Một số vấn đề an ninh thông tin trọng yếu trong kỷ nguyên AI (Phần 1)

1. Tiến bộ công nghệ, các thách thức bảo mật, và an ninh thông tin

Sự tiến bộ nhanh chóng của các nền tảng Công nghệ Thông tin (CNTT) và ngôn ngữ lập trình đã làm thay đổi hình thái vận động và phát triển của xã hội loài người. Không gian mạng và các tiện ích đi kèm ngày càng được mở rộng, dẫn đến sự chuyển dịch dần từ đời sống trong thế giới thực sang đời sống trong thế giới ảo (còn gọi là không gian mạng hay không gian số). Tính đến năm 2023, Internet Vạn Vật (Internet of Things – IoT) đã kết nối với khoảng 15,14 tỷ thiết bị trên toàn cầu (Vailshery, 2023). Trung bình mỗi người trên Trái Đất hiện nay đang sở hữu xấp xỉ 1,89 thiết bị, tăng gần 24 lần so với 20 năm trước (trung bình mỗi người sở hữu 0.08 thiết bị vào năm 2003) (Lu & Da Xu, 2018). Con số này được dự đoán sẽ tăng lên gần gấp đôi vào năm 2030, với khoảng 29,42 tỷ thiết bị được kết nối. Chủng loại các thiết bị sẽ ngày càng đa dạng, được trang bị các hệ thống cảm biến hoặc bộ điều khiển để chúng có thể tương tác với con người tốt hơn, và tích hợp trí tuệ nhân tạo (AI) để phục vụ việc ra quyết định, tìm kiếm, và truyền tải thông tin cho người sử dụng.

Trong bối cảnh các hoạt động kinh tế và xã hội ngày càng được kết nối tốt hơn thông qua IoT, và sắp tới đây là tiềm năng tích hợp AI vào hầu như mọi mặt của đời sống ở cả thế giới thực và thế giới ảo, thì không chỉ cá nhân, mà cả doanh nghiệp và quốc gia cũng sẽ phải đối mặt với các thách thức chưa từng có tiền lệ đối với rủi ro về an ninh thông tin

(Keck *et al.*, 2022).

Khi hệ thống CNTT, đặc biệt là Internet, được tích hợp vào cuộc sống, một lượng thông tin khổng lồ sẽ được tạo ra, lưu trữ và truyền tải, như dữ liệu về thông tin cá nhân, các tương tác trên mạng xã

hội, thông tin doanh nghiệp, dữ liệu giao dịch, hồ sơ bảo hiểm, y tế, v.v.. Một khi những dữ liệu này bị rò rỉ, chúng có khả năng bị lợi dụng để trục lợi và gây ra ảnh hưởng tiêu cực đến cuộc sống của các cá nhân, vận hành của doanh nghiệp, và sự ổn định và phát triển bền vững của quốc gia. Báo cáo Rủi ro Toàn cầu của Diễn đàn Kinh tế Thế giới (WEF) năm 2023 đã xếp thử thách liên quan đến tội phạm mạng và an ninh mạng là một trong 10 rủi ro hàng đầu ở hiện tại và trong cả tương lai (World Economic Forum, 2023). Thật vậy, các cuộc tấn công mạng đã tăng 600% kể từ khi bắt đầu Đại dịch COVID-19 với hơn 5,4 tỷ cuộc tấn công bằng phần mềm độc hại chỉ riêng trong năm 2022 (RiskXchange, 2023). Theo dự báo của Cybersecurity Ventures, các hoạt động tội phạm mạng được dự kiến gây ra thiệt hại xấp xỉ 10,5 nghìn tỷ đô la hàng năm kể từ 2025. Các thiệt hại này bao gồm các tổn thất về dữ liệu, tiền của bị đánh cắp, sự suy giảm năng suất, mất mát tài sản trí tuệ, trộm cắp dữ liệu cá nhân và tài chính, gian lận, gây rối sau cuộc tấn công vào quá trình kinh doanh bình thường, điều tra sau vụ tấn công, khôi phục và xóa dữ liệu và hệ thống đã bị tấn công, và suy giảm uy tín (Morgan, 2022). Đây chỉ là ước tính trực tiếp về tổn hại về kinh tế mà chưa kể các ảnh hưởng gián tiếp lên cả hệ thống kinh tế và xã hội toàn cầu (Chính & Hoàng, 2009).

Khi xã hội đang ngày càng hướng đến sự tiện nghi và hữu ích của các đô thị thông minh, thì việc tích hợp một lượng lớn các thiết bị điện tử và phần mềm vào cuộc sống để quản lý tài sản, tài nguyên, và dịch vụ trở thành một xu hướng tất yếu. Như vậy, thông tin sẽ được thu thập từ từng công dân, thiết bị, tòa nhà, và hệ thống vận hành, để giúp giám sát và quản lý hệ thống giao thông, nhà máy điện, hệ thống cung cấp nước, hệ thống xử lý chất thải, hệ thống truyền tải thông tin, trường học, bệnh viện, đảm bảo an ninh, và các dịch vụ xã hội khác (Musa, 2018; Paiho *et al.*, 2022). Với sự liên kết sâu rộng và phức tạp của các hệ thống thiết bị và phần mềm, các cuộc tấn công mạng có thể làm tê liệt sự vận hành toàn phần hoặc một phần của xã hội và quốc gia một cách nhanh chóng hoặc chiếm quyền kiểm soát của hệ thống nếu vấn đề an ninh mạng không được đảm bảo. Việc *hacker* tấn công vào hệ thống quản lý tiện ích và kiểm soát quyền điều khiển của thiết bị đã từng xảy ra đối với hệ thống Uconnect, tính năng máy tính kết nối Internet dùng cho mục đích giải trí, dẫn đường, cho phép thực hiện cuộc gọi điện thoại, và cung cấp điểm truy cập Wi-fi trên xe, vào năm 2015. *Hacker* đã sử dụng lỗ hổng bảo mật của hệ thống để điều khiển xe tự động tắt máy, tăng và giảm tốc độ của xe, gây mất an toàn cho người sử dụng. Điều này đã khiến hàng loạt các công ty ô tô ở Mỹ ra quyết định thu hồi hơn một triệu xe đang sử dụng hệ thống, gây ra những tổn thất kinh tế rất lớn (Greenberg, 2015).

Ngoài ra, các loại thông tin được tạo ra trên không gian mạng có thể được xem là một loại tài nguyên mới giúp tạo nên các tập dữ liệu cực kỳ lớn và đa dạng (*big data*). Chúng có thể được phân tích nhằm tìm ra các mối liên hệ và xu hướng trong hành vi của con người và các tương tác xã hội. Hiện nay, cứ cách mỗi phút sẽ có 6,3 triệu lượt tìm kiếm thông tin được thực hiện trên Google, hơn 527 nghìn bức ảnh được chia sẻ trên Snapchat, 456 nghìn lượt tweet trên nền tảng X (trước kia là Twitter), hơn 46 nghìn bức ảnh được đăng tải lên Instagram, và khoảng 510 nghìn lượt bình luận được đăng và 293 nghìn trạng thái được cập nhật mới trên Facebook (Marr, 2021; Wise, 2023). Thông qua việc sử dụng các kỹ thuật phân tích phức tạp và thuật toán, lượng lớn dữ liệu này có thể được dùng để cho thấy suy nghĩ, cảm xúc, và hành vi của người dùng trên mạng xã hội, và dùng chúng cho các kế hoạch thao túng tâm lý và hành vi tinh vi (Ho & Vuong, 2023). Đấy là chưa kể, con người ngày càng có cảm xúc với các nhân vật, tài sản, và ứng dụng trong không gian mạng, nên có thể dễ dàng bị ảnh hưởng về tâm lý, tình cảm và hành vi hơn (Mantello *et al.*, 2023; Vuong *et al.*, 2023a, 2023b).

Một minh chứng cho điều trên là vụ việc của Cambridge Analytica, một công ty tư vấn đã thu thập dữ liệu cá nhân từ hàng chục triệu người dùng Facebook và bán nó cho các chiến dịch thao túng tâm lý tình cảm cử tri, nhằm tác động lên kết quả bầu cử chính trị. Vụ bê bối này không chỉ tiết lộ sức mạnh của các bên nắm giữ nguồn tài nguyên thông tin, đặc biệt là các tập đoàn công nghệ, và cách thức mà loại sức mạnh này có thể được sử dụng để tác động đến sự vận hành kinh tế, xã hội, và chính trị (Liaropoulos, 2020; Nilekani, 2018).

Gần đây nhất, sự ra đời của ChatGPT 3.5 vào ngày 30/11/2022 đã đánh dấu cho sự khởi đầu của giai đoạn được nhiều chuyên gia gọi là "thời đại AI". Chỉ sau 1 tháng sau khi được cho ra mắt, ChatGPT

đã thu hút hơn 100 triệu người dùng, biến nó trở thành phần mềm thông dụng, phát triển nhanh nhất trong lịch sử (Hu, 2023). Sự bùng nổ về người dùng đã thúc đẩy việc phát hành các sản phẩm AI cạnh tranh khác, bao gồm Gemini, Ernie Bot, LLaMA, Claude và Grok trong năm 2023. Trên thực tế, công nghệ AI đã được đưa vào ứng dụng rộng rãi trong nhiều khía cạnh đời sống từ trước đây, như nghiên cứu khoa học, chăm sóc sức khỏe, tài chính, giải trí, giáo dục, và giao thông. Một số ứng dụng nổi bậc vận hành bởi AI mà chúng ta sử dụng gần như hằng ngày có thể kể đến như công cụ tìm kiếm web nâng cao (vd: Google Search) và hệ thống đề xuất (được sử dụng bởi YouTube, Amazon và Netflix). Tuy nhiên, khi đấy khả năng vận hành (vd: yêu cầu chuyên môn về CNTT) và khả năng tiếp cận (vd: chi phí đắt đỏ) vẫn là rào cản lớn đối với nhận thức của xã hội về AI cũng như các công năng của chúng ở trong đời sống.

Sự mở rộng và phát triển của các Mô hình Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing – NLP) và Mô hình Ngôn ngữ Lớn (Large Language Model – LLM) thể hiện những đặc điểm giống con người về lập luận, nhận thức, sự chú ý và sáng tạo đã giúp con người vượt qua rào cản về khả năng vận hành (Lappin, 2023; Vuong *et al*., 2023). Những công việc đòi hỏi sự vận hành của các chuyên gia công nghệ thông tin giờ đây có khả năng được hoàn thành bởi người bình thường chỉ thông qua vài mệnh lệnh sử dụng ngôn ngữ hàng ngày. Ngoài ra, AI đang trở nên mạnh mẽ hơn và có giá thành rẻ hơn đáng kể qua thời gian (tính bằng tháng), khiến cho những công việc trước đây không thể thực hiện được do chi phí tính toán quá lớn thì giờ đây đã trở nên phổ biến (Suleyman, 2023). Nói cách khác, AI đang và sẽ tiếp tục mang đến cho văn minh nhân loại loại sức mạnh cực kỳ to lớn, tới mức Sundar Pichai, CEO của Google, cho rằng mức độ quan trọng của AI còn vượt qua cả lửa và điện (Clifford, 2018).

Đi kèm với các tiềm năng to lớn của AI là những lỗ hổng bảo mật mới và các rủi ro về an ninh thông tin phức tạp hơn. Khi xã hội vẫn đang cố gắng chuyển mình sang giai đoạn mới để có thể thích ứng với sự biến đổi công nghệ, cuộc cách mạng AI lại tiếp tục diễn ra đòi hỏi chúng ta phải suy nghĩ lại quỹ đạo của quá trình chuyển đổi xã hội vì nó có thể làm trầm trọng hơn các rủi ro an ninh thông tin được trình bày ở trên. Cụ thể, xã hội cần tiến hóa như thế nào để có thể bắt kịp với sự biến đổi mang tính đột phá tạo ra bởi làn sóng công nghệ AI hiện nay? Làm sao để quản lý và tận dụng sức mạnh của chúng trong khi vẫn đảm bảo được an ninh thông tin khi không gian và thời gian sống trong thế giới ảo của chúng ta ngày càng tăng?

Để góp phần trả lời các câu hỏi trên, phần tiếp theo của bài viết sẽ bàn về các vấn đề và rủi ro ảnh hưởng đến an ninh thông tin trong kỷ nguyên AI, cũng như là vai trò, lợi thế, và cơ hội của việc vận dụng AI cho mục đích đảm bảo an ninh thông tin. Tiếp theo đấy, cách con người tương tác với AI và các quyền tự do vận dụng AI cho mục đích cá nhân có ảnh hưởng như thế nào đối với an ninh thông tin trên không gian mạng sẽ được đưa ra xem xét và tạo nền tảng cho các bàn luận về vai trò của chính phủ, doanh nghiệp, và người dân trong việc đảm bảo an ninh thông tin.

2. Kỷ nguyên trí tuệ nhân tạo và ảnh hưởng với an ninh mạng

2.1. Tác động của AI lên hoạt động tấn công và phòng thủ

Công nghệ AI đã cho thấy tiềm năng vượt trội trong việc tự động hóa các tác vụ, đưa ra dự đoán và nâng cao hiệu quả. Chính vì thế mà AI cũng đã và đang cách mạng hóa lĩnh vực an ninh thông tin. An ninh thông tin là các hoạt động quản lý, giám sát, và bảo vệ được thực hiện để giảm thiểu các rủi ro thông tin. Đối với các hoạt động bảo vệ và phòng thủ thông tin cá nhân, hệ thống máy tính, và cơ sở hạ tầng quan trọng, trọng tâm chính của chúng là đạt được bộ ba CIA, trong khi vẫn đảm bảo được sự vận hành hiệu quả của hệ thống được bảo vệ. Bộ ba CIA bao gồm (Maalem Lahcen *et al*., 2020):

- Tính bảo mật (*Confidentiality-C*): bảo vệ dữ liệu và hệ thống trước các rủi ro đến từ các hoạt động trộm dữ liệu nhắm vào các cơ sở dữ liệu, bản sao lưu, máy chủ ứng dụng, và các hệ thống quản trị.
- Tính toàn vẹn (*Integrity-I*): bảo vệ dữ liệu và hệ thống trước các rủi ro ảnh hưởng đến sự toàn vẹn của thông tin và hệ thống quản trị, bao gồm cướp đoạt quyền kiểm soát, thay đổi dữ liệu

tài chính, trộm tiền, điều hướng các thông tin được lưu trữ, và làm tổn thất thương hiệu của tổ chức.
- Tính sẵn sàng/khả dụng (*Availability-A*): bảo vệ dữ liệu và hệ thống trước các cuộc tấn công từ chối dịch vụ (DDoS), tấn công từ chối dịch vụ có mục tiêu, và rủi ro phá hủy vật lý.

Sự xuất hiện của AI đã làm tăng đồng thời năng lực tấn công mạng của các *hacker* và khả năng phòng thủ và bảo mật của các nhà quản trị mạng lên đáng kể. Nhờ vào khả năng tự động hóa các công việc lặp đi lặp lại và tránh được các điểm mù về nhận thức con người hay gặp phải, các thuật toán học máy có khả năng phân tích một lượng thông tin khổng lồ để tìm ra các lỗ hổng bảo mật mà trước đây không thể phát hiện được (Rao, 2021). Ở chiều hướng phòng thủ, công việc rà soát và tìm kiếm các lỗ hổng bảo mật trước đây mất rất nhiều thời gian và công sức do số lượng các lỗi bảo mật của các nền tảng được ghi nhận là rất lớn. Để tìm được các lỗ hổng chưa được vá thường dựa nhiều vào kinh nghiệm các *hacker* mũ trắng, các kỹ thuật viên bảo mật, và các công cụ quét lỗi (*Vulnerability Scanning Tools*). Điều này dẫn đến việc các hệ thống không được rà soát và vá lỗi kỹ càng nên chúng nhanh chóng bị *hacker* phát hiện và khai thác. Các công cụ dựa trên AI hiện nay có thể được sử dụng để tự động hóa quá trình xác định các lỗ hổng này trong hệ thống phần mềm, mạng và các tài sản kỹ thuật số khác trước khi bị *hacker* tìm ra và khai thác.

Ngoài ra, công cụ sử dụng AI khiến cho các cuộc tấn công ngày càng trở nên đa dạng và tinh vi hơn. Tội phạm mạng sử dụng một số chiến thuật dựa trên AI để xâm nhập vào hệ thống thông tin cá nhân và mạng công ty, chẳng hạn như:

- Phát triển phần mềm độc hại và mã độc tống tiền (*ransomware*) nâng cao.
- Thực hiện các cuộc tấn công ẩn dấu (*stealth attack*).
- Sử dụng AI để đoán mật khẩu phức tạp và phá CAPTCHA.
- Tạo dữ liệu *deepfake* và mạo danh cá nhân trên nền tảng truyền thông xã hội.
- Sử dụng các khung AI để tấn công các hệ thống dễ bị tổn thương.
- Tận dụng Machine Learning (ML) để tăng cường thử nghiệm thâm nhập.

Các công cụ dựa trên AI cũng có thể được sử dụng để tạo ra các cuộc tấn công hỗn hợp có tính mục tiêu cao, thiết kế riêng cho các cá nhân hoặc tổ chức cụ thể (Handa *et al.*, 2019). Chúng cho phép tội phạm mạng xâm nhập và ẩn náu trong mạng của một công ty trong thời gian dài để thực hiện các cuộc tấn công ẩn dấu (*stealth attack*). Trong thời gian này, nó có thể thiết lập các điểm truy cập bí mật vào cơ sở hạ tầng quan trọng của tổ chức. Trong khi chuẩn bị sẵn sàng để bắt đầu một cuộc tấn công rộng hơn, những tên tội phạm này có thể chặn thông tin liên lạc, đánh cắp dữ liệu, phổ biến phần mềm có hại, tạo tài khoản có quyền truy cập cao để xâm nhập các hệ thống khác hoặc triển khai phần mềm *ransomware*.

Hay như phương thức tấn công lừa đảo (*phishing*) cũng trở nên tinh vi hơn với sự trợ giúp của AI. Bạn có thể dễ dàng nhận được một email giả mạo (*fake email*), một cuộc gọi, thậm chí là cuộc gọi video, mạo danh ngân hàng, cơ quan quản lý nhà nước hay kể cả người thân của bạn. Các thông tin giả được tạo lập bởi AI (*deepfake*) có thể hoàn toàn bắt chước các giao thức bảo mật của các cơ quan quản lý hay có giọng nói, hành vi trùng khớp với đối tượng bị mạo danh.

Ở chiều hướng ngược lại, khả năng học hỏi và dự đoán các tình huống hiện tại và tương lai thành thạo của AI khiến nó trở thành một công cụ có khả năng cập nhật, phát triển, và thích nghi với sự thay đổi trong các phương thức tấn công của tội phạm mạng. Ví dụ như khả năng phân tích và phát hiện phần mềm độc hại (*malware*) của AI. Trong vài thập kỷ qua, phần mềm độc hại đã liên tục phát triển và tiến hóa với tốc độ cao, tạo ra nhiều phần mềm độc hại tiên tiến có khả năng thay đổi cấu trúc/mã sau mỗi lần lây nhiễm (như phần mềm đa hình và phần mềm biến đổi hình thức) (Sharma & Sahay, 2014). Điều này giúp chúng có khả năng xuyên thủng các hàng rào bảo mật truyền thống như tường lửa, vô hiệu hóa các hệ thống phát hiện xâm nhập. Để đối phó với chúng, các công nghệ AI đang trở nên dần phổ biến vì chúng không chỉ giúp phát hiện các phần mềm độc hại mà còn giúp dự đoán và cập nhật kiến thức về các dạng phần mềm độc hại mới hoặc không rõ ràng (Rieck *et al.*, 2011). Bên cạnh khả năng phân tích và phát hiện phần mềm độc hại, AI cũng đã

và đang được phát triển để nhận biết và đối phó với các cuộc tấn công lừa đảo, các thông tin rác (*spam*), các hoạt động xâm nhập vào hệ thống quản lý giao thông, và các cuộc tấn công vào hệ thống điện và hệ thống kiểm soát công nghiệp (Handa *et al.*, 2019; Martínez Torres *et al.*, 2019).

2.2. Một số giới hạn của AI

Mặc dù AI đang được xem là giải pháp hàng đầu cho nhu cầu đảm bảo an ninh thông tin ngày càng tăng, nó cũng tồn tại một số giới hạn. Đầu tiên phải kể đến là chi phí cần đầu tư để phát triển hệ thống AI độc lập được thiết kế riêng cho nhu cầu bảo mật. Mặc dù không hoàn toàn chính xác, nhưng ta có thể lấy mô hình ChatGPT-3 của OpenAI làm tham khảo. Các nhà phân tích và nhà công nghệ ước tính rằng quá trình đào tạo một mô hình ngôn ngữ như ChatGPT-3 có thể tiêu tốn hơn 4 triệu USD (Vanian & Leswing, 2023). Chưa kể, để thực hiện quá trình đào tạo này, công ty phải có khả năng tiếp cận tới các chuyên gia, máy móc, dữ liệu, và cơ sở dữ liệu phù hợp. Gần như phần lớn các cá nhân, các doanh nghiệp vừa và nhỏ đều không có khả năng thực hiện.

Tất nhiên, chi phí khi sử dụng mô hình AI cung cấp bởi các công ty công nghệ sẽ thấp hơn nhiều lần. Ví dụ như hệ thống bảo mật Copilot được cung cấp bởi Microsoft. Phần mềm này được phát triển dựa trên GPT-4, mô hình ngôn ngữ lớn nhất hiện tại từ OpenAI – mà Microsoft đã đầu tư hàng tỷ USD – và một mô hình cụ thể về bảo mật mà Microsoft xây dựng bằng cách sử dụng dữ liệu hoạt động hàng ngày mà các hệ thống thu thập được (Novet, 2023). Microsoft có kế hoạch sẽ thu mức phí là 4 USD cho mỗi "đơn vị tính toán bảo mật", và người dùng có thể tùy chọn chỉ mua những gì họ cần cho nhu cầu bảo mật phù hợp (Novet, 2024). Tuy nhiên, chi phí thấp lại cũng đi kèm với rủi ro về an ninh thông tin khác: các thông tin về môi trường bảo mật của người sử dụng sẽ bị các công ty công nghệ thu thập. Chính Microsoft cũng đã thừa nhận rằng: "Hệ thống [Copilot] sẽ biết về môi trường bảo mật của khách hàng, nhưng dữ liệu đó sẽ không được sử dụng để huấn luyện các mô hình" (Novet, 2024). Mặc dù Microsoft cam kết họ sẽ không dùng những dữ liệu thu thập được cho mục đích "huấn luyện mô hình", nhưng còn với các mục đích khác ngoài việc "huấn luyện mô hình" thì họ không nói rõ. Nếu bản thân người sử dụng và doanh nghiệp không quan tâm đến điều này vì quá trình vận hành của họ không bị ảnh hưởng, nhưng khi quy mô mở rộng ra hàng triệu người sử dụng và hàng trăm ngàn công ty thì lượng thông tin thu thập được đấy sẽ có giá trị sử dụng cho các hoạt động gián điệp và thao túng ở quy mô quốc gia và khu vực. Điều này thật đáng sợ khi chúng ta trả tiền cho việc tăng cường khả năng bảo mật, nhưng lại cho phép bên cung cấp dịch vụ bảo mật biết tất cả thông tin về điểm yếu trong hệ thống của chúng ta.

Ngoài ra, khi AI được áp dụng vào công việc đảm bảo an ninh rộng rãi hơn, các lỗ hổng an ninh phi truyền thống cũng sẽ xuất hiện nhiều hơn. AI cung cấp khả năng đưa ra quyết định tự động và liên tục trong khoảng thời gian dài, giúp phát hiện các phần mềm độc hại hoặc các bất thường trong hệ thống. Tuy nhiên, để làm được điều này, AI phải được huấn luyện cách phân biệt các dấu hiệu của các phần mềm độc hại hoặc tiềm ẩn những vẫn hành bất thường. Tội phạm mạng có khả năng lợi dụng giai đoạn huấn luyện này để điều chỉnh đầu ra của mô hình phân loại, từ đấy thao túng hệ thống AI để cho phép các phần mềm hoặc mã độc xâm nhập vào hệ thống (Biggio, Fumera, *et al.*, 2013; Handa *et al.*, 2019). Các dạng tấn công này có thể được chia làm hai loại (Biggio, Corona, *et al.*, 2013):

- Tấn công đầu độc (*poisoning attack*): kẻ tấn công tác động vào dữ liệu huấn luyện để thay đổi quá trình huấn luyện và gây tổn hại đến hiệu suất phân loại của AI.
- Tấn công tránh né (*evasion attack*): kẻ tấn công sử dụng các chiến lược nhằm thăm dò hoặc phân tích ngoại tuyến để tìm ra các thông tin giúp họ thao túng phán đoán của hệ thống phân loại mà không cần phải tác động vào quá trình huấn luyện của AI.

Mặc dù AI có thể cung cấp giải pháp đầy sức mạnh cho mục đích bảo mật, nhưng chúng không phải công cụ vạn năng. AI vẫn phải chịu sự kiểm soát và chi phối của người sử dụng, nên hệ thống bảo mật sẽ luôn có khả năng tồn tại các lỗ hổng bảo mật gây ra bởi lỗi của con người. Những lỗi con người này có thể được phân loại dựa trên hậu quả và ý định của người thực hiện (Maalem Lahcen *et al.*, 2020):

- Lỗi không cố ý (*unintentional human error*): lỗi bắt nguồn tự sự thiếu kiến thức hoặc kỹ năng vận hành.
- Lỗi cố ý (*intentional human error*): lỗi được gây ra bởi một người dùng biết về hành vi rủi ro nhưng vẫn hành động dựa trên nó, hoặc sử dụng hệ thống một cách sai trái. Hành động sai không nhất thiết phải gây ra tổn thất ngay lập tức cho tổ chức, nhưng vẫn có thể vi phạm các luật hiện hành hoặc quyền riêng tư.
- Lỗi độc hại (*malicious human error*): lỗi tệ nhất vì nó được thực hiện với ý định cụ thể dùng để gây hại cho hệ thống.

Do những người vận hành và kiểm soát dữ liệu và hệ thống không nằm trong phạm vi kiểm soát của hệ thống AI, nên các lỗ hổng bảo mật vẫn có thể được tạo ra từ các hành vi phá hoại có chủ đích từ bên trong nội bộ (hoặc bản thân người vận hành) (Maalem Lahcen *et al.*, 2020). Đôi khi các quyết đinh và hành vi của con người là bất hợp lý và không thể đoán trước, do bị ảnh hưởng bởi sự tức giận, bế tắc, và cả bất mãn trong công việc, khiến họ thực hiện phá hoại có chủ đích (lỗi độc hại), can thiệp không an toàn (lỗi cố ý), thực hiện các lỗi "ngây thơ" do không chú ý (lỗi không cố ý), v.v.. (Stanton *et al.*, 2005). Theo báo cáo 2023 của Insider Threat Report, 74% số chuyên gia an ninh mạng được khảo sát cảm thấy an ninh của dữ liệu và hệ thống dễ bị tổn thương trước các mối đe dọa từ nội bộ. 74% chuyên gia được khảo sát cũng cho biết các cuộc tấn công từ trong nội bộ cũng trở nên thường xuyên hơn trong 12 tháng vừa qua (Insiders, 2023).

(Còn tiếp)

Vương Quân Hoàng [1], Lã Việt Phương [1], Nguyễn Hồng Sơn [2], Nguyễn Minh Hoàng [1]

[1] Trung tâm nghiên cứu ISR, Trường Đại học Phenikaa

[2] Nguyên Chánh văn phòng, Hội đồng Lý luận Trung ương

Tài liệu tham khảo

Anh, N. (2024). *Hướng tới mục tiêu "cường quốc an toàn thông tin mạng"*. VnEconomy. Retrieved March 19 from https://vneconomy.vn/huong-toi-muc-tieu-cuong-quoc-an-toan-thong-tin-mang.htm

Bierens, R., Klievink, B., & van Den Berg, J. (2017). A social cyber contract theory model for understanding national cyber strategies. Electronic Government: 16th IFIP WG 8.5 International Conference, St. Petersburg, Russia.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., . . . Roli, F. (2013). Evasion attacks against machine learning at test time. Machine Learning and Knowledge Discovery in Databases: European Conference, Prague, Czech Republic.

Biggio, B., Fumera, G., & Roli, F. (2013). Security evaluation of pattern classifiers under attack. *IEEE Transactions on Knowledge and Data Engineering*, *26*(4), 984-996. https://doi.org/10.1109/TKDE.2013.57

Boucher, D., & Kelly, P. (2003). *The social contract from Hobbes to Rawls*. Routledge.

Chính, P. M., & Hoàng, V. Q. (2009). *Kinh tế Việt Nam: Thăng trầm và đột phá*. Nxb Chính trị quốc gia-Sự thật.

Clifford, C. (2018). *Google CEO: A.I. is more important than fire or electricity*. CNBC. Retrieved March 18 from https://www.cnbc.com/2018/02/01/google-ceo-sundar-pichai-ai-is-more-important-than-fire-electricity.html

Eiden, K., Kaplan, J., Kazimierski, B., Lewis, C., & Telford, K. (2021). *Organizational cyber maturity: A survey of industries*. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/organizational-cyber-maturity-a-survey-of-industries

Green, L. (1998). Power. In *Routledge Encyclopedia of Philosophy*: Taylor and Francis.

Greenberg, A. (2015). *Hackers remotely kill a Jeep on the highway—With me in it*. WIRED. Retrieved March 17 from https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/

Handa, A., Sharma, A., & Shukla, S. K. (2019). Machine learning in cybersecurity: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1306. https://doi.org/10.1002/widm.1306

Henshall, W. (2023). *4 charts that show why AI progress is unlikely to slow down*. Time. Retrieved March 18 from https://time.com/6300942/ai-progress-charts/

Ho, M.-T., & Vuong, Q.-H. (2023). Disengage to survive the AI-powered sensory overload world. *AI and Society*. https://doi.org/10.1007/s00146-023-01714-0

Hobbes, T. (1894). *Leviathan: Or, the matter, form, and power of a commonwealth ecclesiastical and civil* (Vol. 21). G. Routledge and sons.

Hu, K. (2023). *ChatGPT sets record for fastest-growing user base - analyst note*. Reuters. Retrieved 11 May from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Insiders, C. (2023). *2023 insider threat report*. https://istari-global.com/insights/spotlight/2023-insider-threat-report/

Keck, M., Gillani, S., Dermish, A., Grossman, J., & Rühmann, F. (2022). *The role of cybersecurity and data security in the digital economy*. https://policyaccelerator.uncdf.org/all/brief-cybersecurity-digital-economy

Kramer, M. H. (2008). *The quality of freedom*. Oxford University Press.

Lappin, S. (2023). Assessing the strengths and weaknesses of Large Language Models. *Journal of Logic, Language and Information*, *33*, 9-20. https://doi.org/10.1007/s10849-023-09409-x

Liaropoulos, A. (2020). A social contract for cyberspace. *Journal of Information Warfare*, *19*(2), 1-11. https://www.jstor.org/stable/27033617

Locke, J. (1967). *Two treatises of government*. Cambridge university press.

Lu, Y., & Da Xu, L. (2018). Internet of Things (IoT) cybersecurity research: A review of current research topics. *IEEE Internet of Things Journal*, *6*(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Maalem Lahcen, R. A., Caulkins, B., Mohapatra, R., & Kumar, M. (2020). Review and insight on the behavioral aspects of cybersecurity. *Cybersecurity*, *3*, 1-18. https://doi.org/10.1186/s42400-020-00050-w

Mantello, P., Ho, M.-T., Nguyen, M.-H., & Vuong, Q.-H. (2023). Machines that feel: behavioral determinants of attitude towards affect recognition technology—upgrading technology acceptance theory with the mindsponge model. *Humanities and Social Sciences Communications*, *10*, 430. https://doi.org/10.1057/s41599-023-01837-1

Marr, B. (2021). *How much data do we create every day? The mind-blowing stats everyone should read*. Bernard Marr & Co. Retrieved March 18 from https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Martínez Torres, J., Iglesias Comesaña, C., & García-Nieto, P. J. (2019). Machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, *10*(10), 2823-2836. https://doi.org/10.1007/s13042-018-00906-1

Morgan, S. (2022). *Cybercrime to cost the world 8 trillion annually in 2023*. Cybercrime Magazine. Retrieved March 18 from https://cybersecurityventures.com/cybercrime-to-cost-the-world-8-trillion-annually-in-2023/

Musa, S. (2018). Smart cities-a road map for development. *IEEE Potentials*, *37*(2), 19-23. https://doi.org/10.1109/MPOT.2016.2566099

Nash-Hoff, M. (2012). *What does the economy have to do with national security?* IndustryWeek. Retrieved March 19 from https://www.industryweek.com/finance/software-systems/article/21954333/what-does-the-economy-have-to-do-with-national-security

Nguyen, M.-H., Le, T.-T., & Vuong, Q.-H. (2023). Ecomindsponge: A novel perspective on human psychology and behavior in the ecosystem. *Urban Science*, *7*(1), 31. https://doi.org/10.3390/urbansci7010031

Nilekani, N. (2018). Data to the people: India's inclusive internet. *Foreign Affairs*, *97*(5), 19-27.

Novet, J. (2023). *Microsoft introduces an A.I. chatbot for cybersecurity experts*. CNBC. Retrieved March 18 from https://www.cnbc.com/2023/03/28/microsoft-launches-security-copilot-in-private-preview.html

Novet, J. (2024). *Microsoft says new AI security chatbot pricing model lets customers 'buy what they need'*. CNBC. Retrieved March 18 from https://www.cnbc.com/2024/03/13/microsoft-uses-compute-units-to-charge-customers-for-security-copilot.html

Okhrimenko, I., Stepenko, V., Chernova, O., & Zatsarinnaya, E. (2023). The impact of information sphere in the economic security of the country: case of Russian realities. *Journal of Innovation and Entrepreneurship*, *12*(1), 67. https://doi.org/10.1186/s13731-023-00326-8

Paiho, S., Tuominen, P., Rökman, J., Ylikerälä, M., Pajula, J., & Siikavirta, H. (2022). Opportunities of collected city data for smart cities. *IET Smart Cities*, *4*(4), 275–291. https://doi.org/10.1049/smc2.12044

Pansardi, P. (2012). Power and freedom: opposite or equivalent concepts? *Theoria*, *59*(132), 26-44. https://www.jstor.org/stable/41802526

Payne, B. K., & Hadzhidimova, L. (2018). Cyber security and criminal justice programs in the United States: Exploring the intersections. *International Journal of Criminal Justice Sciences*, *13*(2). https://doi.org/10.5281/zenodo.2657646

Rao, Vikram Singh. (2021). *Best AI-based cyber security tools for improved safety*. Echnotification. Retrieved March 19 from https://www.technotification.com/2021/06/best-ai-based-cyber-security-tools.html

Rieck, K., Trinius, P., Willems, C., & Holz, T. (2011). Automatic analysis of malware behavior using machine learning. *Journal of Computer Security*, *19*(4), 639-668. https://doi.org/10.5555/2011216.2011217

RiskXchange. (2023). *Cybersecurity statistics you should know in 2023*. RiskXchange. Retrieved March 19 from https://riskxchange.co/1006415/cybersecurity-statistics-2023/

Rousseau, J.-J. (2016). The social contract. In R. Blaug & J. Schwarzmantel (Eds.), *Democracy: A Reader* (pp. 43-51). Columbia University Press.

Shadmy, T. (2019). The new social contract: Facebook's community and our rights. *Boston University International Law Journal*, *37*, 307.

Sharma, A., & Sahay, S. K. (2014). Evolution and detection of polymorphic and metamorphic malwares: A survey. *International Journal of Computer Applications*, *90*(2), 7–11. https://doi.org/10.48550/arXiv.1406.7061

Sơn, M. (2023). *An ninh mạng: Những xu hướng đáng chú ý trong 6 tháng cuối năm 2023*. VietnamPlus. Retrieved March 19 from https://www.vietnamplus.vn/an-ninh-mang-nhung-xu-huong-dang-chu-y-trong-6-thang-cuoi-nam-2023-post869804.vnp

Stacey, K., & Milmo, D. (2023). *AI developing too fast for regulators to keep up, says Oliver Dowden*. The Guardian. Retrieved March 18 from https://www.theguardian.com/technology/2023/sep/22/ai-developing-too-fast-for-regulators-to-keep-up-oliver-dowden

Stanton, J. M., Stam, K. R., Mastrangelo, P., & Jolton, J. (2005). Analysis of end user security behaviors. *Computers and Security*, *24*(2), 124-133. https://doi.org/10.1016/j.cose.2004.07.001

Suleyman, M. (2023). *How the AI revolution will reshape the world*. Time. Retrieved March 18 from https://time.com/6310115/ai-revolution-reshape-the-world/

Tạp chí An toàn thông tin. (2023). *An toàn thông tin 10 dấu ấn nổi bật trong lĩnh vực bảo mật và an ninh, an toàn thông tin tại Việt Nam năm 2023*. Trung tâm Công nghệ Thông tin và Truyền thông Nghệ An. Retrieved March 19 from https://naict.tttt.nghean.gov.vn/attt/an-toan-thong-tin-10-dau-an-noi-bat-trong-linh-vuc-bao-mat-va-an-ninh-an-toan-thong-tin-tai-viet-nam-nam-2023-597.html

Vailshery, L. S. (2023). *Number of IoT connected devices worldwide 2019-2023, with forecasts to 2030*. Statista. Retrieved March 18 from https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

Vanian, J., & Leswing, K. (2023). *ChatGPT and generative AI are booming, but the costs can be extraordinary*. CNBC. Retrieved March 18 from https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html

Vuong, Q.-H. (2018). The (ir) rational consideration of the cost of science in transition economies. *Nature Human Behaviour*, *2*(1), 5.

Vuong, Q.-H. (2023). *Mindsponge Theory*. Walter de Gruyter GmbH. https://www.amazon.com/dp/B0C3WHZ2B3/

Vuong, Q.-H., *et al*. (2019). Artificial intelligence vs. natural stupidity: Evaluating AI readiness for the vietnamese medical information system. *Journal of Clinical Medicine*, *8*(2), 168. https://doi.org/10.3390/jcm8020168

Vuong, Q.-H., *et al*. (2023a). AI's humanoid appearance can affect human perceptions of Its emotional capability: Evidence from self-reported data in the US. *International Journal of Human–Computer Interaction*, 1-12. https://doi.org/10.1080/10447318.2023.2227828

Vuong, Q.-H., *et al*. (2023b). How AI's self-prolongation influences people's perceptions of its autonomous mind: The case of US residents. *Behavioral Sciences*, *13*(6), 470. https://doi.org/10.3390/bs13060470

Vuong, Q.-H., *et al*. (2023). Are we at the start of the artificial intelligence era in academic publishing? *Science Editing*, *10*(2), 158-164. https://doi.org/10.6087/kcse.310

Vuong, Q.-H., Nguyen, M.-H., & La, V.-P. (2022). *The mindsponge and BMF analytics for innovative thinking in social sciences and humanities*. Walter de Gruyter GmbH. https://www.amazon.com/dp/B0C4ZK3M74/

Wise, J. (2023). *How many Google searches per minute in 2024?* EarthWeb. Retrieved March 18 from https://earthweb.com/how-many-google-searches-per-minute/

World Economic Forum. (2023). *The global risks report 2023*. https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf