Operation Types and Implement Models of International AI Sandboxes

Huang, JenChih | Analyst, The Second Research Division, CIER

With the rapid development of artificial intelligence (AI), AI sandboxes are considered essential mechanisms to support innovation and regulate risks. Depending on the needs and resources of operators, the types and promotion models of AI sandboxes vary. This paper explores different types of AI sandboxes and their implement models to provide a reference for the future planning of AI sandboxes in Taiwan.

Academic institutions primarily combine their existing digital sandbox mechanisms with AI application verification. By providing software and hardware technology resources, they lower the cost threshold for research and avoid unintended impacts or the leakage of results. For instance, Harvard University established an AI sandbox in 2023, managed by Harvard University Information Technology (HUIT). This sandbox offers multiple platforms and large language models (LLMs) for applicants to experiment with. Its operational goal is to understand the benefits of generative AI and LLMs in educational applications and provide a starting platform for university researchers to develop AI innovations.

Tech companies like Meta and Google use AI sandboxes to test new features and tools, ensuring the safety and stability of their commercial applications.

Meta established an AI sandbox in 2023 to test new tools and features for ad delivery. Advertisers can use generative AI to create variations in copy, generate creative backgrounds, and crop displayed images. The goal is to improve ad delivery efficiency while managing potential risks for preventive adjustments.

Google launched the Music AI Sandbox in 2024, integrating its music generation model Lyria to help music creators generate professional-quality melodies. Users only need to input text prompts or upload a short tune, and the program generates music. However, whether this program violates copyright protection regulations and its impact on the music industry requires further observation and verification.

National-level AI sandboxes are primarily managed by government departments or related agencies, supported by public funds and legal guarantees. They aim to facilitate strategic national layouts in AI technology fields and assess their socio-economic security impacts.

Since 2020, Norway has been planning an AI regulatory sandbox managed by the Data Protection Authority (Datatilsynet). Its purpose is to promote ethical and responsible AI development and application from a privacy protection perspective. Participating companies must develop innovative services within the data protection framework and undergo data protection impact assessments, shortening the time from AI solution development to market launch.

The EU's Artificial Intelligence Act (AIA) emphasizes a risk-based regulatory strategy and encourages the establishment of AI sandboxes. Spain became the first EU member state to implement an AI sandbox in 2022, managed by the Agency for the Supervision of Artificial Intelligence (AESIA). The sandbox tests high-risk AI applications to ensure compliance with AIA regulations.

Singapore established the AI Verify Foundation in 2023 to coordinate the testing and verification of AI technology systems and launched a generative AI evaluation sandbox. This sandbox, combining leading global AI technology service providers, promotes the testing and evaluation of generative AI applications, providing guidelines for SMEs to adopt related applications.

AI sandboxes provide a safe, isolated environment for developing, testing, and verifying AI application models and potential impacts. Academic institutions, tech companies, and national-level AI sandboxes each have their own operation models and goals. Our country currently lacks an AI sandbox mechanism. In the future, we should establish AI sandboxes based on existing evaluation mechanisms and specific AI application needs to strengthen AI innovation capabilities and enhance socio-economic well-being.