESTION 4.2:

- **Can you provide other examples of tools, platforms, approaches and campaigns being used by the CCP, the Iranian Government, the Russian Government or their allies?**

RESPONSE 4.2:

CyberCX maintains a vigilant watching brief of the methodologies and technical capabilities being used by malicious actors from across the world. Our Green Cicada report was publicly released because we determined that the scale and speed of the operations needed to be

publicised and responded to in order to prevent the operations growing more mature and effective. Other investigations into state-linked cyber and cyber-enabled campaigns, that we assess have been designed to interfere with public attitudes, include:

- In 2023, possible **Russia-backed actors** targeting of Australian private sector and government organisations with cyber attacks designed to cause operational disruption and garner publicity for Russia, following its invasion of Ukraine. The targeting involved significant distribution-of-denial attacks which were threatened in advance and claimed after the fact on Telegram. The targeting was claimed by a group purporting to be a grassroots "hacktivist" collective, called Anonymous Sudan. However, CyberCX judged that it was unlikely that Anonymous Sudan was an authentic hacktivist group, and that there was a real-chance the activity was affiliated with the Russian state. See: cybercx.com.au/blog/a-bear-in-wolfs-clothing/
- In 2022, CyberCX investigated a "hack and leak" attack against the Nauru Police Force, which included alleged sensitive information relating to Australia's offshore processing of asylum seekers. The leak was published less than three weeks before the Australian Federal Election. While we did not definitively attribute this activity to a particular state, we assessed that it was intended to influence Australian political discourse and judged that it may have been done by a **sock puppet** rather than an authentic non-state hacktivist group, as claimed at the time. See: cybercx.com.au/blog/a-question-of-timing-nauru/
- In 2023, we observed at least one Australian victim of an **Iran-linked** hacking group, which uses the persona CyberAv3ngers, and which targeted Israeli-manufactured water pump technology used by organisations globally and defaced it. CyberAv3ngers shares claims about its attacks on its Telegram channel. We assessed the objective of this activity was to generate publicity and create fear about the security and safety of critical infrastructure. This activity has also been reported on by the US government: IRGC-Affiliated Cyber Actors Exploit PLCs in Multiple Sectors, Including U.S. Water and Wastewater Systems Facilities | CISA

QUESTION 4.3:

- **What should social media companies be doing to combat this kind of malicious activity?**

RESPONSE 4.3:

The above examples show the blurring of lines between cyber and information operations. They underscore the importance for social media companies to:

- label known foreign government or foreign government linked accounts, including state-owned media accounts (which are often used to amplify or share information operations); and
- take action against accounts that engage in or promote criminal activities, including hacking. Even authentic hacktivist groups are cyber criminals (as opposed to legitimate whistleblowers).

CyberCX also urges technology companies to protect their users by taking proactive steps to prevent their platforms from being exploited by malicious uses of AI. For social media companies in particular, this should mean directing sufficient resources to effectively perform the following functions:

- Identify and detect patterns of behaviour by users and groups of users that show the features of a coordinated information operation (as opposed to organic user behaviour).
- Intervene against accounts and groups of accounts that being mobilised as part of an information operation, up to suspending the account.
- Rapidly share evidence of information operations taking place on their platform with relevant law enforcement or intelligence agencies, including information that assists with attribution of the instigators of information operation.
- Collaborate with independent researchers to develop better shared understanding of the evolving tactics, techniques and procedures used by malicious actors on their platforms.
- Allow users to report suspected instances of malicious conduct or suspicious material.

QUESTION 5.1:

Your release states that the Green Cicada system is designed, at least in part, to 'launder' politically divisive narratives by rewording organic content as new posts and replies and to amplify organic divisive content on X through engagement. What kind of divisive content does it seek to amplify in Australia?

RESPONSE 5.1:

We observed Green Cicada engaging on issues that were already trending in Australian political discourse, such as allegations about CFMEU corruption, immigration policy, housing, and Israel-Hamas protests.

QUESTION 5.2:

Is this something that would appear to be real, or organic, to a general user of X?

RESPONSE 5.2:

It is difficult to say, as some content that is immediately recognised as inauthentic to one user may be indistinguishable from authentic content to another. Given that CyberCX observed the Green Cicada Network 'learning' or improving in its output over time, this case study would indicate that disinformation campaigns supported by AI tools will increasingly be able to assist groups to create more convincing fake accounts and false content.

QUESTION 5.3:

How could this kind of system be used to conduct harmful activities in future, including in elections? Will it become harder to detect?

RESPONSE 5.3:

Noting difficulties associated with gauging the actual impact of disinformation on individual's decision making and society-wide outcomes such as election results, it is difficult to say with certainty the extent to which malicious actors will choose to mobilise these tools into the future. More research is required to understand the impact and effectiveness of information operations on Australians so we can better anticipate which particular attack methodologies will be proliferated.

However, we can confidently state that without intervention by governments and better collaboration between digital platforms, the sider techn

ology industry, researchers and government, that there is no doubt that generative AI technology will continue to improve and therefore become harder to detect and disrupt from causing harm when used by malicious actors.