Lỗi của Microsoft khiến các email bí mật bị lộ cho công cụ AI Copilot.

2 ngày trước

Liv McMahon
Phóng viên công nghệ

Ảnh của Getty Images

Microsoft đã thừa nhận một lỗi khiến trợ lý AI của họ truy cập và tóm tắt nhầm một số email bí mật của người dùng.

Ông lớn công nghệ này đã quảng bá Microsoft 365 Copilot Chat như một cách an toàn để các nơi làm việc và nhân viên của họ sử dụng chatbot trí tuệ nhân tạo tạo sinh của mình.

Tuy nhiên, họ cho biết một sự cố gần đây đã khiến công cụ này tiết lộ thông tin cho một số người dùng doanh nghiệp từ các tin nhắn được lưu trữ trong thư nháp và thư đã gửi của họ - bao gồm cả những tin nhắn được đánh dấu là bí mật.

Microsoft cho biết họ đã tung ra bản cập nhật để khắc phục sự cố và khẳng định "không ai được phép truy cập thông tin mà họ chưa được phép xem".

Tuy nhiên, một số chuyên gia cảnh báo rằng tốc độ cạnh tranh gay gắt giữa các công ty trong việc bổ sung các tính năng AI mới đồng nghĩa với việc những sai sót kiểu này là điều khó tránh khỏi.

Copilot Chat có thể được sử dụng trong các chương trình của Microsoft như Outlook và Teams, dùng cho email và chức năng trò chuyện, để nhận câu trả lời cho câu hỏi hoặc tóm tắt tin nhắn.

"Chúng tôi đã xác định và khắc phục sự cố trong đó Microsoft 365 Copilot Chat có thể trả về nội dung từ các email được gắn nhãn 'bí mật' do người dùng soạn thảo và được lưu trữ trong mục Thư nháp và Thư đã gửi của họ trên Outlook phiên bản máy tính để bàn," một phát ngôn viên của Microsoft nói với BBC News.

"Mặc dù các biện pháp kiểm soát truy cập và chính sách bảo vệ dữ liệu của chúng tôi vẫn được giữ nguyên, nhưng hành vi này không đáp ứng được trải nghiệm Copilot mà chúng tôi mong muốn, vốn được thiết kế để loại trừ nội dung được bảo vệ khỏi quyền truy cập của Copilot", họ nói thêm.

"Bản cập nhật cấu hình đã được triển khai trên toàn thế giới cho khách hàng doanh nghiệp."

Lỗi này lần đầu tiên được trang tin công nghệ **Bleeping Computer** đưa tin , trang này cho biết họ đã nhận được một thông báo dịch vụ xác nhận vấn đề.

Thông báo này dẫn lời một thông báo của Microsoft cho biết "các tin nhắn email của người dùng có gắn nhãn bảo mật đang bị xử lý không chính xác bởi tính năng trò chuyện của Microsoft 365 Copilot".

Thông báo cho biết thêm rằng tab công việc trong Copilot Chat đã tóm tắt các tin nhắn email được lưu trữ trong thư mục nháp và thư mục đã gửi của người dùng, ngay cả khi chúng có nhãn bảo mật và chính sách ngăn ngừa mất dữ liệu được cấu hình để ngăn chặn việc chia sẻ dữ liệu trái phép.

Các báo cáo cho thấy Microsoft lần đầu tiên phát hiện ra lỗi này vào tháng Giêng.

Thông báo về lỗi này cũng được chia sẻ trên bảng điều khiển hỗ trợ dành cho nhân viên NHS ở Anh - nơi nguyên nhân gốc rễ được cho là do "lỗi mã lập trình".

Một phần **thông báo trên trang hỗ trợ CNTT của NHS** cho thấy trang này đã bị ảnh hưởng.

Tuy nhiên, họ nói với BBC News rằng nội dung của bất kỳ bản nháp hoặc email đã gửi nào được xử lý bởi Copilot Chat sẽ vẫn thuộc quyền sở hữu của người tạo ra chúng, và thông tin bệnh nhân không bị lộ.

'Rò rỉ dữ liệu là điều khó tránh khỏi'

Các công cụ AI dành cho doanh nghiệp như Microsoft 365 Copilot Chat - dành cho các tổ chức có đăng ký Microsoft 365 - thường có các biện pháp kiểm soát và bảo vệ an ninh nghiêm ngặt hơn để ngăn chặn việc chia sẻ thông tin nhạy cảm của công ty.

Tuy nhiên, đối với một số chuyên gia, vấn đề này vẫn nêu bật những rủi ro khi áp dụng các công cụ trí tuệ nhân tạo tạo sinh trong một số môi trường làm việc nhất định.

Nader Henein, nhà phân tích về bảo vệ dữ liệu và quản trị AI tại Gartner, cho biết "sai sót kiểu này là không thể tránh khỏi", do tần suất ra mắt các "khả năng AI mới và độc đáo".

Ông nói với BBC News rằng các tổ chức sử dụng các sản phẩm trí tuệ nhân tạo này thường thiếu các công cụ cần thiết để tự bảo vệ mình và quản lý từng tính năng mới.

"Trong điều kiện bình thường, các tổ chức sẽ chỉ đơn giản là tắt tính năng này và chờ đến khi hệ thống quản trị được cập nhật," Henein nói.

"Thật không may, áp lực từ làn sóng cường điệu hóa về trí tuệ nhân tạo thiếu căn cứ khiến điều đó gần như không thể," ông nói thêm.

Giáo sư Alan Woodward, chuyên gia an ninh mạng thuộc Đại học Surrey, cho biết điều này cho thấy tầm quan trọng của việc thiết lập các công cụ như vậy ở chế độ riêng tư theo mặc định và chỉ cho phép người dùng tự nguyện sử dụng.

"Chắc chắn sẽ có lỗi trong những công cụ này, nhất là khi chúng phát triển với tốc độ chóng mặt, vì vậy ngay cả khi việc rò rỉ dữ liệu không phải là cố ý thì nó vẫn sẽ xảy ra," ông nói với BBC News.