How to deal with influence operations in the era of generative AI

Makoto Shiono

2024 will be a year when we watch over elections in many countries while paying attention to the progress in artificial intelligence (AI), from emulating conversation through natural language generation to creating realistic-looking videos.

Following a previous article in this series by Kousuke Saitou that stated that distorted information can come not only from foreign countries, but can also be produced and sent out within Japan, this article will discuss influence operations in the country. I will touch on the characteristics of social media, including echo chambers and filter bubbles; how they can lead to polarization; emerging technologies like generative AI that have a high affinity with social media; and Japan's unique media literacy.

Polarization

People come in daily contact with sources of information, including conventional media — television and newspapers, for example — and social media — such as Facebook, Instagram and X. Such social media services are characterized by echo chambers and filter bubbles.

The concept of echo chambers, first presented by Cass Sunstein, a professor at Harvard Law School, describes a situation in which social media users tend to follow other users with ideologies and opinions similar to their own, leading them to be surrounded by information completely tailored to their beliefs. By repeatedly exposing themselves to the same type of information, users lose the opportunity to step out of their comfort zone, only to keep strengthening their own values.

Filter bubbles are defined as an environment in which social media algorithms learn from users' attributes and behavior to provide selected information that is close to their interests and preferences. The emergence of a system to provide information optimized to meet individuals' interests and preferences was prophesied in 1995 by Nicholas Negroponte, co-founder of the MIT Media Lab, who called it "the Daily Me," saying technologies would allow people to filter content so that they only access things they want to read, see and listen to.

Some experts say that in countries such as the United States, where both the people and the media are politically polarized, the characteristics of social media work to widen the divide. There are cases in which data on social media is used to grasp people's attributes and messages are continuously sent to closed information environments to which certain people belong, in order to strengthen their ideology or encourage changes in their behavior.

British consultancy firm Cambridge Analytica was accused of collecting the personal data of millions of Facebook users without their consent in the run-up to the 2016 U.S. presidential election and the 2016 Brexit referendum — a vote on whether the United Kingdom should leave the European Union. Russia is believed to have used major social media platforms to try to

influence the 2016 U.S. presidential election.

Operations resembling marketing

Methods used by Cambridge Analytica and Russia resemble those used in internet marketing — classifying users' attribute data and displaying personalized advertisements. Such methods have been used, for instance, to encourage swing voters in elections to vote for a certain candidate or fuel anxiety among minorities with social discontent to strengthen their personal political ideology.

In such a way, marketing in the business world was brought into politics, and the spread of social media enabled the gathering of massive data — including on users' personal interests and preferences, personal relationships and economic situation — and precision targeting. If such targeting is conducted by foreign countries as influence operations, it can be regarded as information warfare in the cognitive domain, meaning the information becomes subject to surveillance. From the perspective of foreign countries conducting influence operations, they are operations in the gray zone that doesn't fall in the category of either peacetime or wartime.

One of the theories that supports Russia's influence operations is "reflexive control," meaning conveying specially prepared information to an opponent to make that opponent voluntarily make a predetermined decision desired by the initiator of the action. Russia's attempt to keep hold of the areas it has occupied in Ukraine while claiming that its war against the country could end is one such example. Let's look at the effects of inputting AI-generated information into people within social media echo chambers and filter bubbles.

First of all, there is a possibility that people cannot distinguish between AI and humans. Researchers at the University of Notre Dame in the U.S. recently conducted a study using AI bots based on large language models — a type of AI developed for language understanding and text generation. They asked participants to engage in political discourse with humans and AI bots on Mastodon — a microblogging social networking platform — in three rounds, with each round lasting four days. After each round, they were asked to identify which accounts they believed were AI bots. In the experiment, the participants got it wrong 58% of the time.

Meanwhile, not all AI-generated information can be regarded as fake information. Prior to Pakistan's Feb. 8 general election, jailed former Prime Minister Imran Khan's Pakistan Tehreek-e Insaf party used generative AI to create an audio clip of Khan from text he had written from prison and passed to his lawyers, meaning the voice clone conveyed his message. In the U.S., a Democratic congressional candidate conducted a political phone campaign in December using a generative AI bot. AI can conduct customized dialogues without limit, and such AI bots can replace campaign volunteers.

There is also a case of a country's incumbent administration spreading fake information. Venezuela's state-run media last year spread messages supportive of the government by airing videos of an international English news channel that did not exist, featuring avatar newscasters

created with a private company's AI.

Japan's media environment

In Japan, there are cases of social media users spreading fake information to capture attention. Users could be motivated to take such action on platforms such as YouTube and X, where there is a monetary benefit from having higher numbers of views. Following Russia's invasion of Ukraine, fake information on the situation spread in Japanese on YouTube and X.

There were also cases of news in English or other foreign languages translated incorrectly into Japanese and posted on X. It is not clear whether such mistranslations were done intentionally or not, but some of them had been translated completely differently from the original. Balancing misinformation regulation and freedom of expression is a challenge, but we must note that false information could endanger certain rights. Japan's private television broadcasters air political news in tabloid shows as entertainment, and sometimes present anonymous social media posts as voices on the internet.

According to a nationwide survey by Japan Press Research Institute released in October, 87.6% of respondents, the highest percentage, said they see or hear news at least a few times a week on private broadcasters, followed by 74.6% on the internet, 72.1% on public broadcaster NHK, 57.5% in newspapers and 29.9% on the radio. Asked whether they care about where news is sourced from when seeing it on the internet, 47.1% said they do and 52.9% said they don't.

The results indicate that sources of news don't matter for roughly half of the surveyed people, who could be seeing investigative reporting by media and news analyses by personal blogs as having the same level of quality. This means it is possible to exert influence on a group of people with certain interests and preferences within echo chambers and filter bubbles containing those who don't care about where news is sourced from.

In February, a research team led by professor Fujio Toriumi of the University of Tokyo Graduate School of Engineering released a report titled "Anti-vaccine rabbit hole leads to political representation: the case of Twitter in Japan," studying a case in Japan of how people's behavior is affected by information on the internet. The research, using social media data in Japan, analyzes characteristics of people who became anti-vaccine during the COVID-19 pandemic and states that people newly against vaccines, whose views were prompted by the pandemic, displayed a greater affinity for conspiracy theories and spirituality.

Protecting democracy from misinformation

Cases of influence operations using generative AI are growing in foreign countries, and an environment is also being formed which could allow Japan to be affected. There is a concern that massive amounts of fake texts and videos will be created by generative AI 24/7 and at lower costs than by humans. And there is no denying that customized two-way communications or deepfake videos created by

tools like OpenAI's Sora could be misused. Such misinformation will be posted on video platforms with many young users such as TikTok.

Taiwan FactCheck Center, a private organization, detected and analyzed numerous fake videos that circulated on social media prior to the January presidential election in Taiwan. Pessimistically speaking, the Japanese people could either fall into the trap of perception bias due to misinformation and echo chambers, or come to think that no information can be trusted, believing that news contains both true and false content without checking its source. Such a situation could lead to people's declining trust in the government and political systems.

We should be alert to the possibility of foreign countries conducting operations with political intentions against people who have lost trust in their own government. Democracy and elections are built on people's decisions made based on trustworthy facts. An environment in which people cannot trust their own country's elections should be avoided at all costs.

In order to protect Japan's democracy, it is urgently necessary for the government to regulate the use of AI in political activities, platform operators to monitor fake information and nonprofit organizations to step up fact checking systems. And above all, it is essential for the Japanese people themselves to be aware of the need to improve their information literacy and work to prevent the spread of fake information.

Profile

Director of Management & Group Head, Emerging Technologies,
Institute of Geoeconomics
Makoto Shiono
Expertise
Technology and International Politics / Energy Security / Innovation Policy
https://ihj.global/en/experts/experts-4432/