BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

BÁO CÁO SEMINAR

TÊN ĐỀ TÀI

NHỮNG THÀNH TỰU, LỢI ÍCH VÀ RỦI RO CỦA TRÍ TUỆ NHÂN TẠO

Người thực hiện: PGS. TS. Lê Văn Hưng (Mã cán bộ: 0801-01)

Đơn vị: Bộ môn Công nghệ phần mềm
Khoa Công nghệ Thông tin

Hà Nội - 2024

MỤC LỤC

1. Mở đầu

Trí tuệ nhân tạo (AI) đang ngày càng trở thành một yếu tố quan trọng trong cuộc sống hiện đại, tác động mạnh mẽ đến nhiều lĩnh vực như công nghệ, y tế, giao thông và giáo dục. Bài viết này sẽ khám phá tình hình hiện tại và những thành tựu nổi bật của AI, từ các báo cáo của nghiên cứu Một trăm năm về AI của Đại học Stanford (AI100). Chúng ta sẽ tìm hiểu cách AI được triển khai, những tiến bộ đáng chú ý và các thách thức mà xã hội đang đối mặt. Đồng thời, bài viết cũng sẽ đề cập đến những lợi ích to lớn mà AI mang lại, như tăng hiệu quả, cải thiện dịch vụ y tế và giao thông thông minh, cùng với các rủi ro tiềm ẩn như mất việc làm, thiên vị trong quyết định và vấn đề an ninh mạng. Qua đó, chúng ta sẽ có cái nhìn toàn diện về sự phát triển và tương lai của AI, cũng như những biện pháp cần thiết để quản lý và tận dụng công nghệ này một cách hiệu quả và bền vững.

2. Tình hình hiện tại và những thành tựu của Trí tuệ nhân tạo

2.1 Tình hình hiện tại

Nghiên cứu Một trăm năm về AI của Đại học Stanford (AI100) tổ chức các hội đồng chuyên gia để cung cấp báo cáo về tình hình hiện tại của AI. Báo cáo năm 2016 của họ kết luận rằng trong tương lai sẽ có sự gia tăng đáng kể trong việc sử dụng các ứng dụng AI như xe tự lái, chẩn đoán và điều trị y tế, và hỗ trợ chăm sóc người già. Xã hội hiện đang ở ngã rẽ quan trọng trong việc quyết định cách triển khai công nghệ AI theo cách thúc đẩy các giá trị dân chủ như tự do, bình đẳng và minh bạch. AI100 cũng sản xuất một Chỉ số AI tại aiindex.org để giúp theo dõi tiến độ. Dưới đây là một số điểm nổi bật từ các báo cáo năm 2018 và 2019:

• Số lượng công bố: Số lượng bài báo AI tăng gấp 20 lần từ năm 2010 đến 2019 lên khoảng 20.000 bài mỗi năm, với học máy là lĩnh vực phổ biến nhất. Số lượng bài báo học máy trên arXiv.org tăng gấp đôi mỗi năm từ 2009 đến 2017. Thị giác máy tính và xử lý ngôn ngữ tự nhiên là các lĩnh vực phổ biến tiếp theo.

• Quan điểm công chúng: Khoảng 70% bài báo tin tức về AI là trung lập, nhưng số bài báo tích cực tăng từ 12% vào năm 2016 lên 30% vào năm

2018. Các vấn đề phổ biến nhất liên quan đến đạo đức, quyền riêng tư dữ liệu và thiên vị thuật toán.

• Số lượng sinh viên: Số lượng sinh viên đăng ký học AI ở Mỹ tăng gấp 5 lần và trên toàn thế giới tăng 16 lần so với năm 2010. AI trở thành chuyên ngành phổ biến nhất trong Khoa học Máy tính.

• Đa dạng giới tính: Trên toàn thế giới, khoảng 80% giáo sư AI là nam và 20% là nữ, tỷ lệ tương tự cũng được ghi nhận ở sinh viên tiến sĩ và nhân viên trong ngành.

• Tham gia hội nghị: Số lượng tham dự hội nghị NeurIPS tăng 800% kể từ năm 2012 lên 13.500 người. Các hội nghị khác có mức tăng trưởng hàng năm khoảng 30%.

• Công nghiệp: Số lượng startup AI ở Mỹ tăng gấp 20 lần lên hơn 800.

• Quốc tế hóa: Trung Quốc xuất bản nhiều bài báo hơn Mỹ mỗi năm và gần bằng toàn bộ châu Âu. Tuy nhiên, các tác giả Mỹ có ảnh hưởng hơn 50% so với các tác giả Trung Quốc. Singapore, Brazil, Australia, Canada, và Ấn Độ là các quốc gia tăng trưởng nhanh nhất về số lượng nhân viên AI.

• Thị giác máy tính: Tỷ lệ lỗi trong phát hiện đối tượng giảm từ 28% năm 2010 xuống còn 2% năm 2017, vượt qua hiệu suất của con người. Độ chính xác trong trả lời câu hỏi hình ảnh mở cải thiện từ 55% lên 68% kể từ năm 2015, nhưng vẫn kém hơn so với con người ở mức 83%.

• Tốc độ huấn luyện: Thời gian huấn luyện cho tác vụ nhận dạng hình ảnh giảm đi 100 lần chỉ trong hai năm qua. Số lượng sức mạnh tính toán sử dụng trong các ứng dụng AI hàng đầu đang tăng gấp đôi mỗi 3,4 tháng.

• Ngôn ngữ tự nhiên: Độ chính xác trong trả lời câu hỏi trên Bộ dữ liệu Trả lời Câu hỏi Stanford (SQUAD) tăng từ 60 lên 95 từ năm 2015 đến 2019; trên biến thể SQUAD 2, tiến độ tăng nhanh hơn, từ 62 lên 90 chỉ trong một năm. Cả hai điểm số đều vượt qua hiệu suất của con người.

• Hiệu suất cấp độ con người: Đến năm 2019, các hệ thống AI đã đạt hoặc vượt qua hiệu suất của con người trong nhiều lĩnh vực như cờ vua, cờ vây, poker, Pac-Man, Jeopardy!, phát hiện đối tượng ImageNet, nhận dạng

giọng nói, dịch từ tiếng Trung sang tiếng Anh, Quake III, Dota 2, StarCraft II, các trò chơi Atari, phát hiện ung thư da, ung thư tuyến tiền liệt, gấp protein và chẩn đoán bệnh võng mạc tiểu đường.

Khi nào (nếu có) các hệ thống AI sẽ đạt được hiệu suất cấp độ con người trên một loạt các tác vụ rộng lớn? Ford (2018) phỏng vấn các chuyên gia AI và nhận thấy một loạt các năm mục tiêu, từ 2029 đến 2200, với trung bình là năm 2099. Trong một khảo sát tương tự (Grace et al., 2017) 50% số người được hỏi nghĩ rằng điều này có thể xảy ra vào năm 2066, mặc dù 10% nghĩ rằng điều này có thể xảy ra sớm nhất vào năm 2025, và một số ít cho rằng "không bao giờ". Các chuyên gia cũng chia rẽ về việc chúng ta cần các đột phá cơ bản mới hay chỉ cần tinh chỉnh các phương pháp hiện tại. Nhưng đừng quá coi trọng những dự đoán của họ; như Philip Tetlock (2017) chứng minh trong lĩnh vực dự đoán các sự kiện thế giới, các chuyên gia không giỏi hơn so với người không chuyên.

2.2 Những thành tựu của các hệ thống AI hiện tại

2.2.1 Phương tiện giao thông tự động

Lịch sử của các phương tiện giao thông tự động bắt đầu từ những chiếc xe điều khiển từ xa của những năm 1920, nhưng các cuộc biểu diễn đầu tiên về xe tự lái trên đường diễn ra vào những năm 1980. Sau các cuộc biểu diễn thành công về lái xe trên đường đất trong Cuộc thi thách thức DARPA Grand Challenge dài 132 dặm vào năm 2005 và trên đường phố có giao thông trong Cuộc thi thách thức đô thị năm 2007, cuộc đua phát triển xe tự lái bắt đầu thực sự. Năm 2018, các xe thử nghiệm của Waymo đã vượt qua cột mốc 10 triệu dặm lái xe trên đường công cộng mà không gặp tai nạn nghiêm trọng, với việc tài xế con người chỉ can thiệp để kiểm soát xe sau mỗi 6.000 dặm. Không lâu sau, công ty bắt đầu cung cấp dịch vụ taxi tự lái thương mại.

Trong không gian, các drone cánh cố định tự động đã cung cấp các chuyến giao máu xuyên quốc gia ở Rwanda từ năm 2016. Các thiết bị bay bốn cánh quạt thực hiện các động tác nhào lộn đáng chú ý, khám phá các tòa nhà trong khi xây dựng bản đồ 3D, và tự lắp ráp thành các đội hình tự động.

2.2.2 Máy móc có chân

BigDog, một robot bốn chân của Raibert et al. (2008), đã thay đổi cách nhìn của chúng ta về cách robot di chuyển—không còn là dáng đi chậm, cứng nhắc, mà là một thứ gì đó giống động vật hơn và có thể phục hồi khi bị đẩy hoặc khi trượt trên vũng nước đá. Atlas, một robot hình người, không chỉ đi trên địa hình không bằng phẳng mà còn nhảy lên hộp và thực hiện các cú lộn ngược (Ackerman và Guizzo, 2016).

2.2.3 Lập kế hoạch và lập lịch tự động

Nằm cách Trái đất hàng trăm triệu dặm, chương trình Remote Agent của NASA trở thành chương trình lập kế hoạch tự động đầu tiên điều khiển việc lập lịch các hoạt động cho một tàu vũ trụ (Jonsson et al., 2000). Remote Agent tạo ra các kế hoạch từ các mục tiêu cấp cao được chỉ định từ mặt đất và giám sát việc thực hiện các kế hoạch đó—phát hiện, chẩn đoán, và khắc phục các vấn đề khi chúng xảy ra. Hiện nay, bộ công cụ lập kế hoạch EUROPA được sử dụng cho các hoạt động hàng ngày của các rover trên Sao Hỏa của NASA và hệ thống SEXTANT cho phép điều hướng tự động trong không gian sâu, vượt ra ngoài hệ thống GPS toàn cầu.

Trong cuộc khủng hoảng Vịnh Ba Tư năm 1991, lực lượng Hoa Kỳ đã triển khai Công cụ Phân tích và Lập lại Kế hoạch Động, DART (Cross và Walker, 1994), để thực hiện lập kế hoạch hậu cần tự động và lập lịch cho việc vận chuyển. Điều này liên quan đến việc điều phối lên tới 50.000 phương tiện, hàng hóa và con người cùng một lúc, và phải tính đến các điểm xuất phát, đích đến, tuyến đường, khả năng vận chuyển, khả năng cảng và sân bay, và giải quyết xung đột giữa tất cả các tham số. DARPA tuyên bố rằng ứng dụng duy nhất này đã hơn trả lại khoản đầu tư 30 năm của họ vào AI.

Hàng ngày, các công ty gọi xe như Uber và các dịch vụ bản đồ như Google Maps cung cấp hướng dẫn lái xe cho hàng trăm triệu người dùng, nhanh chóng lập kế hoạch tuyến đường tối ưu dựa trên điều kiện giao thông hiện tại và dự đoán.

2.2.4 Dịch tự động

Các hệ thống dịch máy trực tuyến hiện nay cho phép đọc tài liệu bằng hơn 100 ngôn ngữ, bao gồm ngôn ngữ bản địa của hơn 99% dân số thế giới, và xử lý hàng trăm tỷ từ mỗi ngày cho hàng trăm triệu người dùng. Mặc dù không hoàn hảo, chúng thường đủ để hiểu nội dung. Đối với các ngôn ngữ có liên quan chặt chẽ và

có nhiều dữ liệu huấn luyện (như tiếng Pháp và tiếng Anh), các bản dịch trong một lĩnh vực hẹp gần đạt trình độ con người.

2.2.5 Nhận dạng giọng nói

Năm 2017, Microsoft cho thấy rằng Hệ thống Nhận Dạng Giọng Nói Hội Thoại của họ đã đạt tỷ lệ lỗi 5,1%, bằng với hiệu suất của con người trong tác vụ Switchboard, liên quan đến việc chuyển đổi các cuộc trò chuyện điện thoại. Khoảng một phần ba tương tác máy tính trên toàn thế giới hiện nay được thực hiện bằng giọng nói thay vì bàn phím; Skype cung cấp dịch giọng nói thời gian thực trong mười ngôn ngữ. Alexa, Siri, Cortana và Google cung cấp các trợ lý có thể trả lời câu hỏi và thực hiện các nhiệm vụ cho người dùng; ví dụ, dịch vụ Google Duplex sử dụng nhận dạng giọng nói và tổng hợp giọng nói để đặt chỗ nhà hàng cho người dùng, thực hiện cuộc trò chuyện lưu loát thay mặt họ.

2.2.6 Các hệ thống gợi ý

Các công ty như Amazon, Facebook, Netflix, Spotify, YouTube, Walmart và các công ty khác sử dụng học máy để đề xuất những gì bạn có thể thích dựa trên kinh nghiệm quá khứ của bạn và của những người khác giống bạn. Lĩnh vực hệ thống gợi ý có một lịch sử lâu dài nhưng đang thay đổi nhanh chóng do các phương pháp học sâu mới phân tích nội dung (văn bản, âm nhạc, video) cũng như lịch sử và siêu dữ liệu. Lọc thư rác cũng có thể được coi là một dạng gợi ý (hoặc không khuyến nghị); các kỹ thuật AI hiện tại lọc ra hơn 99,9% thư rác, và các dịch vụ email cũng có thể gợi ý người nhận tiềm năng, cũng như văn bản phản hồi có thể có.

2.2.7 Chơi game

Khi Deep Blue đánh bại nhà vô địch cờ vua Garry Kasparov vào năm 1997, những người bảo vệ ưu thế của con người đã đặt hy vọng vào trò chơi cờ vây. Piet Hut, một nhà vật lý thiên văn và người đam mê cờ vây, dự đoán rằng sẽ mất "một trăm năm trước khi một máy tính đánh bại con người ở cờ vây—có thể còn lâu hơn nữa." Nhưng chỉ 20 năm sau, ALPHAGO đã vượt qua tất cả các người chơi con người. Ke Jie, nhà vô địch thế giới, nói rằng "Năm ngoái, nó vẫn còn khá giống con người khi chơi. Nhưng năm nay, nó trở thành một vị thần của cờ vây." ALPHAGO hưởng lợi từ việc nghiên cứu hàng trăm ngàn trò chơi cờ vây trước đây của người chơi con người và từ kiến thức chuyên môn của các người chơi cờ vây làm việc trong nhóm.

Chương trình tiếp theo, ALPHAZERO, không sử dụng đầu vào từ con người (ngoại trừ các quy tắc của trò chơi), và đã có thể học thông qua tự chơi một mình để đánh bại tất cả các đối thủ, con người và máy, ở cờ vây, cờ vua và shogi. Trong khi đó, các nhà vô địch con người đã bị đánh bại bởi các hệ thống AI trong các trò chơi đa dạng như Jeopardy!, poker, và các trò chơi video Dota 2, StarCraft II, và Quake III.

2.2.8 Hiểu biết về hình ảnh

Không hài lòng với việc vượt qua độ chính xác của con người trong tác vụ nhận dạng đối tượng ImageNet đầy thử thách, các nhà nghiên cứu thị giác máy tính đã tiếp cận vấn đề khó hơn về chú thích hình ảnh. Một số ví dụ ấn tượng bao gồm "Một người đang đi xe máy trên con đường đất," "Hai chiếc pizza đang ngồi trên đầu bếp lò nướng," và "Một nhóm người trẻ đang chơi trò chơi frisbee". Tuy nhiên, các hệ thống hiện tại vẫn còn xa sự hoàn hảo: ví dụ, chú thích một "tủ lạnh đầy thức ăn và đồ uống" hóa ra là về một biển báo cấm đỗ xe bị che khuất bởi nhiều nhãn dán nhỏ.

2.2.9 Y tế

Các thuật toán AI hiện nay ngang bằng hoặc vượt qua các bác sĩ chuyên gia trong việc chẩn đoán nhiều bệnh, đặc biệt khi chẩn đoán dựa trên hình ảnh. Các ví dụ bao gồm bệnh Alzheimer, ung thư di căn, bệnh mắt, và bệnh da. Một đánh giá hệ thống và phân tích tổng hợp cho thấy hiệu suất của các chương trình AI, trung bình, tương đương với các chuyên gia y tế. Một nhấn mạnh hiện tại trong AI y tế là trong việc tạo điều kiện cho các hợp tác con người-máy móc. Ví dụ, hệ thống LYNA đạt được 99,6% độ chính xác trong chẩn đoán ung thư vú di căn—tốt hơn so với một chuyên gia con người không được hỗ trợ—nhưng kết hợp vẫn tốt hơn.

Việc chấp nhận rộng rãi các kỹ thuật này hiện bị giới hạn không phải bởi độ chính xác chẩn đoán mà bởi nhu cầu chứng minh cải thiện trong kết quả lâm sàng và đảm bảo tính minh bạch, không thiên vị và quyền riêng tư dữ liệu. Năm 2017, chỉ có hai ứng dụng AI y tế được FDA chấp thuận, nhưng con số này đã tăng lên 12 vào năm 2018, và tiếp tục tăng.

2.2.10 Khoa học khí hậu

Một nhóm các nhà khoa học đã giành giải Gordon Bell 2018 cho mô hình học sâu phát hiện thông tin chi tiết về các hiện tượng thời tiết cực đoan mà trước đây bị chôn vùi trong dữ liệu khí hậu. Họ đã sử dụng một siêu máy tính với phần cứng

GPU chuyên dụng để vượt qua mức exaop (hoạt động trên giây), chương trình học máy đầu tiên đạt được điều này. Rolnick et al. trình bày một danh mục dài 60 trang về các cách mà học máy có thể được sử dụng để đối phó với biến đổi khí hậu.

Đây chỉ là một vài ví dụ về các hệ thống trí tuệ nhân tạo tồn tại ngày nay. Không phải ma thuật hay khoa học viễn tưởng—mà là khoa học, kỹ thuật và toán học, mà cuốn sách này cung cấp một giới thiệu.

3. Những rủi ro của Trí tuệ nhân tạo

Francis Bacon, một triết gia được công nhận là người tạo ra phương pháp khoa học, đã nhận xét trong "The Wisdom of the Ancients" (1609) rằng "các nghệ thuật cơ học có tác dụng mơ hồ, vừa có thể gây hại vừa có thể chữa trị." Khi AI ngày càng đóng vai trò quan trọng trong các lĩnh vực kinh tế, xã hội, khoa học, y tế, tài chính và quân sự, chúng ta nên cân nhắc đến những tổn hại và lợi ích mà nó có thể mang lại—theo ngôn ngữ hiện đại, các rủi ro và lợi ích. Các chủ đề tóm tắt ở đây được đề cập chi tiết hơn trong các Chương 27 và 28.

Để bắt đầu với những lợi ích: nói một cách đơn giản, toàn bộ nền văn minh của chúng ta là sản phẩm của trí tuệ con người. Nếu chúng ta có quyền truy cập vào trí tuệ máy móc vượt trội, ngưỡng của tham vọng của chúng ta được nâng lên đáng kể. Tiềm năng của AI và robot để giải phóng nhân loại khỏi công việc lặp đi lặp lại và tăng đáng kể sản xuất hàng hóa và dịch vụ có thể báo trước một kỷ nguyên của hòa bình và thịnh vượng. Khả năng tăng tốc nghiên cứu khoa học có thể dẫn đến các phương pháp chữa bệnh và giải pháp cho biến đổi khí hậu và thiếu hụt tài nguyên. Như Demis Hassabis, CEO của Google DeepMind, đã gợi ý: "Đầu tiên giải quyết AI, sau đó sử dụng AI để giải quyết mọi thứ khác."

Trước khi chúng ta có cơ hội “giải quyết AI”, nghĩa là hoàn thiện các vấn đề liên quan đến Trí tuệ nhân tạo, chúng ta sẽ phải đối mặt với rủi ro từ việc sử dụng AI sai cách, một cách cố ý hoặc không. Một số rủi ro này đã rõ ràng, trong khi những rủi ro khác có khả năng xảy ra dựa trên các xu hướng hiện tại.

3.1 Vũ khí giết người tự động

Đây là các loại vũ khí được Liên Hợp Quốc định nghĩa là vũ khí có thể xác định, chọn và loại bỏ mục tiêu con người mà không cần can thiệp của con người. Mối quan tâm chính với các loại vũ khí này là khả năng mở rộng: việc không cần giám sát của con người có nghĩa là một nhóm nhỏ có thể triển khai một số lượng vũ khí không giới hạn chống lại các mục tiêu con người được xác định bởi bất kỳ tiêu chí nhận dạng nào có thể thực hiện được. Các công nghệ cần thiết cho vũ khí tự động

tương tự như các công nghệ cần thiết cho xe tự lái. Các cuộc thảo luận chuyên gia không chính thức về các rủi ro tiềm ẩn của vũ khí tự động chết người bắt đầu tại Liên Hợp Quốc vào năm 2014, chuyển sang giai đoạn tiền hiệp ước chính thức của Nhóm Chuyên gia Chính phủ vào năm 2017.

3.2 Giám sát và thuyết phục

Mặc dù việc theo dõi các cuộc gọi điện thoại, các nguồn cấp dữ liệu camera video, email và các kênh nhắn tin khác tốn kém, tẻ nhạt và đôi khi gây tranh cãi về mặt pháp lý đối với nhân viên an ninh, AI (nhận dạng giọng nói, thị giác máy tính và hiểu ngôn ngữ tự nhiên) có thể được sử dụng một cách mở rộng để thực hiện giám sát hàng loạt các cá nhân và phát hiện các hoạt động quan tâm. Bằng cách điều chỉnh luồng thông tin cho từng cá nhân thông qua mạng xã hội, dựa trên các kỹ thuật học máy, hành vi chính trị có thể bị thay đổi và kiểm soát đến một mức độ nào đó – điều đã trở nên rõ ràng trong các cuộc bầu cử bắt đầu từ năm 2016.

3.3 Quyết định thiên vị

Việc sử dụng cẩu thả hoặc cố ý các thuật toán học máy cho các tác vụ như đánh giá đơn xin tạm tha và đơn xin vay có thể dẫn đến các quyết định thiên vị theo chủng tộc, giới tính hoặc các danh mục được bảo vệ khác. Thường thì dữ liệu bản thân phản ánh sự thiên vị phổ biến trong xã hội.

3.4 Tác động đến việc làm

Các lo ngại về việc máy móc làm mất các công việc đã tồn tại hàng thế kỷ. Tuy nhiên, câu chuyện này không đơn giản: máy móc thực hiện một số tác vụ mà con người có thể thực hiện, nhưng chúng cũng làm cho con người năng suất hơn và do đó có thể làm tăng cơ hội việc làm, và làm cho các công ty có lợi nhuận cao hơn và do đó có thể trả lương cao hơn. Chúng có thể làm cho một số hoạt động trở nên khả thi về mặt kinh tế mà nếu không sẽ không thực hiện được. Việc sử dụng chúng thường dẫn đến tăng cường sự giàu có nhưng có xu hướng chuyển sự giàu có từ lao động sang vốn, làm gia tăng sự bất bình đẳng. Các tiến bộ công nghệ trước đây—như phát minh của máy dệt cơ khí—đã dẫn đến các gián đoạn nghiêm trọng đối với việc làm, nhưng cuối cùng con người tìm được các loại công việc mới để làm. Mặt khác, có thể rằng AI sẽ thực hiện những loại công việc mới đó. Chủ đề này đang nhanh chóng trở thành một trọng tâm lớn cho các nhà kinh tế và chính phủ trên khắp thế giới.

3.5 Các ứng dụng với mức độ an toàn cao

Khi các kỹ thuật AI tiến bộ, chúng ngày càng được sử dụng trong các ứng dụng có tính chất an toàn quan trọng như lái xe và quản lý nguồn cung cấp nước của các thành phố. Các tai nạn chết người đã xảy ra và nêu bật khó khăn của việc xác minh chính thức và phân tích rủi ro thống kê cho các hệ thống được phát triển bằng cách sử dụng các kỹ thuật học máy. Lĩnh vực AI sẽ cần phát triển các tiêu chuẩn kỹ thuật và đạo đức ít nhất là tương đương với những tiêu chuẩn phổ biến trong các ngành kỹ thuật và y tế khác nơi mà mạng sống con người đang bị đe dọa.

3.6 An ninh mạng

Các kỹ thuật AI có hữu ích trong việc phòng thủ chống lại các cuộc tấn công mạng, chẳng hạn như bằng cách phát hiện các mô hình hành vi bất thường, nhưng chúng cũng sẽ góp phần làm tăng sức mạnh, khả năng tồn tại và khả năng phát tán của phần mềm độc hại. Ví dụ, các phương pháp học tăng cường đã được sử dụng để tạo ra các công cụ rất hiệu quả cho các cuộc tấn công tống tiền và lừa đảo cá nhân hóa.

Khi các hệ thống AI trở nên mạnh mẽ hơn, chúng sẽ đảm nhận nhiều vai trò xã hội mà trước đây do con người đảm nhận. Cũng như con người đã sử dụng các vai trò này trong quá khứ để gây rối, chúng ta có thể mong đợi rằng con người có thể sử dụng các hệ thống AI trong các vai trò này để gây rối thậm chí nhiều hơn. Tất cả các ví dụ được đưa ra ở trên chỉ ra tầm quan trọng của việc quản lý và, cuối cùng, đưa ra các quy định. Hiện tại, cộng đồng nghiên cứu và các tập đoàn lớn tham gia nghiên cứu AI đã phát triển các nguyên tắc tự quản lý cho các hoạt động liên quan đến AI. Các chính phủ và tổ chức quốc tế đang thiết lập các cơ quan tư vấn để đưa ra các quy định thích hợp cho từng trường hợp sử dụng cụ thể, chuẩn bị cho các tác động kinh tế và xã hội, và tận dụng các khả năng của AI để giải quyết các vấn đề xã hội lớn.

Khi các hệ thống AI được ứng dụng rộng rãi trong thế giới thực, ta cần phải xem xét một loạt các rủi ro và hậu quả đạo đức.

Trong tương lai dài hạn, chúng ta đối mặt với vấn đề khó khăn trong việc kiểm soát các hệ thống AI siêu thông minh có thể phát triển theo những cách không thể dự đoán. Giải quyết vấn đề này dường như cần thiết phải thay đổi nhận thức của chúng ta về AI.

4. TÀI LIỆU THAM KHẢO

Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach 4th Edition. Peason, 2021.