TẠP CHÍ NGHIÊN CỨU
PHẬT HỌC
CƠ QUAN NGÔN LUẬN CỦA TRUNG ƯƠNG GIÁO HỘI PHẬT GIÁO VIỆT NAM
ISSN: 2734-9195

Quản trị AI Agent: Khoảng trống kiểm soát và bài toán trách nhiệm

ISSN: 2734-9195 10:10 23/05/2026

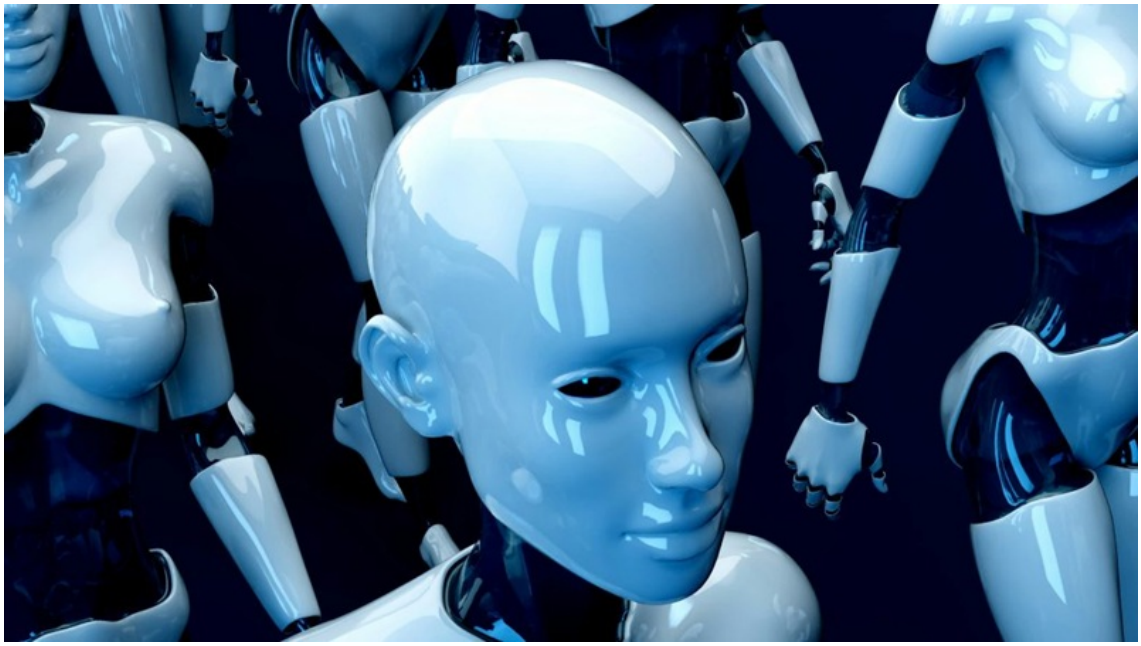
Quản trị AI không chỉ là kiểm soát hệ thống, mà còn là quá trình: Nhận diện tham vọng và giới hạn của con người, Nuôi dưỡng trí tuệ và trách nhiệm, Giữ vững sự cân bằng giữa hiệu quả và đạo đức.

Trong bối cảnh **trí tuệ nhân tạo** (AI) ngày càng hiện diện sâu rộng trong hệ thống tài chính, cảnh báo đáng chú ý vừa được đưa ra từ Cơ quan Quản lý Thận trọng Úc (APRA): năng lực quản trị và kiểm soát các AI agent (tác nhân AI), những hệ thống có khả năng tự động hành động vẫn còn nhiều khoảng trống đáng lo ngại.

Cảnh báo này xuất hiện khi các ngân hàng và quỹ hưu trí tại Úc đang gia tăng ứng dụng AI, không chỉ trong vận hành nội bộ mà còn trực tiếp tương tác với khách hàng. Điều đó đặt ra một câu hỏi lớn: **chúng ta đang kiểm soát, hay đang dần phụ thuộc vào AI?**

AI đã hiện diện nhưng sự trưởng thành chưa đồng đều

Theo APRA, trong cuộc rà soát các tổ chức tài chính lớn vào cuối năm 2025, AI đã được triển khai ở tất cả các đơn vị được khảo sát. Tuy nhiên, mức độ trưởng thành về quản lý rủi ro và khả năng vận hành an toàn lại rất khác nhau.



Ảnh: julien Tromeur

Hội đồng quản trị các tổ chức này nhìn nhận AI như một công cụ thúc đẩy năng suất và nâng cao trải nghiệm khách hàng. Nhưng nghịch lý nằm ở chỗ: **sự quan tâm chiến lược chưa đi kèm với năng lực kiểm soát rủi ro tương xứng**.

Một số hội đồng vẫn phụ thuộc nhiều vào các bản trình bày từ nhà cung cấp, thiếu sự thẩm định độc lập đối với những rủi ro cốt lõi như:

+ Hành vi khó dự đoán của mô hình AI.

+ Tác động dây chuyền khi AI gặp lỗi trong các hệ thống trọng yếu.

Đây chính là biểu hiện của **“vô minh trong nhận thức công cụ”**, khi con người sử dụng một phương tiện mạnh mẽ nhưng chưa hiểu rõ bản chất và giới hạn của nó.

Rủi ro không chỉ là công nghệ - mà là nhận thức

Một trong những sai lầm phổ biến mà APRA chỉ ra là việc nhiều tổ chức đối xử với rủi ro AI giống như rủi ro công nghệ truyền thống. Tuy nhiên, AI không đơn thuần là một công cụ tĩnh, mà là hệ thống có khả năng học hỏi, biến đổi và thậm chí “lệch chuẩn” (bias).

Các khoảng trống được xác định bao gồm:

+ Thiếu cơ chế theo dõi hành vi mô hình.

+ Quản lý thay đổi (change management) chưa đầy đủ.

+ Không có quy trình “khai tử” (decommissioning) AI rõ ràng.

+ Thiếu danh mục kiểm kê đầy đủ các hệ thống AI.

+ Không xác định rõ cá nhân chịu trách nhiệm cho từng hệ thống.

Đặc biệt, APRA nhấn mạnh vai trò con người phải tham gia vào các quyết định có rủi ro cao.

Điều này gợi nhắc đến nguyên lý căn bản trong Phật giáo: **Trí tuệ không thể được thay thế hoàn toàn bởi công cụ.**

Dù AI có thể xử lý dữ liệu nhanh hơn, nhưng chính kiến (right view) - khả năng thấy đúng bản chất sự việc vẫn là yếu tố chỉ có thể được nuôi dưỡng từ con người.

Khi AI trở thành “tác nhân”: Ai chịu trách nhiệm?

Một vấn đề nổi bật khác là sự xuất hiện của các AI agent tự động, những hệ thống có thể thực hiện hành động thay cho con người, từ xử lý hồ sơ vay vốn đến phát hiện gian lận.

Tuy nhiên, điều này đặt ra một câu hỏi căn bản về trách nhiệm và danh tính: AI hành động “thay mặt ai”? Ai chịu trách nhiệm khi quyết định sai lầm xảy ra?

Tổ chức FIDO Alliance hiện đang phát triển các tiêu chuẩn xác thực mới cho các hành động do AI khởi tạo. Họ nhấn mạnh rằng các mô hình xác thực hiện tại vốn được thiết kế cho con người, không phù hợp với các hành động được “ủy quyền” cho phần mềm.

Một số giải pháp đang được đề xuất, như: Giao thức thanh toán bằng AI của Google. Khung “ý định có thể xác minh” của Mastercard.

Đây là vấn đề liên quan đến nghiệp và trách nhiệm (karma & accountability). Một hành động dù được thực hiện bởi AI, nhưng nếu xuất phát từ ý chí và hệ thống do con người thiết kế, thì **nghiệp vẫn thuộc về con người**.

An ninh mạng trong thời đại AI: Khi rủi ro trở nên “vô hình” hơn

APRA cũng cảnh báo rằng việc triển khai AI đang làm thay đổi môi trường an ninh mạng, tạo ra các con đường tấn công mới như: Prompt injection (tấn công thông qua đầu vào), Tích hợp không an toàn giữa các hệ thống.

Ngoài ra, các hệ thống quản lý danh tính và quyền truy cập chưa kịp thích ứng với các "thực thể không phải con người" như AI agent.

Một điểm đáng lo ngại khác là: Sự phụ thuộc vào một nhà cung cấp AI duy nhất, trong khi thiếu kế hoạch thay thế hoặc rút lui.

Điều này phản ánh một dạng **"chấp chước công nghệ"**, khi con người đặt niềm tin quá lớn vào một hệ thống mà thiếu phương án dự phòng.

Trong giáo lý Phật giáo, mọi pháp đều mang tính **vô thường (anicca)**. Một hệ thống AI hôm nay có thể hiệu quả, nhưng ngày mai có thể lỗi thời hoặc rủi ro.

Do đó, tư duy linh hoạt và không phụ thuộc là nguyên tắc quan trọng trong quản trị công nghệ.

Hướng đi nào cho "đạo đức AI" tỉnh thức?

Các hướng dẫn từ Trung tâm An ninh Internet (CIS) đã bắt đầu xây dựng các tiêu chuẩn bảo mật cho AI, bao gồm: Kiểm soát dữ liệu nhạy cảm trong mô hình ngôn ngữ lớn (LLM). Quản lý truy cập của các thực thể phi con người. Bảo mật trong môi trường tương tác mạng.

Tuy nhiên, như thực tế cho thấy, khung kỹ thuật thôi là chưa đủ.

Điều cần thiết hơn là một nền tảng đạo đức và nhận thức sâu sắc, điều mà Phật giáo có thể đóng góp:

+ *Chính niệm (mindfulness):* nhận diện rõ ràng giới hạn của AI.

+ *Chính kiến (right view):* hiểu đúng bản chất công cụ và hệ quả của việc sử dụng.

+ *Trách nhiệm (karma):* không chuyển giao hoàn toàn quyết định cho máy móc.

+ *Vô chấp (non-attachment):* không phụ thuộc tuyệt đối vào công nghệ.

Cảnh báo từ APRA không chỉ là vấn đề kỹ thuật, mà là lời nhắc về một thực tế sâu xa hơn: *Công nghệ càng thông minh, con người càng cần tỉnh thức.*

AI có thể trở thành công cụ hỗ trợ đắc lực, nhưng nếu thiếu quản trị, AI cũng có thể khuếch đại sai lầm ở quy mô lớn.

Từ lăng kính Phật học, quản trị AI không chỉ là kiểm soát hệ thống, mà còn là quá trình: Nhận diện tham vọng và giới hạn của con người, Nuôi dưỡng trí tuệ và trách nhiệm, Giữ vững sự cân bằng giữa hiệu quả và đạo đức.

Khi đó, AI không còn là mối đe dọa, mà trở thành một phương tiện thiện xảo: *phục vụ con người, thay vì dẫn dắt con người.*

Tác giả: **Muhammad Zulhusni/**Chuyển ngữ và biên tập: **Uyển Nhi**

Nguồn: https://www.artificialintelligence-news.com/news/ai-agent-governance-control-gaps/