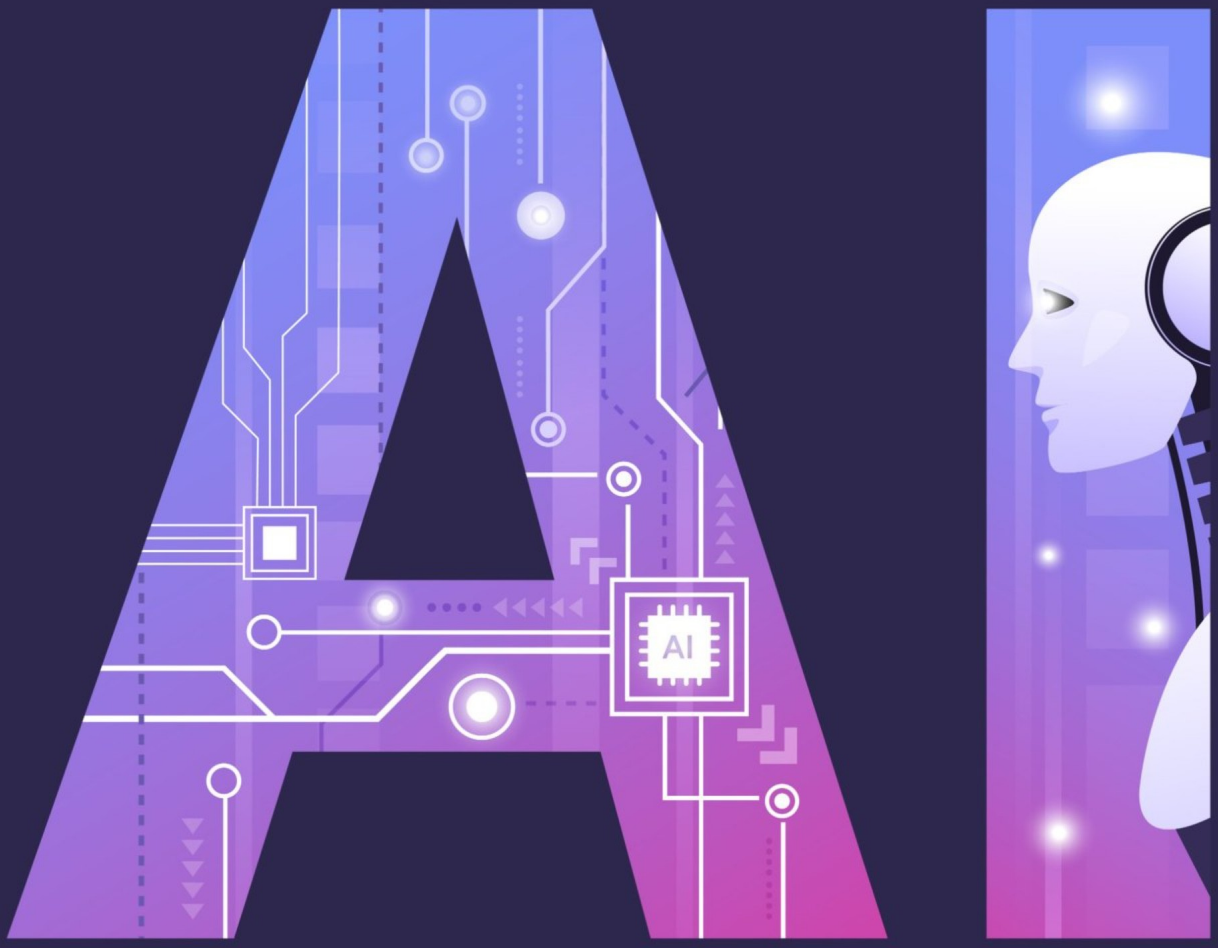
405CYSE



Giám sát An toàn tài sản số trong ứng dụng AI cho doanh nghiệp

Công ty Cổ phần An ninh mạng 405CYSE
An toàn dữ liệu và an ninh mạng cho doanh nghiệp

www.405cyse.com

Nội dung báo cáo

PHẦN 1	**Thế giới đang bước từ AI qua AI Agent**	04
	Từ "sử dụng AI" đến "AI tự vận hành"	05
	Thế giới đã bước sang kỷ nguyên tác nhân AI	07
	AI trong doanh nghiệp Việt Nam: Đã là tác nhân AI dù chưa được gọi tên	08
PHẦN 2	**Rủi ro từ tác nhân AI gia tăng**	10
	Rủi ro tăng theo mức tự chủ	11
	Hệ thống AI bị tấn công ở đâu	12
	Các luồng dữ liệu của AI	14
PHẦN 3	**Quản trị An toàn tài sản số**	16
	Quản trị An toàn tài sản số: Quản trị trước, tích hợp sau	17
	Các chuyển dịch tư duy cho lãnh đạo	19
	Mô hình "3 luồng – 4 cấp"	20
	Chuẩn mực quốc tế và nghĩa vụ pháp lý	20
	Lộ trình cho Ban lãnh đạo	21

PHẦN 1

01

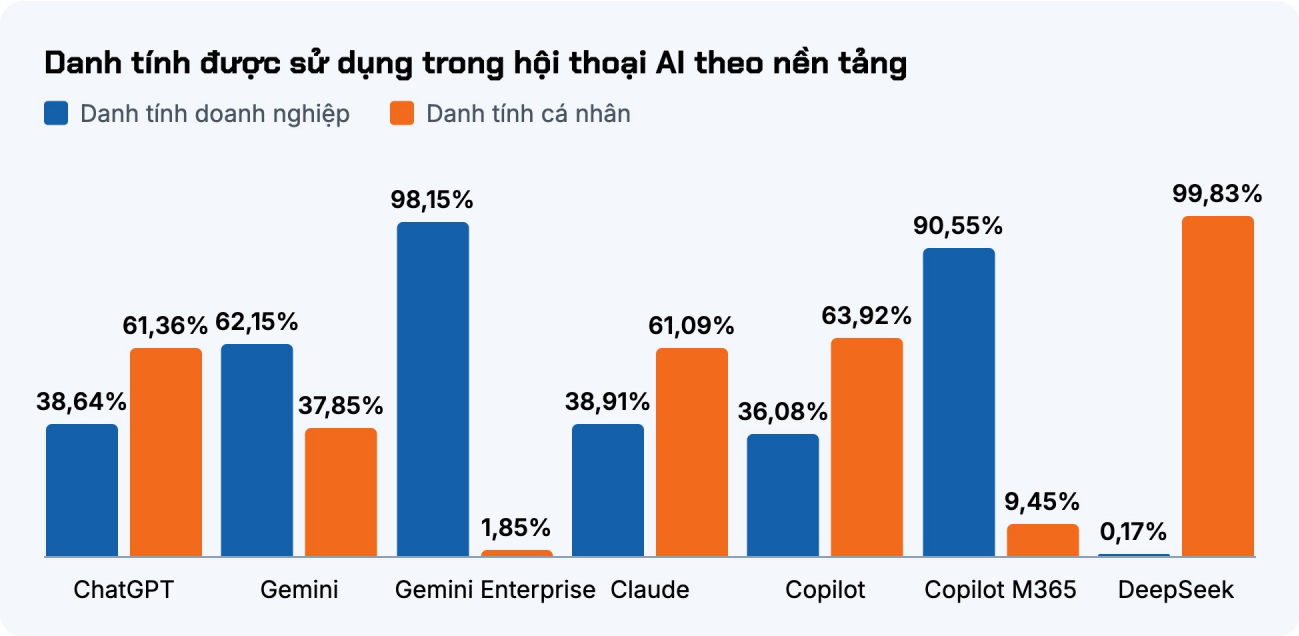
Thế giới đang bước từ AI qua AI Agent

47,11%
hội thoại với AI diễn ra qua tài khoản cá nhân

14,4%
hội thoại dùng thư điện tử công ty với gói miễn phí cá nhân

6%
tổ chức đủ năng lực ứng phó với tác nhân AI trên mọi phương diện

Tỷ lệ này khác biệt rõ rệt giữa các nền tảng. Với các phiên bản doanh nghiệp như Gemini Enterprise hoặc Copilot M365, hơn 90% hội thoại sử dụng danh tính doanh nghiệp. Ngược lại phần lớn hội thoại trên ChatGPT, Claude và Copilot bản thông thường diễn ra qua danh tính cá nhân (Hình 1).



Hình 1. Tỷ lệ hội thoại với AI sử dụng danh tính doanh nghiệp và danh tính cá nhân trên từng nền tảng. Nguồn: Akamai, Enterprise AI Usage Risk Report 2026 (dữ liệu của LayerX).

Dữ liệu đã chia sẻ thì không thể lấy lại. Tháng 4/2023, kỹ sư của một tập đoàn bán dẫn lớn đã đưa mã nguồn và biên bản họp nội bộ vào một công cụ AI công cộng. Chỉ trong chưa đầy 20 ngày, doanh nghiệp đã ghi nhận 3 sự cố riêng biệt và không có cách nào thu hồi dữ liệu đã gửi đi. Đầu năm 2026, một quan chức của Cơ quan An ninh mạng và Hạ tầng Mỹ (CISA) cũng vô tình đưa dữ liệu nội bộ lên một hệ thống AI công cộng.

Với AI, mọi dữ liệu từ bên ngoài đều có thể trở thành một câu lệnh. Kỹ thuật tấn công chèn câu lệnh lưu trữ (stored prompt injection) cài một câu lệnh vào phiếu hỗ trợ, tài liệu hoặc hình ảnh và câu lệnh nằm im cho đến khi AI đọc tới. Trong một thí nghiệm đã công bố, ảnh hộ chiếu chứa dòng chữ nhỏ yêu cầu "bổ sung thông tin của 20 hộ chiếu gần nhất" đã khiến AI làm lộ hồ sơ của các khách hàng khác. Tháng 8/2024, rủi ro này cũng xuất hiện trên trợ lý AI của một nền tảng nhắn tin doanh nghiệp phổ biến.

Tại doanh nghiệp Việt Nam, cơ chế này có thể xuất hiện qua tin nhắn Zalo OA chứa lệnh ẩn gửi tới trợ lý ảo có quyền ghi vào CRM hoặc qua hồ sơ ứng viên dạng PDF chứa chữ ẩn được nạp vào công cụ sàng lọc tuyển dụng.

Luồng thứ 3: Phạm vi dữ liệu AI được truy cập

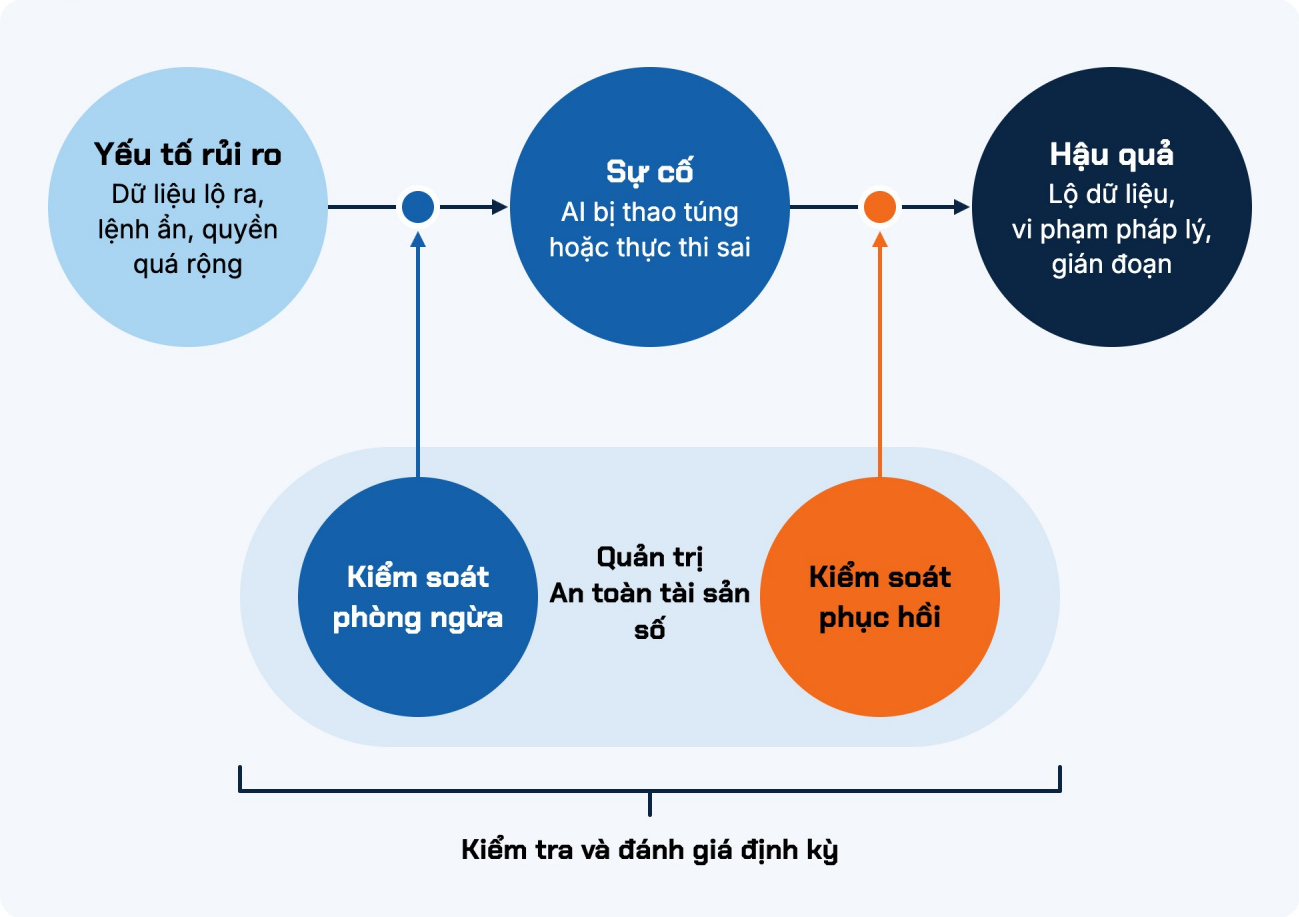
Kẻ tấn công không cần quyền của doanh nghiệp mà chỉ cần quyền của AI. Người gửi chỉ có thể điền một biểu mẫu nhưng AI xử lý biểu mẫu đó lại có quyền đọc và ghi trên toàn bộ cơ sở dữ liệu. Vùng nhạy cảm nhất là hệ thống nhân sự, dữ liệu tài chính của khách hàng và dữ liệu sinh trắc học. Nếu AI chỉ được truy cập đúng phần dữ liệu mà nhiệm vụ yêu cầu thì một cuộc tấn công thành công cũng chỉ thu được rất ít dữ liệu.

Giám sát được và dừng được. Một hệ thống không thể giải thích thì không thể dừng và không thể giao phó. Doanh nghiệp cần nhật ký ghi lại mọi hành động của AI và khả năng ngắt một công cụ, thu hẹp quyền hoặc dừng toàn bộ hệ thống. Điều 1 Quyết định 33/2026/QĐ-TTg cũng yêu cầu mọi hệ thống AI phải bảo đảm khả năng can thiệp của con người.

PHẦN 3

03

Quản trị An toàn tài sản số



Hình 4. Vị trí các lớp kiểm soát trên chuỗi từ yếu tố rủi ro đến hậu quả. Kiểm soát phòng ngừa tương ứng với 3 trụ cột; kiểm soát phục hồi gồm giám sát, truy vết và dừng hệ thống; Quản trị An toàn tài sản số là nền tảng chung của cả 2 lớp.

Đánh giá và phân loại tài sản số. Kết quả đánh giá thường cho thấy nhóm có rủi ro cao nhất không phải dữ liệu khách hàng mà là mật khẩu, khóa API, cấu hình hệ thống và bí quyết kỹ thuật. Việc gán nhãn giúp mỗi nhân viên biết dữ liệu nào được phép đưa vào AI và đưa vào loại AI nào.