NPI 中曽根平和研究所

November 5, 2024

Threat of Generative AI in Cyber and Information Warfare: Focusing on Information Manipulation Cyberattacks

Research Project for Risk in the Information Sphere
Implementation Report

On September 18, 2024, the Research Project for Risk in the Information Sphere at Nakasone Peace Institute held a discussion based on a report by Dr. Nagasako Tomoko, Researcher at the Office of Cyber Domain Awareness, Information-technology Promotion Agency, Japan (IPA). The summary is as follows.

Dr. Nagasako presented a report entitled "Threat of Generative AI in Cyber and Information Warfare: Focusing on Information Manipulation Cyberattacks."

First, in the context of cyber and information warfare, she gave an overview of the current situation in which the "human cognitive domain" is regarded as a "battlefield," making it even easier for adversaries to engage in influence operations. In addition to the "functional destruction" and "information theft" types of conventional cyberattacks, "information manipulation" cyberattacks using disinformation and specific narratives have emerged and have been waged increasingly actively. In particular, it was pointed out that the use of images and videos created by generative AI is on the rise. Since such information and cognitive warfare is conducted even in peacetime, there are concerns about the impact not only on the countries targeted in actual contingencies but also on third countries and international public opinion. The dissemination of generative AI has raised concerns about the expansion of this type of cyberattack in that it has made information manipulation-type threats more sophisticated and easier.

It was also pointed out that the threat of "hybrid" cyberattacks, which combine information theft and information manipulation, is increasing. For example, during a visit of the Speaker of the U.S. House of Representatives Nancy Pelosi to Taiwan, a digital signage system was hacked, an act believed to be from China, and messages warning against U.S. involvement in Taiwan and stressing "one China" were displayed on the signage system. It is clear that today an integrated approach to cybersecurity, information warfare, and cognitive warfare is required. In addition, regarding trends in information and cognitive warfare in Asia, while Russia was actively attacking before the 2020s, China has become more active in recent years, creating fake media on social media and in other formats and spreading disinformation through a combination of these formats and platforms.

Next, the threat of generative AI was explained. First, Dr. Nagasako pointed out that the increase in the number of generative AI services has created a situation in which various types of content can be easily generated in large quantities even without advanced skills. In addition, it is now possible to create more realistic content, making it possible to create and disseminate persuasive articles containing a large amount of disinformation in a short period of time. On the other hand, because there are technical limitations of generative AI, such as hallucination, it is thought that it has not yet reached the level of unmatched accuracy. As part of countermeasures, the need to improve people's literacy was noted. At present, the points to identify deepfakes include ambiguity of borders in landscape images, ambiguity in facial expressions in scenes with multiple people, and ambiguity of details in human images, such as fingers and hair.

As further examples of attacks using generative AI, she cited cases related to elections in various countries. She pointed out the quality of a disinformation video in which, for example, U.S. President Joe Biden called for cooperation in the military draft. If viewers watch it without being aware of the points for verification, they might accept it as true—believing that President Biden made the statement—due to its high quality, so it could be difficult to identify it as deepfake at first glance.

In the case of the manipulation to smear Taiwan's Democratic Progressive Party (DPP) Chairman Lai Ching-te, it has become easy to generate a large number of meme images with variations using generative AI and to create videos filled with sensational elements that can easily be spread on social media. In addition to such overseas influence operations and election interference as threats to democracy, there are increasing numbers of cases in which videos and still images generated by generative AI are used to criticize specific parties or candidates in domestic issues, such as political party disputes. It was noted that election campaigns using generative AI should be autonomous and restrained, as they could increase voters' distrust in elections and may be used by China, Russia, and other countries in their influence operations.

In addition, Dr. Nagasako cited the example of a video that falsely claimed that Ukrainian President Volodymyr Zelenskyy had issued a surrender statement and pointed out that not only will disinformation increase dramatically during a contingency, but there is also concern that the misuse of generative AI will further deepen social confusion. Other examples from Japan included deepfake images and videos of Chief Cabinet Secretary Kato Katsunobu and Prime Minister Kishida Fumio. With the exception of the case of Prime Minister Kishida's support for Ukraine, it is believed that these deepfakes were prompted by a mischievous motivation. However, it was pointed out that it is necessary to create a system to prevent the spread of deepfakes because they may be misused for influence campaigns beyond their intended purpose.

Among the actors involved in these attacks were the Russian group Doppelganger, which is reportedly using generative AI to generate large numbers of images in multiple languages; the Chinese group Storm 1376, which is tasked with creating and spreading deepfake videos and images and meme images; as well as the Spamouflage campaign. "Operation Overlord" was also discussed as an example of an attack using generative AI that could be of concern. Operation Overlord is a Russian influence campaign method that combines fake media and mass reporting of disinformation in which Russian actors spread disinformation which they themselves post while at the same time requesting fact-checking of the information. This method is concerning for the following reasons: (1)

it damages public trust by giving the impression that a large amount of disinformation is spread in cyberspace, (2) it may encourage the spread of pro-Russian narratives by disclosing the results of fact-checking, and (3) it may place a burden on the fact-checking entity and impede the verification function of the disinformation.

As countermeasures against information warfare in the AI age, Dr. Nagasako cited trends in various countries regarding AI governance, including the adoption of the AI Process at the G7 Hiroshima Summit and the establishment of the AI Safety Institutes (AISI) in the U.S. and the U.K. She also pointed out that these countermeasures are still in the process of development. She also introduced technologies such as digital watermarking, AI image recognition, originator profiles, and cyber vaccines to prevent the misuse of generative AI. She also pointed out the importance of not only ex-post-facto checking but also “prebunking” in advance of debunking to enhance the resilience of the public.

On the other hand, it was pointed out that excessive warnings of disinformation and influence operations could lead to people distrusting elections and the press, which in turn could lead to the decline of democracy. Since there is also manipulation using fact-checking, it was mentioned that both a well-balanced alert system and information dissemination together with the cultivation of public literacy to receive such information are indispensable.

In response to the above report, the Q&A session included questions and comments on “availability of recognition tools for generative AI videos,” “need for research on dissemination channels,” “techniques for detecting deepfakes other than AI,” “guardrails by generative AI service providers,” “problem of overestimation of disinformation,” and “influence of generative AI content.”

The views expressed in this work are solely those of the author and do not necessarily reflect the views of Nakasone Peace Institute.