CƠ CHẾ THỬ NGHIỆM CÓ KIỂM SOÁT ĐỐI VỚI NỀN TẢNG TRÍ TUỆ NHÂN TẠO: THAM CHIẾU KHUNG CHÍNH SÁCH QUẢN TRỊ RỦI RO AI CỦA SINGAPORE VÀ KINH NGHIỆM CHO VIỆT NAM

HOÀNG THỊ HẠNH NGUYÊN
NGUYỄN TẤN KHOA
TỪ BỈNH KHANG
TRẦN KHÁNH NGÂN
Trường Đại học Sài Gòn

Nhận bài ngày 16/6/2026. Sửa chữa xong 10/7/2026. Duyệt đăng 15/7/2026.

Abstract

This article analyzes the Regulatory Sandbox mechanism as a flexible institutional solution to address legal gaps in the wake of AI advancement under the 2025 AI Law. By benchmarking experiences from Singapore, it clarifies core features such as risk-based governance and identifies cross-sectoral coordination challenges. Consequently, key recommendations are proposed: formalizing "Exit Reports," establishing technology insurance, and aligning with international standards to foster a secure and sustainable innovation ecosystem.

Keywords: *Regulatory sandbox; artificial intelligence; Singapore.*

1. Đặt vấn đề

Sự bùng nổ của trí tuệ nhân tạo (AI) đã tạo ra một cuộc cách mạng công nghệ toàn cầu mang lại những công nghệ tiện lợi phục vụ hoạt động sản xuất, kinh doanh,... nhưng cũng đồng thời đặt ra những giới hạn trong pháp luật truyền thống. Với đặc tính tự học liên tục, trí tuệ nhân tạo (AI) vận hành dựa trên việc khai thác dữ liệu quy mô lớn trong quá trình hoạt động tiềm ẩn rủi ro xâm phạm quyền riêng tư, dữ liệu cá nhân. Nếu pháp luật không có cơ chế điều chỉnh phù hợp, sự thiếu hụt quy định hiện hành sẽ tạo ra những "khoảng trống" pháp lý đe doạ đến quyền và lợi ích hợp pháp của chủ thể dữ liệu. Thực hiện Nghị quyết số 57-NQ/TW năm 2024 của Bộ Chính trị về đột phá phát triển khoa học, công nghệ, đổi mới sáng tạo và chuyển đổi số quốc gia giai đoạn 2024 – 2030, một trong các nhiệm vụ cốt lõi là "Khẩn trương sửa đổi, bổ sung, hoàn thiện đồng bộ các quy định pháp luật về khoa học công nghệ". Trên tinh thần đó, Điều 21 Luật Trí tuệ nhân tạo 2025 đã đề cập đến "Cơ chế thử nghiệm có kiểm soát đối với trí tuệ nhân tạo" (Sandbox AI) tạo ra một hành lang pháp lý đảm bảo sự cân bằng cho đổi mới sáng tạo và bảo vệ dữ liệu cá nhân. Bài viết tập trung phân tích các đặc điểm, khung pháp lý về cơ chế thử nghiệm có kiểm soát trí tuệ nhân tạo tại Điều 21 Luật trí tuệ nhân tạo 2025, nhóm tác giả tiến hành đối chiếu kinh nghiệm với pháp luật Singapore trong vận hành mô hình khung pháp lý Sandbox AI từ đó đề ra thực trạng đề xuất giải pháp hoàn thiện cơ chế hậu kiểm và miễn trừ trách nhiệm pháp lý có điều kiện, hướng tới một môi trường vận hành AI an toàn.

2. Nội dung nghiên cứu

2.1. Định nghĩa và đặc điểm về cơ chế thử nghiệm có kiểm soát

Cơ chế thử nghiệm có kiểm soát tên tiếng Anh là "Regulatory Sandbox" hiện nay pháp luật Việt Nam chưa có định nghĩa chính thức đối với cơ chế thử nghiệm có kiểm soát đối với nền tảng trí tuệ nhân tạo nhưng về bản chất đây là cách tạo ra một môi trường thử nghiệm trực tiếp mà ở đó các cơ quan tổ chức có thể tiến hành đưa vào các sản phẩm trí tuệ nhân tạo trong một thời gian, phạm vi nhất định nhằm

Email: hthnguyen801@gmail.com.

mục đích đánh giá tác động, nhận diện rủi ro và kiểm chứng tính hiệu quả của công nghệ trước khi triển khai rộng rãi1.

Khác với các thử nghiệm thông thường, cơ chế thử nghiệm có kiểm soát đối với nền tàng trí tuệ nhân tạo là công cụ quản trị rủi ro chủ động. Cơ chế này cho phép các nhà lập pháp và doanh nghiệp nhận diện sớm các lỗ hổng trong vận hành hệ thống, đặc biệt là các nguy cơ xâm phạm dữ liệu cá nhân hoặc các sai lệch về nội dung do AI tạo ra. Đặc biệt hơn, cơ chế thử nghiệm có kiểm soát tạo ra một hành lang pháp lý an toàn cho các nhà thực nghiệm được miễn trừ trách nhiệm pháp lý có điều kiện khi tham gia nghiên cứu. Về bản chất mô hình này biến mối quan hệ của cơ quan quản lý nhà nước và doanh nghiệp từ "giám sát - tuân thủ" sang quan hệ "đối tác - thử nghiệm" (Trần, Q. T., 2026, tr. 110). Để nhận diện rõ nét cơ chế này đối với nền tảng trí tuệ nhân tạo, cần xét đến các đặc điểm đặc trưng sau:

Thứ nhất, tính giới hạn thời gian thử nghiệm (Trần Đắc Hiến và cộng sự, 2021). Đặc điểm đặc trưng của cơ chế thử nghiệm có kiểm soát là có tính giới hạn ở một thời gian xác định, giới hạn này tùy thuộc vào từng lĩnh vực, giai đoạn thử nghiệm và khả năng quản lý của các chuyên gia, thông thường diễn ra từ 3 đến 24 tháng. Khoảng thời gian này vừa đủ để các chuyên gia đưa vào công nghệ, tìm ra lỗ hổng bảo mật dữ liệu, khắc phục trước khi được cơ quan quản lý giám sát đánh giá sự an toàn.

Thứ hai, có sự tham gia quản lí, giám sát từ cơ quan nhà nước (European Institute of Public Administration, 2021). Đặc điểm cốt lõi của cơ chế thử nghiệm có kiểm soát chính là sự chuyển đổi vai trò của cơ quan nhà nước từ "người gác cổng" thụ động chuyên giám sát, phạt vi phạm sang "người đồng hành" chủ động với doanh nghiệp. Cơ quan nhà nước cùng doanh nghiệp đưa ra những giới hạn cho việc khai thác thông tin cá nhân của người dùng, nếu trí tuệ nhân tạo sử dụng dữ liệu sai mục đích hệ thống báo cáo định kỳ sẽ gửi thông báo và các đợt kiểm tra kỹ thuật trực tiếp từ cơ quan nhà nước. Bằng việc tạo ra một "phòng thí nghiệm", cơ quan nhà nước có thể nới lỏng có kiểm soát một số quy định pháp luật về bảo vệ dữ liệu cá nhân, để trí tuệ nhân tạo có thể sử dụng một lượng vừa đủ tài nguyên dữ liệu cá nhân cơ bản hay nhạy cảm của người dùng (phải có sự đồng ý của chủ thể tham gia thử nghiệm), phục vụ cho việc đào tạo, phát triển trí tuệ nhân tạo trong việc bảo vệ dữ liệu cá nhân khỏi các sự xâm phạm bất hợp pháp từ những trí tuệ nhân tạo khác. Có thể nói đây là cách dùng trí tuệ nhân tạo phòng chống trí tuệ nhân tạo để bảo vệ người dùng.

Thứ ba, tính thử nghiệm và điều chỉnh (OECD, 2023). Khác phương pháp các nhà lập pháp truyền thống vốn đòi hỏi sự hoàn thiện ngay từ khi ban hành. Cơ chế thử nghiệm có kiểm soát đối với nền tảng trí tuệ nhân tạo cho phép trí tuệ nhân tạo thực hiện theo một quy trình vận hành thực tế được thuật toán sẵn dựa trên thông tin vận hành đó nhàv lập pháp có thẻ biết được những mâu thuẫn hay lỗ hỏng trong chính sách để điều chỉnh trước khi phổ biến rộng rãi.

2.2. Quy định của pháp luật Việt Nam về cơ thế thử nghiệm có kiểm soát đối với trí tuệ nhân tạo

Trong bối cảnh trí tuệ nhân tạo phát triển, việc xây dựng cơ chế pháp lý vừa khuyến khích đổi mới sáng tạo, vừa kiểm soát rủi ro là yêu cầu cấp thiết. Sandbox – cơ chế thử nghiệm có kiểm soát – được xem là công cụ pháp lý quan trọng, cho phép doanh nghiệp triển khai công nghệ mới trong môi trường linh hoạt dưới sự giám sát của Nhà nước. Tại Việt Nam, từ cuối năm 2025, hệ thống pháp luật về cơ chế này đã được hình thành hoàn chỉnh.

Luật Trí tuệ nhân tạo ban hành ngày 10/12/2025, hiệu lực từ 01/3/2026, thiết lập khuôn khổ pháp lý quản lý AI. Điều 21 Luật này quy định cơ chế thử nghiệm có kiểm soát phải tuân theo pháp luật về khoa học, công nghệ và đổi mới sáng tạo, khẳng định tính thống nhất với Luật Khoa học và Công nghệ 2013 (sửa đổi 2023), Luật Đầu tư 2020, Luật Công nghiệp công nghệ số. Hoạt động thử nghiệm AI được coi là hoạt động nghiên cứu, phát triển công nghệ đặc thù, hưởng chính sách hỗ trợ, ưu đãi về tài chính, hạ tầng và nhân lực.

Khoản 2 Điều 21 quy định kết quả thử nghiệm là cơ sở để cơ quan nhà nước xem xét xác nhận sự

1) EU Artificial Intelligence Act. Chapter VI: Measures in Support of Innovation. Article 57: AI Regulatory Sandboxes https://artificialintelligenceact.eu/article/57/

phù hợp, miễn hoặc giảm một số nghĩa vụ pháp lý. Điều 22 Nghị định của Chính phủ số 142/2026/NĐ-CP ngày 30/4/2026 quy định chi tiết một số điều và biện pháp thi hành Luật Trí tuệ nhân tạo (Nghị định 142/2026/NĐ-CP) đã cụ thể hóa trình tự đánh giá, theo đó kết quả đạt chuẩn có thể rút ngắn thủ tục cấp phép hoặc miễn giảm nghĩa vụ tuân thủ. Sandbox trở thành cầu nối giữa thử nghiệm và áp dụng pháp luật, giảm thiểu rủi ro pháp lý cho doanh nghiệp, thúc đẩy đổi mới sáng tạo. Ông Hoàng Minh Hiếu nhấn mạnh: mục đích chính của Sandbox là gỡ bỏ rào cản pháp lý, cho phép thử nghiệm để thu thập dữ liệu chứng minh hiệu quả, thúc đẩy cải cách quy định dài hạn (Thu Giang, 2025).

Việc công nhận kết quả cơ chế thử nghiệm có kiểm soát như cơ sở để xác định sự phù hợp hoặc miễn, giảm nghĩa vụ pháp lý thể hiện giá trị pháp lý thực chất của cơ chế thử nghiệm có kiểm soát. Đây không chỉ là một môi trường thử nghiệm, mà còn là một công cụ điều chỉnh nghĩa vụ pháp lý của các chủ thể tham gia. Như vậy, Sandbox trở thành cầu nối giữa thử nghiệm công nghệ và áp dụng pháp luật, giúp giảm thiểu rủi ro pháp lý cho doanh nghiệp đồng thời khuyến khích đổi mới sáng tạo. Trong phiên làm việc ngày 27/11 của Quốc Hội, ông Hoàng Minh Hiếu - Ủy viên Thường trực Ủy ban Pháp luật và Tư pháp của Quốc hội nêu ý kiến rằng: "Mục đích chính của cơ chế thử nghiệm có kiểm soát đối với AI là gỡ bỏ rào cản pháp lý đang cản trở các mô hình AI, cho phép thử nghiệm trong phạm vi giới hạn để lấy dữ liệu chứng minh hiệu quả, từ đó thúc đẩy cải cách quy định lâu dài" (Thu Giang, 2025).

Khoản 3 Điều 21 trao quyền giám sát, tạm dừng hoặc chấm dứt thử nghiệm khi phát hiện dấu hiệu gây hại cho quyền lợi hợp pháp hoặc an ninh quốc gia, thể hiện nguyên tắc quản lý dựa trên rủi ro (risk-based approach) và nguyên tắc phòng ngừa. Quản lý dựa trên rủi ro thiết kế biện pháp pháp lý tỷ lệ thuận với mức độ rủi ro, thay vì quy định cứng nhắc đồng đều. Nguyên tắc phòng ngừa cho phép can thiệp sớm ngay cả khi chưa có bằng chứng đầy đủ, bảo vệ lợi ích công cộng trước công nghệ có tính không chắc chắn cao. Điều 25 Nghị định 142/2026/NĐ-CP bổ sung quy định về báo cáo định kỳ, xử lý sự cố, phân loại rủi ro, giúp giám sát hiệu quả.

Khoản 4 Điều 21 giao Chính phủ quy định chi tiết điều kiện, trình tự, thủ tục, thời hạn, phạm vi thử nghiệm, quyền và nghĩa vụ các bên, trách nhiệm cơ quan quản lý. Trên cơ sở đó, Nghị định 142/2026/NĐ-CP xây dựng hệ thống quy định toàn diện: nguyên tắc thực hiện (Điều 21), phân loại cấp độ rủi ro (Điều 22), thẩm quyền chấp thuận (Uỷ ban nhân dân cấp tỉnh, Bộ ngành, Bộ Công an) (Điều 23), điều kiện tham gia, hồ sơ điện tử qua Cổng Dịch vụ công quốc gia (Điều 24), bảo hiểm trách nhiệm dân sự đối với hệ thống rủi ro cao.

Tổng thể, sự kết hợp giữa Luật Trí tuệ nhân tạo 2025 và Nghị định 142/2026/NĐ-CP thiết lập khung pháp lý đồng bộ cho cơ chế thử nghiệm có kiểm soát. Sandbox không chỉ mang tính thử nghiệm mà còn có giá trị pháp lý trong việc công nhận kết quả và điều chỉnh nghĩa vụ pháp lý. Đây là bước tiến quan trọng, vừa thúc đẩy đổi mới sáng tạo, vừa bảo đảm an toàn và trật tự pháp lý.

Mặc dù Luật Trí tuệ nhân tạo 2025 và Luật Công nghiệp công nghệ số là những cột mốc quan trọng, khoảng cách từ đạo luật khung đến thực thi vi mô vẫn tồn tại nhiều thách thức kỹ thuật và pháp lý. Đánh giá thực trạng cho thấy một số khoảng trống trọng yếu cần được giải quyết ở cấp độ nghị định và thông tư hướng dẫn:

Thứ nhất là bài toán nan giải về "Trách nhiệm pháp lý giới hạn". Trong luật doanh nghiệp truyền thống, quy chế pháp nhân và trách nhiệm hữu hạn đã bảo vệ tài sản cá nhân của nhà đầu tư, giải phóng sức sáng tạo kinh doanh (Shabir Korotana, 2025). Tương tự, Section 230 của Đạo luật Chuẩn mực Truyền thông Hoa Kỳ đã bảo vệ các nền tảng khỏi trách nhiệm đối với nội dung do người dùng tạo ra (Megan Cistulli, 2025). Tuy nhiên, đối với AI, việc thử nghiệm các mô hình phức tạp tiềm ẩn nguy cơ sai lệch hoặc tự chủ vượt quyền. Khung pháp lý hiện hành của Việt Nam chưa có quy định chi tiết về giới hạn miễn trừ trách nhiệm trong môi trường Sandbox. Nếu doanh nghiệp vẫn phải đối mặt với trách nhiệm dân sự khổng lồ hoặc chế tài hành chính nặng nề cho sự cố vô ý, họ sẽ e ngại tham gia; ngược lại, miễn trừ tuyệt đối có thể trở thành vỏ bọc trốn tránh nghĩa vụ, đe dọa an toàn người tiêu dùng.

Thứ hai là sự thiếu vắng của cơ chế phối hợp liên ngành thực chất. Khi thử nghiệm thuật toán AI chẩn

đoán y khoa, doanh nghiệp phải tuân thủ đồng thời quy định của nhiều bộ, ngành (Khoa học và Công nghệ, Y tế, Công an…). Thực tiễn quản lý theo ngành dọc dễ dẫn đến đùn đẩy trách nhiệm hoặc yêu cầu tuân thủ chồng chéo, làm mất đi tính linh hoạt và nhanh chóng – ưu điểm cốt lõi của Sandbox (Phan, T. T., 2024).

Thứ ba là rào cản trong việc chuyển hóa kết quả Sandbox thành lợi thế thương mại. Việc thiếu bộ tiêu chuẩn kỹ thuật số lượng hóa để xác nhận thành công của quá trình thử nghiệm khiến việc ra khỏi Sandbox và bước vào thị trường đại chúng vẫn là quy trình xin-cho kéo dài, làm nản lòng nhà đầu tư nước ngoài (Phan, T. T., 2024).

2.3. Chiến lược và mô hình vận hành khung thử nghiệm có kiểm soát trong lĩnh vực trí tuệ nhân tạo (Sandbox AI) tại Singapore

2.3.1. Định hướng quản trị rủi ro trong trí tuệ nhân tạo

Singapore khẳng định vị thế dẫn đầu về AI thông qua chiến lược quốc gia NAIS 2019 đến NAIS 2.0 (12/2023)2, hướng tới xây dựng hệ sinh thái AI đáng tin cậy và có trách nhiệm. Quốc gia này tập trung đầu tư vào hạ tầng, nhân lực và khung quản trị (Theo baoquocte.vn, 2024). Trên cơ sở đó, cách tiếp cận của Singapore trong việc định hình hành lang pháp lý cho AI thể hiện tư duy lập pháp linh hoạt và thích ứng, ưu tiên các cơ chế điều chỉnh mềm và thử nghiệm trước khi ban hành quy định ràng buộc. Tư duy này được thể hiện thông qua các định hướng chủ đạo sau:

Thứ nhất, chuyển từ việc áp dụng các chế tài hành chính ngay từ đầu sang mô hình quản trị linh hoạt. Thay vì cố gắng bao quát mọi kịch bản công nghệ bằng một hệ thống luật định, mô hình này điều chỉnh ranh giới pháp lý theo tính chất và quy mô của từng ứng dụng AI. Việc ưu tiên sử dụng "luật mềm" (soft law) bao gồm các bộ quy tắc đạo đức, khung hướng dẫn quản trị đã tạo ra một môi trường thử nghiệm tối ưu (Jason Grant Allen et al., 2025). Điều này giúp doanh nghiệm thực hiện các đổi mởi trong tầm kiểm soát và giúp cơ quan quản lý tích lũy kinh nghiệm trước khi luật hóa chính thức.

Thứ hai, xây dựng tư duy quản trị AI lấy niềm tin làm nền tảng, qua đó thúc đẩy đổi mới sáng tạo. Trong Chiến lược AI Quốc gia 2.0 (NAIS 2.0), Singapore khẳng định quản trị không phải là rào cản, mà là bệ phóng cho AI phát triển vì lợi ích công cộng3. NAIS 2.0 thiết lập các tiêu chuẩn an toàn và khung đánh giá rủi ro thông qua sự hợp tác giữa Chính phủ, doanh nghiệp và giới nghiên cứu. Cách tiếp cận này cho phép các nhà lập pháp Singapore duy trì sự linh hoạt, giúp hệ thống pháp lý có thể thích ứng kịp thời với tốc độ biến đổi của các công nghệ như AI tạo sinh.

2.3.2. Cơ chế quản trị và vận hành Sandbox AI tại Singapore

Sandbox AI của Singapore được phân loại theo mức độ rủi ro nhằm cân bằng giữa yêu cầu thúc đẩy đổi mới sáng tạo và bảo đảm an toàn pháp lý. Chính phủ thiết lập các môi trường chuyên biệt thay vì một khung thử nghiệm chung cho mọi loại hình công nghệ. Điển hình là việc vận hành các Sandbox về dữ liệu, cho phép doanh nghiệp khai tập dữ liệu đã được ẩn danh để huấn luyện mô hình mà không vi phạm các quy định bảo mật. Đặc biệt, sự ra đời của Sandbox dành riêng cho AI tạo sinh (Generative AI Sandbox) vào đầu năm 2024 đã đtập trung giải quyết các bài toán về bản quyền và tính xác thực dữ liệu trong kỷ nguyên của các mô hình ngôn ngữ lớn.

Về cơ chế vận hành, Singapore áp dụng cơ chế quản trị dựa trên mức độ rủi ro. Hệ thống AI tham gia vào Sandbox được đối soát qua một ma trận rủi ro căn cứ trên xác suất và mức độ nghiêm trọng của tác hại với người dùng. Từ đó, doanh nghiệp phải xác lập mức độ can thiệp phù hợp của con người thông qua ba trạng thái giám sát: (Human-out-of-the-loop) Human-in-the-loop (con người trực tiếp tham gia vào quá trình ra quyết định của AI); Human-over-the-loop (con người giám sát và có can thiệp khi cần thiết); Human-out-of-the-loop (AI vận hành tự động, con người không trực tiếp tham gia quá trình ra quyết định) cho các ứng dụng có rủi ro thấp (Infocomm Media Development Authority and Personal

2) Government of Singapore (2023). *National AI Strategy 2.0 (NAIS 2.0): AI for the public good for Singapore and the world.*

3) "This requires us to steer AI for the Public Good. We must harness AI in a sustainable way to create positive impact - for new opportunities, better jobs, and safer, more meaningful connections."

Data Protection Commission Singapore, 2020). Điểm khác biệt trong cơ chế này là việc chuyển nghĩa vụ báo cáo bằng văn bản sang các bằng chứng kỹ thuật số. Sự ra đời của AI Verify Foundation (Tổ chức Kiểm định Trí tuệ Nhân tạo) được chính thức công bố bởi Cơ quan Phát triển Truyền thông Thông tin Singapore (IMDA) vào 5/2023, là bước chuyển dịch lớn trong quản trị công nghệ toàn cầu. Đây là bộ công cụ kiểm thử mã nguồn mở (open-source toolkit) đầu tiên trên thế giới có khả năng định lượng, gúp doanh nghiệp cụ thể hóa các nguyên tắc AI trừu tượng như tính minh bạch, tính công bằng, tính giải thích được. Nền tảng này đã biến AI Verify trở thành một "công cụ thực thi pháp lý kỹ thuật số", giúp doanh nghiệp chuyển đổi các tiêu chuẩn đạo đức thành các chỉ số kỹ thuật cụ thể và có thể kiểm chứng được (Infocomm Media Development Authority and AI Verify Foundation, 2023).

Cuối cùng, sau thử nghiệm, các tổ chức phải thực hiện báo cáo tác động về hành vi thực tế của thuật toán. Dựa trên kết quả này, cơ quan quản lý sẽ quyết định cho phép triển khai thương mại hóa đại trà, yêu cầu điều chỉnh thêm hoặc đình chỉ nếu AI có rủi ro an ninh dữ liệu. Toàn bộ quy trình này tạo thành một vòng lặp phản hồi liên tục, nơi dữ liệu thực từ Sandbox được sử dụng để tinh chỉnh các quy định pháp luật chính thức, đảm bảo hành lang quản trị luôn theo kịp tốc độ phát triển của công nghệ.

2.3.3. Cơ chế hậu kiểm và bảo vệ dữ liệu người dùng

Sandbox AI của Singapore cân bằng giữa khuyến khích đổi mới và bảo vệ cá nhân, không miễn trừ các trách nhiệm pháp lý về bảo mật và quyền riêng tư:

Thứ nhất, nghĩa vụ tuân thủ các nguyên tắc bảo vệ dữ liệu cá nhân. Mặc dù Sandbox cung cấp một môi trường linh hoạt, nhưng các thực thể AI vẫn phải tuân thủ nghiêm ngặt Luật Bảo vệ dữ liệu cá nhân (PDPA)4. Theo hướng dẫn của Ủy ban Bảo vệ dữ liệu cá nhân (PDPC), các tổ chức cần đảm bảo các tập dữ liệu dùng cho huấn luyện mô hình là đầy đủ, chính xác và không có các định kiến độc hại. Doanh nghiệp được khuyến khích đảm bảo tính minh bạch, thông báo rõ ràng cho cá nhân khi AI đưa ra các quyết định có tác động đáng kể5.

Thứ hai, xác lập cơ chế trách nhiệm dựa trên sự tham gia của con người. Singapore gắn liền trách nhiệm pháp lý và mức độ kiểm soát của con người đối với AI. Các tổ chức phải xác định rõ vai trò và trách nhiệm cho từng nhân sự hoặc bộ phận giám sát6. Trách nhiệm giải trình yêu cầu các tổ chức phải thiết lập các kênh phản hồi, cho phép cá nhân có quyền khiếu nại hoặc yêu cầu giải thích về các quyết định do AI đưa ra7. Với các hệ thống có rủi ro cao, Singapore khuyến khích áp dụng mô hình "con người trong vòng lặp" (human-in-the-loop) để đảm bảo con người giữ vai trò quyết định cuối cùng về mặt đạo đức và pháp lý.

Thứ ba, cơ chế bảo vệ quyền rút lui. Singapore xây dựng niềm tin thông qua việc trao quyền cho người dùng. Các tổ chức được khuyến khích cung cấp tùy chọn "rút khỏi thử nghiệm" (opt-out) cho người dùng, dựa trên mức độ rủi ro kỹ thuật. Nếu không thể cung cấp tùy chọn này, tổ chức bắt buộc phải có các cơ chế để con người rà soát trực tiếp, đánh giá lại sai sót của AI8. Việc tuân thủ Khung quản trị mẫu trong Sandbox giúp giảm thiểu các chế tài hành chính nếu xảy ra sự cố ngoài ý muốn.

Tổng kết lại, cơ chế Sandbox tại Singapore là hình mẫu về quản trị thích ứng. Bằng cách kết hợp chiến lược NAIS 2.0 và Khung quản trị mẫu của IMDA/PDPC, Singapore đã cân bằng giữa đổi mới sáng tạo và bảo vệ quyền lợi con người. Mô hình này lấy niềm tin làm nền tảng, định lượng trách nhiệm bằng bằng chứng thực tế, và trao quyền cho cá nhân qua cơ chế rút lui (opt-out) hoặc rà soát quyết định. Đây là kinh nghiệm quý giá về một hệ sinh thái quản trị linh hoạt, nơi luật pháp không còn là rào cản mà là bệ phóng cho các công nghệ tiên phong vì lợi ích cộng đồng.

2.4. Một số đề xuất cho Việt Nam

Mặc dù Luật Trí tuệ nhân tạo 2025 đã thiết lập nền tảng pháp lý cơ bản cho cơ chế thử nghiệm có

4) Personal Data Protection Act 2012 (Singapore), s 11.
5) MDA and PDPC, *Model AI Governance Framework, p. 55*.
6) MDA and PDPC, *Model AI Governance Framework, p. 22*.
7) MDA and PDPC, *Model AI Governance Framework, p. 57*.
8) MDA and PDPC, *Model AI Governance Framework, p. 56*.

kiểm soát, việc chuyển hóa các quy phạm khung thành thực tiễn vận hành đòi hỏi các giải pháp cấu trúc và kỹ thuật pháp lý chuyên sâu. Để Sandbox AI tại Việt Nam là không gian kiến tạo giá trị thay vì chỉ là vùng đệm miễn trừ rủi ro, hệ thống giải pháp cần được định hình đồng bộ trên các phương diện sau:

2.4.1. Thiết lập Khung quản trị AI mẫu có tính định lượng

Việt Nam cần chuyển dịch sang mô hình quản trị linh hoạt dựa trên rủi ro thông qua việc ban hành "Khung quản trị AI mẫu", học tập kinh nghiệm phối hợp giữa IMDA và PDPC của Singapore. Khung này đóng vai trò là bộ tiêu chuẩn tự nguyện giúp chuyển hóa các nguyên tắc đạo đức AI trừu tượng thành thông số kỹ thuật và yêu cầu vận hành thực tế. Nội dung cốt lõi bao gồm phân định cấu trúc quản trị nội bộ, phân cấp mô hình kiểm soát (human-in-the-loop hoặc human-over-the-loop) và quy trình quản lý dữ liệu an toàn.

Việc doanh nghiệp chứng minh tuân thủ nghiêm ngặt Khung quản trị mẫu này trong suốt thời gian vận hành Sandbox sẽ là căn cứ pháp lý thể hiện "nỗ lực cẩn trọng hợp lý". Đây là cơ sở để cơ quan quản lý nhà nước xem xét giảm nhẹ hoặc miễn trừ trách nhiệm pháp lý có điều kiện khi xảy ra sự cố kỹ thuật ngoài ý muốn.

2.4.2. Thể chế hóa giá trị pháp lý của "Báo cáo thoát" như một hộ chiếu công nghệ

Thành công thực sự của Sandbox đo bằng tỷ lệ doanh nghiệp "tốt nghiệp" và tự tin thương mại hóa sản phẩm ra thị trường. Do đó, việc thừa nhận giá trị pháp lý thực chất của "Báo cáo thoát" – hồ sơ chứng minh toàn diện mức độ an toàn, hiệu quả của công nghệ sau thử nghiệm – là mắt xích quan trọng nhất. Văn bản này phải được định hình là một "hộ chiếu công nghệ" chính thức thay vì chỉ là bản tổng kết để lưu hồ sơ.

Đối với hệ thống AI rủi ro cao sở hữu "Báo cáo thoát" đạt chuẩn, pháp luật cần đặc cách áp dụng quy trình đánh giá sự phù hợp vắn tắt. Cơ quan quản lý sẽ sử dụng chính dữ liệu thực tế từ Báo cáo thoát để cấp phép thay vì thẩm định lại từ đầu. Đột phá này giúp cắt giảm triệt để thủ tục "tiền kiểm" phức tạp, tốn kém; đảm bảo tính liên tục của chu trình "thử nghiệm thực tế – chứng nhận nhanh chóng – thương mại hóa kịp thời" và xóa bỏ rào cản "xin - cho".

2.4.3. Phân bổ rủi ro dân sự thông qua "Bảo hiểm công nghệ" và "Đồng thuận có thông tin"

Việc giới hạn trách nhiệm trong Sandbox nhằm bảo vệ quyền lợi hợp pháp của bên thứ ba chịu tác động. Việt Nam cần thiết lập cơ chế phân bổ rủi ro dân sự nhị phân:

Về tài chính: Nhà phát triển tham gia Sandbox bắt buộc phải lập Quỹ dự phòng rủi ro hoặc mua gói "Bảo hiểm trách nhiệm công nghệ" chuyên biệt để bảo đảm đền bù thiệt hại thỏa đáng cho người dùng, tránh nguy cơ đổ vỡ tài chính doanh nghiệp.

Về quyền lợi người dùng: Áp dụng nguyên tắc "Đồng thuận có thông tin", buộc doanh nghiệp phải minh bạch thuật toán, thông báo rủi ro tiềm ẩn và cung cấp quyền rút khỏi thử nghiệm cho người dùng. Thỏa thuận này cấu thành văn bản pháp lý ràng buộc về việc chấp nhận rủi ro, cụ thể hóa yêu cầu về tính giải trình của Luật Trí tuệ nhân tạo 2025 và ngăn chặn Sandbox biến thành "vùng trắng" pháp lý.

2.4.4. Định vị Sandbox như một thiết chế chiến lược thu hút FDI công nghệ cao

Để tạo ra "lợi thế cạnh tranh thể chế" và xóa bỏ rào cản bất định pháp lý đối với dòng vốn FDI vào công nghệ sâu, Việt Nam cần chủ động nội luật hóa và tích hợp các chuẩn mực kiểm định toàn cầu do Tổ chức Hợp tác và Phát triển Kinh tế (OECD) hoặc Viện Kỹ sư điện và điện tử (IEEE) đề xuất vào khung đánh giá Sandbox nội địa. Tính tương thích này sẽ biến Việt Nam thành cứ điểm an toàn, linh hoạt để các tập đoàn đa quốc gia và quỹ đầu tư mạo hiểm thiết lập trung tâm nghiên cứu và phát triển, đạt hiệu quả tối ưu khi gắn kết với các chính sách ưu đãi từ Luật Công nghiệp công nghệ số.

2.4.5. Dân chủ hóa hạ tầng đổi mới sáng tạo thông qua chia sẻ nguồn lực thực chất

Để xóa bỏ sự bất bình đẳng về dữ liệu và năng lực điện toán giữa startup nội địa với các tập đoàn xuyên quốc gia, Sandbox phải vận hành như một không gian hỗ trợ tài nguyên thực chất. Chính sách cần cho phép hệ sinh thái khởi nghiệp được quyền tiếp cận ưu tiên đối với hạ tầng siêu máy tính quốc

gia, cơ sở dữ liệu lớn của Chính phủ đã được ẩn danh, làm sạch và các mô hình ngôn ngữ lớn (LLM) dùng chung. Việc phá vỡ thế độc quyền hạ tầng này sẽ giúp doanh nghiệp công nghệ Việt Nam từng bước tự chủ về lõi thuật toán.

3. Kết luận

Sự tiến hóa của AI bộc lộ giới hạn cấu trúc của hệ thống luật pháp truyền thống, đòi hỏi chuyển dịch từ quản trị thụ động sang quản trị kiến tạo và thực nghiệm. Cơ chế thử nghiệm có kiểm soát (Sandbox) không chỉ là miễn trừ pháp lý tạm thời, mà là không gian đối thoại thể chế, nơi lợi ích đổi mới sáng tạo và bảo vệ quyền con người được hiệu chỉnh qua bằng chứng thực tiễn.

Kinh nghiệm của Singapore cho thấy quốc gia này linh hoạt với tiêu chuẩn kỹ thuật và red-teaming (thử nghiệm tấn công giả lập). Tại Việt Nam, Luật Công nghiệp công nghệ số, Luật Trí tuệ nhân tạo 2025 và Luật Thủ đô 2024 tạo nền tảng pháp lý vững chắc, định danh rõ ràng cơ chế Sandbox, phản ánh tầm nhìn chiến lược gia nhập nhóm dẫn dắt công nghệ toàn cầu. Tuy nhiên, hiệu quả phụ thuộc vào việc tháo gỡ nút thắt vi mô: ban hành nghị định giải quyết trách nhiệm pháp lý giới hạn, hiện thực hóa báo cáo thoát thành lợi thế thương mại, và tương thích tiêu chuẩn kỹ thuật nội địa với chuẩn mực quốc tế. Khi sandbox đủ dẻo dai để hấp thụ rủi ro số và đủ vững chãi để bảo vệ chuẩn mực đạo đức, nó sẽ trở thành công cụ tối ưu giúp Việt Nam quản trị AI an toàn và thu hút FDI công nghệ cao bền vững.

Tài liệu tham khảo

European Institute of Public Administration (2021). *Sandboxes for Responsible Artificial Intelligence*. https://www.eipa.eu/publications/briefing/sandboxes-for-responsible-artificial-intelligence/

Government of Singapore (2023). *National AI Strategy 2.0 (NAIS 2.0): AI for the public good for Singapore and the world.*

Infocomm Media Development Authority and AI Verify Foundation (2023). *AI Verify Testing Framework: For Traditional and Generative AI*. https://aiverifyfoundation.sg

Infocomm Media Development Authority and Personal Data Protection Commission Singapore (2020). *Model Artificial Intelligence Governance Framework (2nd ed)*.

OECD (2023). *Regulatory sandboxes in artificial intelligence* (OECD Digital Economy Papers No. 356). Paris, France: OECD Publishing. doi:10.1787/849f2a00-en

Phan, T. T. (2024, January 10). Pháp luật về sandbox đổi mới sáng tạo ở Việt Nam: Yêu cầu hoàn thiện nhằm thu hút FDI công nghệ cao. *Tạp chí Pháp lý.* https://phaply.net.vn/phap-luat-ve-sandbox-doi-moi-sang-tao-o-viet-nam-yeu-cau-hoan-thien-nham-thu-hut-fdi-cong-nghe-cao-a261507.html.

Megan Cistulli (2025). Generative AI Meets Section 230: The Future of Liability and Its Implications for Startup Innovation. *The University of Chicago Business Law Review*. https://businesslawreview.uchicago.edu/print-archive/generative-ai-meets-section-230-future-liability-and-its-implications-startup.

Shabir Korotana (2025). Decentralized autonomous organizations: adapting legal structures and proposing a new model of DAOLLP. *Capital Markets Law Journal*. https://academic.oup.com/cmlj/article/20/3/kmaf011/8249442.

Thu Giang (2025). *Xây dựng Luật AI: Đặt cơ chế thử nghiệm có kiểm soát ở vị trí trung tâm.* https://baochinhphu.vn/xay-dung-luat-ai-dat-co-che-thu-nghiem-co-kiem-soat-o-vi-tri-trung-tam-102251127181338124.htm, ngày 27/11/2025.

Theo baoquocte.vn (2024). Kinh nghiệm Singapore về quản trị trí tuệ nhân tạo (AI) và bài học cho Việt Nam. *Tạp chí Pháp luật & Phát triển*. https://phapluatphattrien.vn/kinh-nghiem-singapore-ve-quan-tri-tri-tue-nhan-tao-ai-va-bai-hoc-cho-viet-nam-d588.html

Trần, Q. T. (2026). Bản chất pháp lý và cơ chế vận hành của khung pháp lý thử nghiệm có kiểm soát: Thực trạng và hướng hoàn thiện. *Tạp chí Quản lý nhà nước*, *360*, tr. 110.

Trần Đắc Hiến, Trần Thị Thu Hà, Nguyễn Mạnh Quân, Nguyễn Lê Hằng, & Phùng Anh Tiến (2021). *Cơ chế thử nghiệm (Regulatory Sandbox): Từ lý thuyết đến thực tiễn áp dụng trên thế giới.* Bản tin Chiến lược phát triển số, 7. Cục Thông tin Khoa học và Công nghệ Quốc gia.

Jason Grant Allen et al. (2025), Governing Intelligence: Singapore's Evolving AI Governance Framework. *Cambridge Forum on AI: Law and Governance*. https://doi.org/10.1017/cfl.2024.12.

William Fry (2024). *The time to AI Act is now: A practical guide to regulatory sandboxes under the AI Act.* https://www.williamfry.com/knowledge/the-time-to-ai-act-is-now-a-practical-guide-to-regulatory-sandboxes-under-the-ai-act/.