ADMM Cybersecurity and Information Centre of Excellence

UPDATE ON THE CYBER DOMAIN

Issue 6/24 (June)

The Tactics, Techniques, and Procedures of AI-Driven Disinformation and Misinformation Campaigns

INTRODUCTION

1. The World Economic Forum's Global Risks Report 2024 ranks misinformation[1] and disinformation[2] as the number one threat the world faces in the next two years. The proliferation of AI has exacerbated this issue, with malicious actors exploiting generative AI and deepfake technology to manipulate public opinion and deceive individuals on a massive scale. This report will focus on how AI can augment threat actors in their Tactics, Techniques, and Procedures (TTPs) to disseminate disinformation and misinformation, and also offer countermeasures to disrupt and mitigate the impact of disinformation and misinformation campaigns.

HOW AI IS USED IN TACTICS

2. Tactics refer to the specific goals attackers are trying to achieve through their actions. The examples below show how AI are augmenting threat actors in their campaigns:

a. **Undermining trust in institutions**. AI technologies, such as deepfakes, can create highly realistic and convincing content that is difficult to distinguish from genuine information. This realism makes it easier for malicious actors to deceive the public, as fabricated content can appear credible and authoritative. For

[1] Misinformation – false or inaccurate information, especially that which is deliberately intended to deceive.
[2] Disinformation – false information which is intended to mislead, especially propaganda issued by a government organisation to a rival power or media.

example, on 16 Mar 2022, a group of hackers intercepted one of the Ukrainian live broadcasts with AI-fabricated content, showing Ukraine's President Volodymyr Zelenskyy calling for his soldiers to surrender to the Russians. This disinformation campaign aimed to erode the morale of Ukrainian troops and civilians by creating a deceptive image of a high-level surrender. The goal was to weaken resistance against the invading forces and disrupt the unity and resolve within the Ukrainian defence apparatus. During critical times, such as conflicts or crises, people are more susceptible to psychological manipulation due to heightened emotions and stress. AI-driven misinformation that exploits these vulnerabilities can exacerbate such fears, undermining public confidence in institutions and leaders.

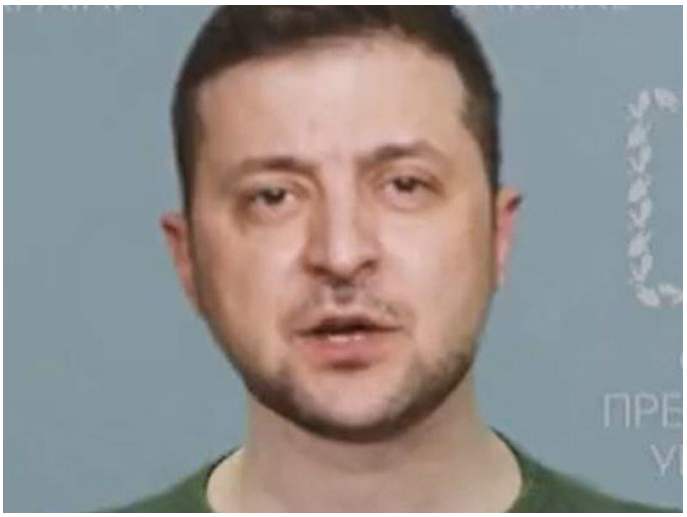

Fig 1: Zoomed-in picture of the deepfaked Zelenskyy (taken from Sky News)

b. **Influencing Political Outcomes**. AI-driven misinformation plays a significant role in influencing political outcomes by manipulating public opinion, shaping narratives, and undermining the democratic process. For example, as Moldova gears up for its presidential election and a referendum on joining the EU this year, the country has seen an increase in the amount of misinformation being circulated. Various deepfakes of pro-EU President Maia Sandu have been circulating around since 2023. Examples include a video of her announcing her resignation, as well as a video of her instituting a ban (that would have been deeply unpopular) on the consumption of a well-known brand of tea in Moldova. Moldovan Facebook pages have also seen an uptick in AI-generated pro-Kremlin content, trying to sway public opinion away from closer ties with the EU and back toward Russia.

HOW AI IS USED IN TECHNIQUES

3. Techniques are the different means in which cyber attackers can achieve their objective. Below are some examples of how AI is incorporated into techniques to assist in the attackers' campaigns:

a. **Deepfake Technology**. AI algorithms can be employed in deepfake technology to create highly realistic forged videos and audio recordings, often featuring individuals saying or doing things they never did. These manipulated media can deceive viewers and spread false information at an alarming rate. For example, in Jan 2024, there was a fake, AI-generated audio message of Joe Biden attempting to dissuade people from voting in the New Hampshire primaries. This exemplifies the potential of deepfake technology in undermining democratic processes and manipulating public perception.

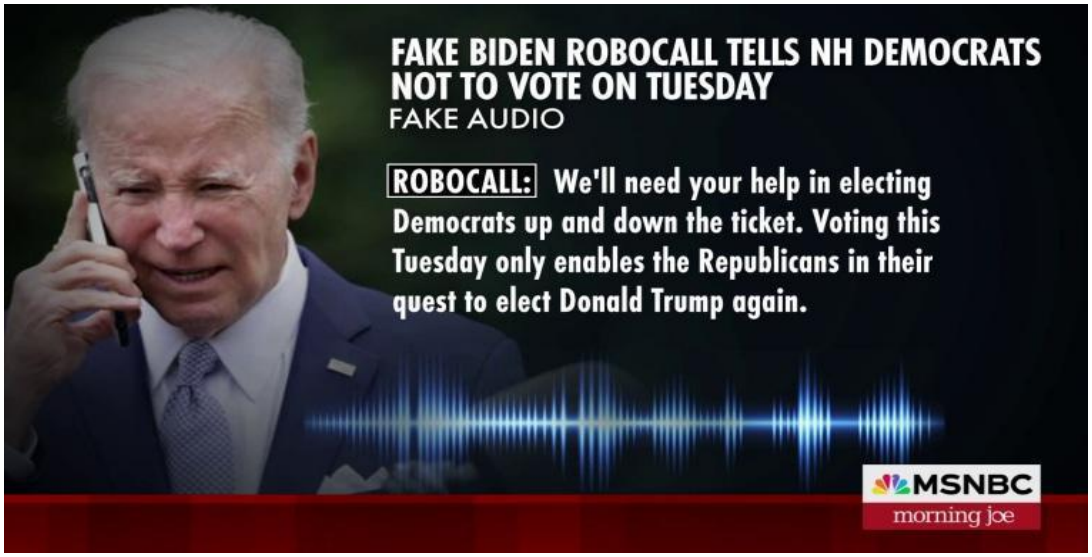

Fig 2: Fake Biden Robocall (Source: MSNBC)

b. **Automated Bot Networks to Amplify Disinformation and Misinformation**. Malicious actors can deploy automated bot networks on social media platforms to rapidly disseminate false information and manipulate online discussions. Platforms like Facebook, Twitter, and YouTube prioritise content that garners high engagement—likes, shares, and comments. By artificially boosting these metrics through coordinated activity from fake accounts, bots, and trolls, disinformation and misinformation can be made to appear popular and relevant, thus increasing its visibility. A study done analysing tweets regarding the US 2010 midterm elections details several examples of how bots are used to create illusions of virality and widespread consensus. This illusion can further influence public opinion and media coverage.

c. **Social Engineering Tactics Using AI**. AI has been exploited to personalise and tailor deceptive messages for specific individuals or groups. By analysing vast amounts of data, AI algorithms can craft convincing narratives that exploit psychological vulnerabilities, increasing the effectiveness of influence campaigns. For instance, a synthetic video surfaced on social media in May 2024, purportedly recounting the narrative of an internet troll farm in Kyiv instructed to "do everything to prevent Donald Trump from winning the elections". This video was strategically disseminated to American audiences supporting President Biden, aiming to sow seeds of doubt and discord among his supporters. Such manipulative tactics underscore the insidious nature of AI-powered social engineering.

HOW AI IS USED IN PROCEDURES

4. AI-driven disinformation campaigns often involve a series of structured and strategic steps. The Centre for Security and Emerging Technology (CSET) has identified key stages of disinformation campaigns, illustrated in a framework termed the RICHDATA pyramid. The pyramid organises and analyses the techniques and methodologies employed in these campaigns.

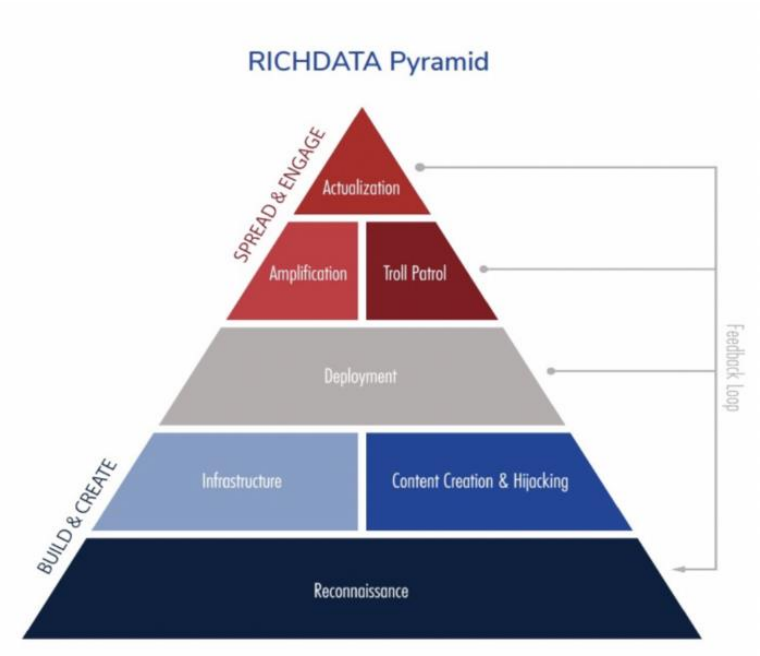

Fig 3: RICHDATA framework (Source: CSET)

5. There are three main phases of a disinformation campaign:

a. **Research and Reconnaissance to Identify Vulnerabilities and Targets**. This initial phase involves gathering intelligence on individuals, organisations, or systems that can be exploited to spread misinformation. The simplest of these are phishing campaigns. With the large amount of personal information available to threat actors via social media, AI tools such as WormGPT can be used to analyse the target individual to create a profile. This profile can then be used to create phishing content that is highly personalised to each target. For example, scammers first collected publicly available video footage related to a multinational company located in Hong Kong. This footage was then used to create deepfakes that were used to specifically target a finance employee and trick him into releasing US$25 million. The employee was led to believe that he had been in a video call with his colleagues, when in actual fact they were merely deepfakes that both looked and sounded like them.

b. **Development and Deployment of AI Disinformation Campaigns**. After identifying targets and vulnerabilities, adversaries develop sophisticated AI-driven disinformation campaigns. These campaigns create and disseminate false narratives through various channels, including social media, news websites, and messaging platforms, to achieve specific objectives such as sowing discord, undermining trust, or influencing public opinion.

c. **Iterative Adaptation of Tactics Based on Response and Feedback.** As disinformation campaigns unfold, malicious actors continuously monitor their effectiveness and adapt their tactics based on the response and feedback received. They may adjust the content, timing, or dissemination strategies to maximise impact and evade detection by security measures or content moderation efforts. For example, over the years, Russia has continuously adapted its disinformation tactics in Ukraine. Preceding the invasion of Crimea, Russia focused on spreading false narratives about Ukraine's actions and legitimacy using content mainly created by people working in "troll farms". Russia has since adapted its disinformation strategies to include new technologies: this report details in an earlier section an example where deepfakes were used to create increasingly believable misinformation targeted at Ukraine as part of the ongoing Russia-Ukraine war. This adaptability ensures that as old tactics become less effective, new ones can take their place to maintain influence and confusion.

DETECTION AND COUNTERMEASURES

5. Detecting and countering AI-driven disinformation and misinformation require a multifaceted approach involving technological, societal, and policy-based strategies. By understanding the TTPs employed by malicious actors, more effective detection and countermeasure strategies can be developed. Such solutions should be designed to identify and neutralise disinformation and misinformation at various stages of its lifecycle, from inception to dissemination.

a. **Advanced AI-Powered Detection to Identify False or Misleading Content**. Sophisticated AI algorithms can be employed to detect and analyse patterns indicative of deepfakes across various online platforms. These algorithms utilise machine learning techniques to continuously improve their accuracy and effectiveness in identifying false or misleading content. By analysing linguistic cues, metadata, and contextual information, AI-powered detection systems can flag suspicious content for further review by human moderators or automated verification processes. For example, one approach involves training an AI to detect the frequency of blinking in deepfake videos. This is based on the observation that many training datasets lack sufficient images of people with their eyes closed, leading to unnatural or infrequent blinking in the generated content. Other detection methods include examining the head or facial movements, which can be overly smooth or unnatural in deepfakes.

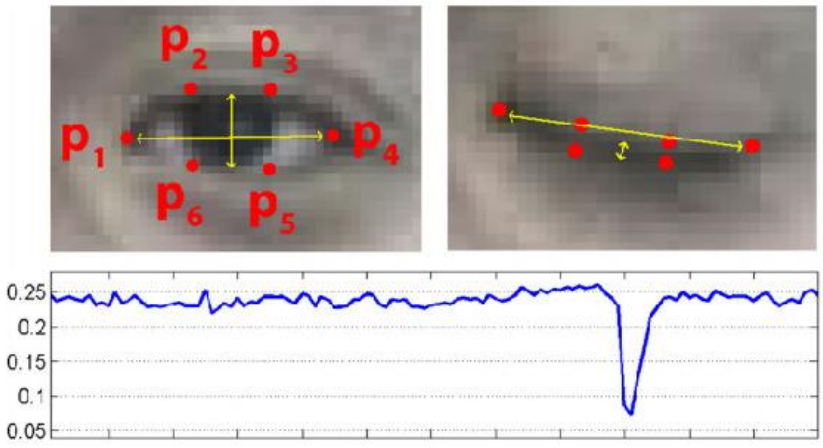

Fig 4: AI algorithm analysing data points to detect deepfake videos (Source: IEEE)

b. **Collaborations between Governments and Social Media Platforms**. By pooling resources, expertise, and data, these stakeholders can develop robust frameworks for verifying the accuracy and authenticity of information shared online. This collaborative approach may involve sharing insights, best practices, and technological tools for fact-checking, content moderation, and debunking false

narratives. For example, at the Munich Security Conference 2024, major technology companies including Meta, OpenAI, TikTok, Microsoft, and Amazon signed a pact to voluntarily adopt precautions to prevent tools from being used to disrupt democratic elections around the world. By combining the technological capabilities of private companies with the regulatory oversight and policy-making power of governments, these partnerships can create a comprehensive defence mechanism against the spread of false information.

c. **Limiting access to data by threat actors**. Producing a sophisticated Machine Learning (ML) system requires large amounts of data. Currently, many platforms such as Reddit or X (formerly known as Twitter) have their content available for access through Application Programming Interfaces (APIs), which makes it easy to scrape for this data by way of a wide variety of open-source tools. Most platforms also do not have policies regulating who they sell user data to. In this way, threat actors are able to obtain the data they need to train their models. To combat this, platforms should make data more difficult to access via their APIs and adopt policies to regulate the sale of data to unverified buyers.

d. **Creating standards for the detection and labelling of AI content.** Much of the strength of AI-created disinformation is its ability to masquerade as content created by humans. A universal way to tag and label AI-generated content would allow humans to always know when they are engaging with AI content and increase transparency on platforms. Several companies have joined together to form the Coalition for Content Provenance and Authenticity (C2PA), which aims to counter the increasing prevalence of misinformation by creating technical standards by which the source and history of digital content can be certified. Under such a framework, it would be impossible to pass off digital content produced by AI as that produced by a human. Widespread adoption of such a framework will lend credibility to organisations that produce digital content and make it easier to distinguish legitimate content.

CONCLUSION

6. In conclusion, the proliferation of AI-driven disinformation and misinformation poses a significant threat to societies worldwide, as evidenced by its manipulation of public opinion and erosion of trust in institutions. Through an understanding of the TTPs employed by malicious actors, it becomes apparent that proactive detection and

countermeasures are imperative to mitigate the impact of such disinformation and misinformation campaigns.

Contact Details

All reports can be retrieved from our website at www.acice-asean.org/resource/.

For any queries and/or clarifications, please contact ACICE, at ACICE@defence.gov.sg.

Prepared by:
ADMM Cybersecurity and Information Centre of Excellence

• • • •

REFERENCES

1. MITRE ATT&CK and DNS – Infoblox

   https://blogs.infoblox.com/security/mitre-attck-and-dns/

2. Generative AI and Custom Disinformation – Wired
   https://www.wired.com/story/generative-ai-custom-disinformation/

3. AI and the Future of Disinformation Campaigns - Georgetown CSET
   https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/

4. Ukraine war: Deepfake video of Zelenskyy telling Ukrainians to 'lay down arms' debunked - Sky News
   https://news.sky.com/story/ukraine-war-deepfake-video-of-zelenskyy-telling-ukrainians-to-lay-down-arms-debunked-12567789

5. Moldova fights to free itself from Russia's AI-powered disinformation machine – Politico
   https://www.politico.eu/article/moldova-fights-free-from-russia-ai-power-disinformation-machine-maia-sandu/

6. Fake Biden Robocall and AI in Elections – MSNBC
   https://www.msnbc.com/opinion/msnbc-opinion/fake-biden-robocall-ai-elections-rcna140570

7. Detecting and Tracking Political Abuse in Social Media - J. Ratkiewicz, M. D. Conover, M. Meiss, B. Gonc ̧alves, A. Flammini, F. Menczer
   https://ojs.aaai.org/index.php/ICWSM/article/view/14127/13976

8. The golden age of scammers: AI-powered phishing - Sinch Mailgun
   https://www.mailgun.com/blog/email/ai-phishing/

9. Finance worker pays out $25 million after video call with deepfake 'chief financial officer' – CNN
   https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

10. DeepVision: Deepfakes Detection Using Human Eye - Semantic Scholar
    https://www.semanticscholar.org/paper/DeepVision%3A-Deepfakes-Detection-Using-Human-Eye-Jung-Kim/90fb8a792c4a17e7dfde7ec290581ebcd94b6116/figure/

11. Coalition for Content Provenance and Authority (C2PA)
    https://c2pa.org