TẠP CHÍ ĐIỆN TỬ
THƯƠNG HIỆU & CÔNG LUẬN

Kỷ nguyên “hacker AI” cận kề: Khi trí tuệ nhân tạo có thể tự tìm lỗ hổng, tự tấn công và làm tê liệt ngân hàng trong vài phút

21:19 19/04/2026

Những mô hình AI an ninh thế hệ mới đang khiến giới công nghệ toàn cầu vừa kỳ vọng vừa lo ngại: chúng có thể tự phát hiện lỗ hổng bảo mật mà con người mất hàng chục năm chưa tìm ra, đồng thời cũng có thể bị biến thành công cụ tấn công mạng cực mạnh. Khi AI có khả năng tự xây dựng kịch bản xâm nhập và khai thác điểm yếu mà không cần con người can thiệp, nhiều quốc gia đã bắt đầu coi đây là mối đe dọa an ninh cấp quốc gia.

Cuộc đua trí tuệ nhân tạo đang bước sang một giai đoạn mới, nơi AI không chỉ hỗ trợ con người xử lý dữ liệu hay sáng tạo nội dung, mà còn có thể trực tiếp tham gia vào mặt trận an ninh mạng với năng lực vượt xa các công cụ bảo mật truyền thống.

Sau khi Anthropic công bố hệ thống AI an ninh quy mô lớn mang tên “Mythos”, OpenAI cũng giới thiệu mô hình chuyên biệt về an ninh mạng có tên “GPT-5.4-Cyber”.

Điểm gây chú ý là các hệ thống này cho thấy năng lực phát hiện lỗ hổng bảo mật ở cấp độ mà giới chuyên gia an ninh đánh giá là chưa từng có.

Theo mô tả, Mythos có thể xác định các lỗ hổng nghiêm trọng trong hệ điều hành mà nhiều chuyên gia bảo mật phải mất tới hàng chục năm vẫn chưa phát hiện được.

Trong khi đó, GPT-5.4-Cyber được cho là có thể phân tích rủi ro an ninh chỉ thông qua tệp thực thi, không cần tiếp cận mã nguồn hay sơ đồ thiết kế hệ thống.



Ảnh minh hoạ

Điều này đồng nghĩa AI không còn chỉ đọc dữ liệu theo cách truyền thống, mà đã bắt đầu hiểu cấu trúc vận hành của phần mềm ở mức đủ để nhận diện điểm yếu tiềm ẩn.

Tuy nhiên, chính khả năng này đang tạo ra một nỗi lo mới: nếu AI có thể trở thành "lá chắn", nó cũng có thể bị biến thành "mũi giáo".

Các chuyên gia cảnh báo rằng việc chuyển đổi một mô hình AI phòng thủ thành công cụ tấn công không quá phức tạp.

Một khi được điều chỉnh mục tiêu, AI hoàn toàn có thể tự xây dựng đường tấn công, xác định mục tiêu dễ bị khai thác, thử nhiều phương án xâm nhập và thực hiện chuỗi hành động gần như tự chủ.

Khái niệm "hacker AI" – tức AI có thể tấn công mạng mà không cần con người chỉ đạo từng bước – vì thế đang dần từ giả thuyết trở thành kịch bản thực tế.

Trước nguy cơ này, nhiều quốc gia đã bắt đầu phản ứng mạnh.

Hoa Kỳ, Vương quốc Anh và Canada được cho là đã rà soát lại hệ thống phòng thủ số ở cấp quốc gia.

Tại Hàn Quốc, cơ quan giám sát tài chính đã triệu tập cuộc họp khẩn với lãnh đạo an ninh của nhiều ngân hàng lớn để đánh giá nguy cơ từ các cuộc tấn công sử dụng AI.

Mối lo không chỉ nằm ở các tổ chức tài chính lớn mà còn ở toàn bộ hệ sinh thái thanh toán số.

Trong thời đại ngân hàng số, khi phần lớn giao dịch tài chính được thực hiện trên smartphone, bất kỳ lỗ hổng nào cũng có thể tạo ra hiệu ứng dây chuyền rất lớn.

Một cuộc tấn công AI đủ mạnh có thể khiến tiền trong tài khoản biến mất chỉ trong vài phút hoặc làm gián đoạn hệ thống thanh toán diện rộng.

Những sự cố gần đây tại Hàn Quốc càng khiến nỗi lo này trở nên cụ thể hơn.

Toss Bank từng gặp sự cố nhập sai tỷ giá, dẫn đến vụ giao dịch ngoại tệ sai lệch gây thiệt hại lớn chỉ trong 7 phút.

Bithumb từng ghi nhầm khoản nạp lên tới hàng chục nghìn tỷ won vì lỗi nhập liệu nội bộ.

Trong khi đó, Upbit cũng từng trở thành mục tiêu của các cuộc tấn công mạng gây thất thoát lớn.

Dù chưa phải do AI trực tiếp gây ra, các sự cố này cho thấy mức độ mong manh của hệ thống tài chính số khi đối mặt với lỗi kỹ thuật hoặc lỗ hổng bảo mật.

Nếu AI được khai thác vào mục đích xấu, tốc độ và quy mô tấn công có thể vượt xa khả năng phản ứng của con người.

Nhiều chuyên gia an ninh cho rằng tư duy “khả năng thấp nên chưa cần đầu tư mạnh” không còn phù hợp.

Thay vào đó, an ninh mạng cần được đặt ngang tầm an ninh quốc gia.

Một trong những giải pháp được nhấn mạnh là xây dựng hệ thống phòng thủ tự động bằng AI – tức dùng AI để chống lại AI.

Các hệ thống này phải đủ khả năng phát hiện hành vi bất thường theo thời gian thực, tự động chặn truy cập nguy hiểm và phản ứng ngay khi xuất hiện mô hình tấn công mới.

Bên cạnh đó, cơ chế “ngắt mạch mạng” cũng được xem là biện pháp bắt buộc.

Nếu phát hiện giao dịch bất thường hoặc luồng truy cập nguy hiểm, hệ thống cần có quyền ngắt ngay chuỗi xử lý để tránh thiệt hại lan rộng.

Ngoài giải pháp kỹ thuật, nhiều ý kiến cho rằng cần áp dụng chế tài mạnh hơn với doanh nghiệp không đáp ứng chuẩn an ninh số.

Bởi trong thời đại hacker AI, chỉ một mắt xích yếu cũng có thể trở thành cánh cửa mở ra rủi ro cho cả hệ thống.

Điều đáng lo nhất là nếu một cuộc tấn công AI đủ mạnh xuyên thủng được hạ tầng số quốc gia, khoảng thời gian để "chữa cháy" có thể gần như không còn nữa

Phương Thảo *(t/h)*