NPI 中曽根平和研究所

December 27, 2024

Issues Concerning Cooperation between Digital Platforms (DPFs) and the Government in Countering Foreign Malign Influence (FMI)

Research Project for Risk in the Information Sphere

Implementation Report

On November 20, 2024, the Research Project for Risk in the Information Sphere at Nakasone Peace Institute held a discussion based on a report by Executive Chief Fellow Mr. Fuse Satoru at Institute for International Socio-Economic Studies, Ltd. (IISE). The summary is as follows.

Foreign Malign Influence (FMI) refers mainly to "harmful influence operations, mainly through disinformation, that attempt to guide government decision-making or public opinion in a certain direction." As a premise, it should be noted that internationally the definitions of "disinformation" and "influence operations" have not yet been clearly established. Further, in Japan, there are no clear definitions of "domestic-origin or foreign-origin," and slander or illegal information may be included in the definition of "disinformation or influence operations." In this report, we will discuss "disinformation spread by foreign powers" and "the spread of information that may not be said to be disinformation, but which is harmful propaganda or has significant social impact" on social media. Digital Platforms (DPFs) are assumed to be Google (including YouTube), Meta, X, and LY Corporation, known in Japan as "LINE Yahoo."

Importance of DPF in FMI

Cooperation with the DPF is important in combating FMI, including disinformation. The DPF has the ability to collect information about the source of attacks when FMI occurs, is capable of content moderation such as deleting such comments and accounts, and it has a massive research budget to develop technology to detect FMI. It is no exaggeration to say that the DPF holds the power of life and death in cyberspace, so what kind of cooperative relationship can the government develop with the DPF? Based on this question, we will compare the policies of Japan and the U.S.

DPF regulations regarding disinformation in Japan and the U.S.

The U.S. has a policy of strong protection of freedom of speech and expression, and government intervention is kept to a minimum, leaving it to the initiative of the DPF (Soft Law). The EU does not directly mandate the removal of harmful content (hate messages, child pornography, terrorism, illegal images, etc.), but it does mandate that the DPF explain how it deals with harmful content if requested by the government. Failure to comply will result in a fine equivalent to a maximum of 6% of worldwide sales (Hard Law). Japan, like the U.S., has a soft law that leaves it up to the DPF, but discussions have focused almost exclusively on how to remove harmful content (violence, terrorism, fraud, child pornography, etc.) and slander, and there are no measures to deal with the crucial issue of disinformation and influence operations from foreign countries.

Comparison of FMI content moderation of DPFs in Japan and the U.S.

In the U.S., Meta has a track record of removing disinformation and accounts originating from China and Russia. Google and X have also detected and deleted disinformation from these countries. In the U.S., DFPs have advanced capabilities in terms of sensitivity, scale of response, and speed. In effect, the DFPs are seen as being in charge of countermeasures against disinformation.

In Japan, there are cases in which Yahoo! Japan has removed harmful information (violence, fraud, slander, damage information, etc.) from the comments section of Yahoo! News. However, there is a tendency to avoid being the main decision maker for deleting information. Once the decisions of fact-checking organizations and governments are made public, DPFs will remain dependent on the decisions, leaving them with the issue of insufficient capacity to understand and detect influence operation from overseas.

When it comes to seeking cooperation from foreign-affiliated DPFs operating in the Japanese market, the situation is not promising. For global companies, it may be difficult from an efficiency standpoint to respond individually to each country's circumstances and demands. As a result, in Japan, too, consideration regarding ways to draw cooperation from foreign-affiliated DPFs by using hard law-like regulations rather than solely relying on their initiative is beginning. For example, a recent discussion in a study group of the Ministry of Internal Affairs and Communications (MIC) has proposed requesting content moderation from foreign-affiliated DPFs, for "content for which there is a request or application from administrative agencies with jurisdiction over administrative laws and regulations" and "disinformation that is not infringing or illegal but is harmful or has a significant social impact." This proposal may be seen as even more justified when considered under the "Security Exceptions" in Article 4 of the Agreement Between the United States of America and Japan Concerning Digital Trade. This is because Article 4 can be read as it could exclude the exemption of the DPF in serious cases involving the existence or security of a state.

Recent changes and initiatives in the U.S.

Since U.S. policy can serve as a model for Japan, we will overview the U.S. approach. The U.S. government has given wide discretion to DPFs and prioritized their autonomy based on the logic that "the DPF is not a publisher and merely distributes the opinions of users." However, in recent years, a series of bills has been introduced, mainly by the Democratic Party, calling for DPFs to be regulated. This debate has involved a partisan divide, as Republicans strongly criticized the regulations as government censorship, as symbolized by the deletion of former President Trump's account on the platform then named Twitter (currently X).

Currently, the U.S. government is also involved in FMI, with the Cybersecurity and Infrastructure Security Agency (CISA), which is responsible for cybersecurity and protection of critical infrastructure, and the Federal Bureau of Investigation (FBI) as the main actors. CISA and the FBI detect harmful information in social media spaces and request its deletion by DPFs. However, even in such a case, the stance remains that the government provides information, but all actions are based on the independent judgment of the DPFs. The FBI has officially stated that regardless of how DPFs respond, the FBI will not take countermeasures. Therefore, in this situation, there are several points that need to be clarified, such as "Where and by what method will the U.S. government identify the significant signs?" and "the DFP will determine its response in light of its own operational guidelines, so what is the speed at which it will respond?" We believe that this information is necessary for the Japanese government to equip itself with the necessary capabilities.

Although information about the U.S. intelligence agencies' online surveillance methods, capabilities, and systems is undisclosed and not public, declassified documents can provide some insight. For example, the case concerning the 2022 U.S. midterm elections clearly suggests that the government may have grasped trends that could not be ascertained simply by looking at open-source information on social media. The National Security Agency (NSA) appears to have knowledge not only of meta-information but also of personal posts, images, videos, and more. Furthermore, the NSA has pointed out the need to intercept communications on undersea cables. It is unclear whether interception of communications is a future measure or whether it is already in operation, but it is likely to be a point of contention for Japan. One clear publicly available piece of information is an investigative report by the United States Government Accountability Office (U.S. GAO). The report lists the following capabilities of U.S. government agencies: (1) social media content analysis (detection of disinformation posted by foreign countries, keywords, location information, etc.), (2) network analysis (tracking of messages and user information on social media, diffusion trends and impact), (3) natural language processing (machine learning to detect disinformation content), and (4) synthetic content detection (deepfakes, detection and analysis of AI-generated video, audio, and text). From these documents, it can be inferred that the U.S. government is responding to FMI with the combined power of open-source information, information from platforms, and intercepted communications information.

Issues for Japan

In light of the above, we believe that there are three issues related to cooperation by the Government of Japan with DPFs in the countermeasures against FMI: (1) the ability to identify and specify the actors, detect signs, and infer the intentions of actors; (2) automated moderation by DPF to neutralize disinformation; and (3) methods of requesting moderation to the DPF, such as prior listing of targets. On the other hand, protection of freedom of speech and expression is extremely important. It is necessary to clarify the criteria for determining the harmfulness and situations (e.g., national elections, discussions on constitutional revision, Taiwan contingency, etc.) judged to be "non-negotiable as a nation," and to hold discussions to promote mutual understanding between the public and DPF.

Based on the report, in the Q&A session with the Research Project members, questions and comments were exchanged on "Japan's assessment of the DPFs," "synergistic effects of content moderation regulations on harmful information and other digital platform regulations," "the applicability of securitization in Japan," "the DPF's ability to identify and specify disinformation and how the government should regulate and request removal of harmful information," and "how the U.S. government has established a cooperative framework with the DPF and how cooperation between governments of the U.S. and Japan can be sought."