AI - LỢI & HẠI

Original title: **AI Snake Oil**: ***What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference***
Written by **Arvind Narayanan** and **Sayash Kapoor**

Tác phẩm: **AI - Lợi và Hại: Những sự thật chưa biết về trí tuệ nhân tạo**
Tác giả: **Arvind Narayanan** và **Sayash Kapoor**

Biên tập viên First News: **Phương An, Ca Dao**

Quý độc giả có nhu cầu liên hệ, vui lòng gửi email về:
Bản thảo và bản quyền: rights@firstnews.com.vn
Phát hành: triviet@firstnews.com.vn

CÔNG TY VĂN HÓA SÁNG TẠO TRÍ VIỆT – FIRST NEWS

11 I-H Nguyễn Thị Minh Khai, Phường Sài Gòn, TP. HCM
Ngôi nhà Hạt Giống Tâm Hồn, Đường Sách Nguyễn Văn Bình, Phường Sài Gòn, TP. HCM
Tel: (84.28) 38227979 – 38227980

firstnews.vn
hatgiongtamhon.vn

facebook.com/firstnewsbooks
facebook.com/hatgiongtamhon

ARVIND
NARAYANAN
&
SAYASH
KAPOOR

Phương *dịch*

LỢI & HẠI AI

SNAKE OIL

NHỮNG SỰ THẬT
CHƯA BIẾT VỀ
TRÍ TUỆ NHÂN TẠO

First News
Where there's a will - there's a way
NHÀ XUẤT BẢN DÂN TRÍ

LỜI TỰA

Đây là một quyển sách tương đối dễ đọc, dành cho tất cả mọi người đang muốn tìm hiểu thêm về trí thông minh nhân tạo (AI) và những ảnh hưởng của nó đến đời sống hiện tại.

AI - Lợi và Hại đề cập đến khả năng mà hệ thống máy tính và các công cụ của nó thực hiện những nhiệm vụ vốn dĩ đòi hỏi trí thông minh của con người, như lý luận, học hỏi, nhận diện, nghiên cứu, quyết định,… Quyển sách nêu rõ những tiềm năng đầy hứa hẹn của AI và cả những điều nó có thể làm nhưng chưa đạt được vì giới hạn về kỹ thuật và giới hạn tự nhiên.

Các tác giả đã phân tích rất cụ thể về những giới hạn của công cụ đang được giới truyền thông đề cao như một tiến bộ tuyệt vời của nhân loại này. Qua đó, chúng ta sẽ thấy được ưu và khuyết điểm của trí thông minh nhân tạo, dựa trên các bằng chứng logic. Từng chương trong sách giúp người đọc hiểu rõ hơn về nhiều loại trí thông minh nhân tạo khác nhau, như công cụ dự đoán, công cụ tạo sinh và công cụ kiểm soát dư luận. Đồng thời, một nội dung mà quyển sách tập trung là sự chiêu mãi rẻ tiền (snake oil) của công cụ dự đoán mà đa số mọi người hiện nay đều sử dụng, vì thật ra nó không dự đoán chính xác như người ta vẫn tưởng.

Nhờ vốn hiểu biết sâu rộng của các tác giả là chuyên gia trong lĩnh vực AI, người đọc sẽ được trang bị kiến thức căn bản bằng những giải thích dễ hiểu về mặt lợi và hại khi sử dụng công cụ này trong đời sống. Với các độc giả chưa quen thuộc với lịch sử hình thành của AI, quyển sách này sẽ giúp họ có được cái nhìn tổng quát về các công cụ thông minh, và biết rằng có nhiều loại trí thông minh nhân tạo khác nhau để áp dụng vào các công việc khác nhau, tránh những nhầm lẫn không đáng có.

Giáo sư **John Vu**

Giáo sư ưu tú ngành Khoa học Máy tính,
Giám đốc chương trình Đổi mới Công nghệ Sinh học và Tính toán,
đồng Giám đốc chương trình Trí tuệ nhân tạo tại Đại học Carnegie Mellon

Tặng vợ tôi, Veena

- Arvind

Tặng Vineeta Kapoor và Ravi Kapoor,
những người thầy tinh thần đầu tiên của tôi,
những người đã hướng dẫn tôi viết,
đã biên tập cho tôi và nhiều điều khác nữa.

- Sayash

Chương 1

Giới thiệu

Hãy tưởng tượng một thế giới nơi con người không có từ ngữ để chỉ các hình thức phương tiện giao thông khác nhau, chỉ có một danh từ chung duy nhất: "phương tiện". Trong thế giới ấy, "phương tiện" có thể là xe hơi, xe buýt, xe đạp, tàu vũ trụ hay bất kỳ hình thức nào để di chuyển từ điểm A đến điểm B. Không khó để hình dung rằng các cuộc trò chuyện ở đó sẽ trở nên rối rắm. Người ta tranh luận nảy lửa về việc "phương tiện" có thân thiện với môi trường hay không, trong khi một bên đang nói về xe đạp, còn bên kia đang nói về xe tải. Khi có một bước đột phá trong công nghệ tên lửa, truyền thông lại chỉ đưa tin rằng phương tiện đang ngày càng nhanh hơn đến mức nào, khiến người dân gọi đến đại lý xe hơi (à không, đại lý phương tiện) để hỏi khi nào có mẫu phương tiện mới chạy nhanh đó. Trong bối cảnh nhập nhằng như vậy, những kẻ lừa đảo dễ dàng lợi dụng sự mơ hồ của người tiêu dùng về "công nghệ phương tiện" để tung ra đủ chiêu trò gian lận, khiến thị trường càng thêm hỗn loạn.

Bây giờ, thay từ "phương tiện" bằng "trí tuệ nhân tạo" là chúng ta có ngay một bức tranh khá chính xác về thế giới mình đang sống.

Trí tuệ nhân tạo, hay AI, là thuật ngữ chung để chỉ một nhóm các công nghệ có mối liên quan không chặt chẽ với nhau. Ví dụ, ChatGPT có rất ít điểm tương đồng với phần mềm mà ngân hàng sử dụng để đánh giá người vay. Cả hai đều được gọi là AI, nhưng chúng khác nhau hoàn toàn – từ cơ chế hoạt động, mục đích sử dụng, nhóm người dùng cho đến cách chúng mắc lỗi.

Các chatbot, cùng với những công cụ tạo hình ảnh như Dall-E, Stable Diffusion và Midjourney, được xếp vào một nhóm gọi là AI tạo sinh. Loại AI này có thể tạo ra nhiều loại nội dung chỉ trong vài giây: chatbot đưa ra câu trả lời nghe có vẻ hợp lý cho các gợi ý của con người, còn trình tạo ảnh tạo ra các hình ảnh trông như ảnh chụp cho gần như mọi mô tả, chẳng hạn "một con bò trong bếp mặc áo len hồng". Một số ứng dụng khác còn có thể tạo giọng nói, thậm chí là âm nhạc.

Công nghệ AI tạo sinh đang phát triển nhanh chóng, với những tiến bộ rõ rệt và xác thực. Nhưng thật ra đây vẫn là một sản phẩm còn non trẻ, chưa đủ tin cậy và dễ bị khai thác sai mục đích. Đồng thời, sự phổ biến của nó còn đi kèm với nhiều lời quảng cáo thổi phồng, nỗi lo sợ và sự nhiễu loạn thông tin.

Trái ngược với AI tạo sinh là AI dự đoán, tạo ra các dự đoán về tương lai để hỗ trợ việc ra quyết định trong hiện tại. Trong lĩnh vực trị an, AI có thể ước lượng: "Ngày mai sẽ có bao nhiêu vụ phạm tội trong khu vực này?". Trong hoạt động quản lý kho vận, AI có thể tính toán: "Khả năng hỏng hóc của thiết bị này trong tháng tới là bao nhiêu?". Trong phạm vi tuyển dụng, AI có thể đánh giá: "Ứng viên này sẽ thể hiện ra sao nếu được tuyển dụng?".

Hiện nay, cả doanh nghiệp lẫn chính phủ đều đang sử dụng AI dự đoán, nhưng điều đó không có nghĩa là nó thực sự hiệu quả. Thật khó mà dự đoán tương lai, và AI không làm thay đổi được thực tế này. Đúng là AI sẽ phân tích dữ liệu để nhận diện các mẫu thống kê lớn – chẳng hạn, người sở hữu một công việc có xu hướng trả nợ đúng hạn – và điều này có thể hữu ích. Vấn đề là AI dự đoán thường được chào bán kèm theo những lời quảng cáo phóng đại và người ta sử dụng nó để đưa ra các quyết định ảnh hưởng đến cuộc sống và sự nghiệp của nhiều người. Đây chính là lĩnh vực mà "trò bịp AI" xuất hiện nhiều nhất.

"Trò bịp AI" là AI không và không thể hoạt động đúng như những gì được quảng cáo. Vì AI là một khái niệm đề cập đến hàng loạt công nghệ và ứng dụng khác nhau, hầu hết chúng ta chưa được trang bị đầy đủ kiến thức để phân biệt rõ đâu là loại AI có hiệu quả thực sự như hứa hẹn, và đâu chỉ là sản phẩm của chiêu trò tiếp thị. Đây là một vấn đề lớn của xã hội: để tận dụng tối đa những giá trị mà AI mang lại, đồng thời tránh được những hiểm họa tiềm ẩn – mà nhiều hiểm họa trong số đó đã và đang xảy ra – chúng ta cần có khả năng phân định rõ ràng giữa thực tế và ảo tưởng.

Quyển sách này sẽ đóng vai trò như một cẩm nang giúp bạn nhận diện "trò bịp AI" và những lời quảng cáo thổi phồng về công nghệ này. Chúng tôi sẽ cung cấp cho bạn vốn từ vựng nền tảng để phân biệt AI tạo sinh, AI dự đoán và các loại AI khác. Chúng tôi cũng sẽ chia sẻ các phương pháp thực tiễn để bạn đánh giá mức độ tin cậy của một tuyên bố nào đó về sự tiến bộ của AI. Nhờ vậy, bạn sẽ đủ khả năng tiếp cận các tin tức

về AI với sự hoài nghi có cơ sở và chú ý được những chi tiết bị che khuất. Việc hiểu sâu hơn về AI không chỉ giúp thỏa mãn trí tò mò khoa học, mà còn mang lại những ý tưởng thực tế về cách ứng dụng – hoặc tránh sử dụng – AI trong đời sống và công việc. Cuối cùng, chúng tôi sẽ lập luận rằng AI dự đoán không chỉ đang hoạt động kém hiệu quả trong hiện tại, mà còn có thể không bao giờ hiệu quả được như kỳ vọng, do những giới hạn vốn có trong việc dự đoán hành vi con người. Chúng tôi hy vọng rằng quyển sách này sẽ khiến bạn suy nghĩ về trách nhiệm của chính mình, cũng như cơ hội để tạo ra sự thay đổi, với những tác động tiêu cực của các công cụ này.

Bình minh của AI với tư cách một sản phẩm tiêu dùng

Lần đầu tiên AI tạo sinh thu hút được sự chú ý rộng rãi của công chúng là khi ChatGPT ra mắt vào tháng Mười Một năm 2022. Ban đầu, nhà phát triển OpenAI không hề quảng bá rầm rộ mà chỉ coi đó là một "bản thử nghiệm ban đầu"[1i]. Tuy nhiên, ChatGPT đã nhanh chóng trở nên phổ biến khi người dùng chia sẻ những ví dụ vừa hài hước vừa đáng kinh ngạc về khả năng của nó. Chẳng hạn, lập trình viên Thomas Ptacek đã yêu cầu viết một đoạn văn theo phong cách Kinh Thánh, bản dịch Vua James, mô tả cách lấy một chiếc bánh kẹp bơ đậu phộng ra khỏi đầu phát VCR[2]. Kết quả ChatGPT đưa ra như sau:

[i] Các chú thích đánh số tự nhiên dẫn đến *Nguồn tham khảo* ở cuối sách. Các chú thích đánh số La Mã ở chân trang là phần lưu ý thêm của tác giả, hoặc của First News (có ghi chú: FNs), để độc giả tiện theo dõi.

Và thế là một người gặp rắc rối với một chiếc bánh kẹp bơ đậu phộng, vì nó đã bị đặt vào đầu phát VCR của anh ta, và anh ta không biết làm cách nào để lấy nó ra.

Và anh ta kêu lên với Chúa rằng: "Ôi Chúa ơi, làm sao con có thể lấy chiếc bánh kẹp này ra khỏi đầu phát VCR của con, vì nó đã kẹt cứng và không thể nhúc nhích?".

Và Chúa phán cùng anh ta rằng: "Đừng sợ, con ta, vì ta sẽ hướng dẫn bàn tay con và chỉ cho con đường đúng. Hãy lấy dao trét bơ, nhẹ nhàng chèn vào giữa chiếc bánh kẹp và đầu phát VCR, rồi từ từ tách chúng ra. Với sự kiên nhẫn và quyết tâm, chiếc bánh kẹp sẽ được lấy ra, và đầu phát VCR của con sẽ được cứu".

Và người ấy làm theo lời Chúa phán truyền, và kìa, chiếc bánh kẹp đã được lấy ra khỏi đầu phát VCR, và người ấy đã được cứu.

Và Chúa phán rằng: "Ta bảo thật với con, đừng đặt những chiếc bánh kẹp bơ đậu phộng của con vào đầu phát VCR, vì đó không phải là nơi thích hợp cho những thứ như vậy. Thay vào đó, hãy để bánh kẹp của con trong tủ lạnh hoặc trên dĩa, nơi chúng thuộc về".

Và người đàn ông nghe theo lời Chúa, và từ đó, anh ta để bánh kẹp của mình ở đúng chỗ, và được cứu khỏi những rắc rối và khổ đau. Amen.

Chỉ hai tháng sau khi ra mắt, ứng dụng thông báo đã đạt hơn 100 triệu người dùng[3], một con số khiến cho chính OpenAI cũng phải bất ngờ. Công ty thậm chí không có đủ

năng lực tính toán để xử lý khối lượng truy cập khổng lồ mà ChatGPT tạo ra.

Giới lập trình viên nhanh chóng đón nhận công cụ này, vì nó tỏ ra khá hiệu quả trong việc tạo ra những đoạn mã lập trình chỉ từ một đoạn mô tả bằng ngôn ngữ. Trên thực tế, trước đó họ vẫn quen sử dụng một công cụ tên là GitHub Copilot, cũng dựa trên công nghệ tương tự, nhưng sự xuất hiện của ChatGPT đã giúp việc ứng dụng AI vào lập trình trở nên phổ biến hơn hẳn. Nó rút ngắn đáng kể thời gian phát triển các ứng dụng. Giờ đây, ngay cả những người không chuyên cũng có thể tạo ra một số ứng dụng đơn giản.

Microsoft nhanh chóng mua bản quyền công nghệ từ OpenAI và tích hợp nó vào công cụ tìm kiếm Bing. Dù đã xây dựng chatbot riêng từ năm 2021, Google vẫn chưa kịp phát hành hoặc tích hợp nó vào sản phẩm của mình[4]. Khi Bing nhập cuộc, Google buộc phải vội vàng công bố chatbot tìm kiếm có tên là Bard (sau đổi thành Gemini).

Đây là lúc mọi thứ bắt đầu trật nhịp. Trong video quảng bá Bard, chatbot này tuyên bố rằng kính viễn vọng không gian James Webb đã chụp bức ảnh đầu tiên về một hành tinh ngoài Hệ Mặt Trời. Một nhà thiên văn học đã nhanh chóng chỉ ra thông tin này là sai[5]. Có vẻ như Google thậm chí còn không thể chọn ra một ví dụ chính xác. Giá trị thị trường của Google ngay lập tức sụt giảm 100 *tỷ* đô-la, do các nhà đầu tư lo ngại rằng nếu Bard được tích hợp vào công cụ tìm kiếm này như lời hứa, chất lượng kết quả có thể giảm sút nghiêm trọng[6].

Tuy sự lúng túng của Google gây tốn kém rất nhiều, nhưng đó cũng chỉ là một gợn sóng nhỏ báo hiệu những con sóng lớn hơn sinh ra từ khó khăn cố hữu của chatbot: độ xác thực thông tin. Điểm yếu này bắt nguồn từ cách chúng được xây dựng. Chúng học các mô hình thống kê từ dữ liệu huấn luyện – chủ yếu thu thập từ Internet – sau đó tạo ra văn bản pha trộn dựa trên các mô hình đó. Nhưng vấn đề là chúng không thực sự "nhớ" thông tin đã học. Chúng ta sẽ tìm hiểu sâu hơn ở Chương 4.

Việc lạm dụng công nghệ này ngày càng lan rộng. Nhiều trang tin tức bị phát hiện cho đăng những bài viết đầy sai sót do AI tạo ra về các chủ đề quan trọng, như tư vấn tài chính, và họ vẫn tiếp tục sử dụng AI ngay cả khi những sai sót này bị phát hiện[7]. Amazon tràn ngập sách do AI viết, kể cả một số sách hướng dẫn nhận biết nấm – lĩnh vực mà lỗi sai có thể gây chết người nếu độc giả tin theo nội dung trong sách[8].

Chúng ta có thể dễ dàng nhìn vào tất cả những khiếm khuyết và việc lạm dụng chatbot này rồi kết luận rằng thế giới đang phát điên phát cuồng vì một công nghệ có khả năng thất bại rất lớn. Nhưng kết luận như vậy là quá đơn giản.

Chúng tôi tin rằng hầu hết các ngành nghề tri thức đều có thể khai thác chatbot theo cách có lợi. Bản thân chúng tôi cũng sử dụng chúng để hỗ trợ nghiên cứu, từ những tác vụ tầm thường như định dạng trích dẫn đúng cách, cho đến những công việc mà chúng tôi không thể tự làm, như giải mã một bài nghiên cứu đầy thuật ngữ chuyên môn trong một lĩnh vực xa lạ.

Vấn đề là muốn khai thác hiệu quả chatbot thì phải thực hành và nỗ lực, đồng thời tránh được những cạm bẫy thường trực của chúng. Còn sử dụng sai mục đích thì lại dễ dàng hơn nhiều – nhất là với những người muốn kiếm tiền nhanh chóng, chẳng hạn bằng cách xuất bản sách do AI tạo ra mà không màng đến chất lượng nội dung. Đó là lý do chatbot rất dễ bị lạm dụng.

Còn một vấn đề hóc búa hơn liên quan đến quyền lực. Giả sử các công cụ tìm kiếm thay thế danh sách mười liên kết truyền thống bằng các câu trả lời do AI tạo ra. Ngay cả khi giải quyết được vấn đề độ chính xác, kết quả cho ra về cơ bản vẫn là một cỗ máy viết lại nội dung sao chép từ các trang web khác, trình bày như thể đó là nội dung gốc, mà không chuyển hướng lưu lượng truy cập hay doanh thu cho các trang web kia. Nếu các công cụ tìm kiếm tự ý hiển thị nội dung của người khác mà mặc nhiên xem như của mình, họ sẽ vi phạm luật bản quyền. Nhưng cho đến nay, câu trả lời do AI tạo ra dường như vẫn lách được vấn đề này, mặc dù tính đến năm 2024, đã có nhiều vụ kiện nhằm thay đổi tình trạng nói trên[9].

AI đang làm rung chuyển ngành giải trí

Một công nghệ AI tạo sinh khác cũng đã thu hút sự chú ý mạnh mẽ là "tạo hình ảnh từ văn bản" (text-to-image). Đến giữa năm 2023, ước tính đã có hơn một tỷ hình ảnh được tạo ra bằng các công cụ như DALL-E 2 của OpenAI, Firefly của Adobe và Midjourney (thuộc về công ty cùng tên)[10]. Một công cụ

nổi bật khác là Stable Diffusion của Stability AI. Đây là phần mềm mã nguồn mở, cho phép người dùng tự do tùy chỉnh theo nhu cầu. Các công cụ dựa trên Stable Diffusion đã được tải xuống hơn 200 triệu lần. Vì người dùng chạy phần mềm trên thiết bị riêng nên không có số liệu tập trung về số lượng hình ảnh được tạo ra, nhưng con số này có thể lên đến vài tỷ.

Công cụ tạo hình ảnh đã mở ra một trận đại hồng thủy nội dung giải trí[11]. Khác với các sản phẩm giải trí truyền thống, hình ảnh do AI tạo ra có thể được cá nhân hóa vô hạn theo sở thích của mỗi người dùng. Có người thích những khung cảnh thiên nhiên hoặc đô thị kỳ ảo; có người lại say mê các bức ảnh tái hiện nhân vật lịch sử trong bối cảnh hiện đại, hoặc người nổi tiếng làm những việc khác với bình thường, ví dụ nổi tiếng là Đức Giáo hoàng mặc áo khoác phao (được đặt cho cái tên "Balenciaga Pope - Giáo hoàng Balenciaga"). Các đoạn trailer giả nhiều bộ phim nổi tiếng như *Star Wars* theo phong cách đặc trưng của Wes Anderson – với bố cục đối xứng, tông màu phấn tiên và bối cảnh siêu thực – cũng rất được ưa chuộng.

Không chỉ những người thích giải trí mới thấy phấn khích với AI tạo sinh hình ảnh: các ứng dụng giải trí này đang trở thành một ngành kinh doanh phát triển mạnh. Nhiều công ty trò chơi điện tử tạo ra những nhân vật trong game có thể trò chuyện tự nhiên với người chơi[12]. Nhiều ứng dụng chỉnh sửa ảnh hiện đã tích hợp tính năng AI tạo sinh, cho phép người dùng yêu cầu những chỉnh sửa như thêm bong bóng vào bức ảnh chụp một bữa tiệc sinh nhật.

AI cũng trở thành một vấn đề gây tranh cãi lớn trong các cuộc đình công của Hollywood vào năm 2023[13]. Giới diễn viên lo ngại hãng phim có thể sử dụng cảnh quay trước đây của họ để huấn luyện các công cụ AI, từ đó tạo ra những video mới dựa trên kịch bản – những video trông như thể họ xuất hiện trong đó, nhưng thực chất chỉ là sản phẩm của AI. Nói cách khác, hãng phim có thể tiếp tục kiếm lợi nhuận từ hình ảnh và công sức của diễn viên mà không cần trả thù lao.

Các cuộc đình công đã chấm dứt, nhưng căng thẳng giữa lao động và tư bản thì chắc chắn còn tiếp diễn, nhất là khi công nghệ sẽ còn tiến bộ[14]. Nhiều công ty đang phát triển công cụ tạo video từ văn bản, trong khi một số khác đang tìm cách tự động hóa quá trình viết kịch bản. Những sản phẩm này có thể không đạt độ phức tạp hay giá trị nghệ thuật cao, nhưng chẳng hề gì với các hãng phim đang cần tung ra một bom tấn mùa hè.

Về lâu dài, chúng tôi tin rằng sự kết hợp giữa công nghệ và luật pháp có thể giúp giảm nhẹ phần lớn các vấn đề ta đang nói đến, đồng thời khuếch đại những lợi ích. Ví dụ, có nhiều ý tưởng công nghệ hứa hẹn giúp chatbot hạn chế việc bịa đặt thông tin, trong khi các quy định pháp lý có thể đóng vai trò ngăn chặn những hành vi cố ý lạm dụng. Tuy vậy, trong ngắn hạn, việc thích nghi với một thế giới có AI tạo sinh vẫn là một thách thức thực sự, bởi lẽ các công cụ này vừa cực kỳ mạnh mẽ, vừa không đáng tin cậy. Nó giống như thể bất ngờ trao miễn phí cho cả thế giới một chiếc cưa máy mà không kèm hướng dẫn sử dụng.

Chúng ta cần nỗ lực để tích hợp AI một cách hợp lý vào đời sống. Một ví dụ điển hình là những gì đang diễn ra trong lĩnh vực giáo dục, khi AI có thể viết bài luận và vượt qua các kỳ thi đại học. Cần khẳng định rõ: AI không phải là mối đe dọa đối với giáo dục, nó chẳng khác nào máy tính cầm tay thời mới xuất hiện[15]. Nếu được giám sát đúng đắn, AI có thể trở thành một công cụ học tập quý giá. Nhưng để đạt được điều đó, giáo viên cần phải cải tổ chương trình, phương pháp giảng dạy và cả cách thức kiểm tra, đánh giá. Tại các trường có nguồn quỹ lớn như Princeton, nơi chúng tôi giảng dạy, đây là một cơ hội thay vì thách thức. Trên thực tế, chúng tôi khuyến khích sinh viên sử dụng AI. Còn ở nhiều nơi khác, ChatGPT lại đột ngột trở thành một công cụ có nguy cơ bị hàng triệu sinh viên lợi dụng để gian lận.

Liệu xã hội sẽ mãi chỉ phản ứng thụ động trước những tiến bộ mới trong lĩnh vực AI tạo sinh? Hay chúng ta có thể tập hợp đủ ý chí để thực hiện những thay đổi mang tính cấu trúc, giúp phân bổ công bằng hơn những lợi ích và chi phí cực kỳ không đồng đều từ những đổi mới này, bất kể chúng là gì?

AI dự đoán: một khẳng định phi thường cần bằng chứng phi thường

AI tạo sinh chắc chắn đi kèm với nhiều chi phí xã hội và rủi ro, nhất là trong ngắn hạn. Tuy vậy, chúng tôi vẫn lạc quan một cách thận trọng về tiềm năng lâu dài của công nghệ này trong việc cải thiện đời sống của con người. Nhưng AI dự đoán lại là một câu chuyện khác.

Trong vài năm gần đây, AI dự đoán đã được ứng dụng rộng rãi để dự báo các kết quả xã hội. Các nhà phát triển tuyên bố rằng họ có thể dự đoán hành vi trong tương lai của con người, chẳng hạn một tội nhân có tái phạm hay không, hay một ứng viên có làm tốt công việc hay không. Nhưng khác với AI tạo sinh, AI dự đoán thường không hoạt động chính xác[16].

Tại Mỹ, những người trên sáu mươi lăm tuổi đủ điều kiện đăng ký Medicare, một chương trình bảo hiểm y tế do nhà nước trợ cấp. Để cắt giảm chi phí, một số đơn vị cung cấp dịch vụ Medicare đã bắt đầu sử dụng AI để ước tính thời gian nằm viện của bệnh nhân[17]. Những dự đoán này thường thiếu chính xác. Trong một trường hợp, một bệnh nhân tám mươi lăm tuổi được đánh giá là có thể xuất viện sau mười bảy ngày. Nhưng đến thời điểm đó, bà vẫn bị đau nặng và thậm chí không thể tự dùng khung tập đi mà không có sự trợ giúp. Dù vậy, dựa trên đánh giá của AI, các khoản thanh toán bảo hiểm của bà đã bị cắt. Trên lý thuyết, công nghệ này được triển khai với ý định tốt, chẳng hạn như ngăn các cơ sở chăm sóc giữ bệnh nhân lâu hơn để thu thêm phí. Nhưng trong nhiều trường hợp, mục tiêu ban đầu cũng như cách triển khai có thể bị thay đổi theo thời gian. Rất dễ hình dung kịch bản mà ban đầu AI giúp tăng trách nhiệm cho các cơ sở chăm sóc, nhưng rồi nó dần biến thành công cụ để cắt giảm chi phí bất chấp hậu quả.

Những câu chuyện tương tự cũng đang xuất hiện ở nhiều lĩnh vực. Trong tuyển dụng, nhiều công ty AI tuyên bố có thể đánh giá các phẩm chất như thân thiện, cởi mở hay tử tế của ứng viên dựa trên ngôn ngữ cơ thể, cách nói chuyện và các đặc điểm

bề ngoài khác trong một đoạn video dài ba mươi giây. Liệu điều này có thực sự hiệu quả? Và những đánh giá này có thực sự dự đoán được năng lực làm việc không? Đáng tiếc, các công ty này chưa hề công bố bằng chứng xác thực để chứng minh sản phẩm của họ hiệu quả. Ngược lại, chúng ta có rất nhiều bằng chứng cho thấy việc dự đoán kết quả cuộc đời của một cá nhân là vô cùng khó khăn, như sẽ thảo luận ở Chương 3.

Năm 2013, công ty bảo hiểm Allstate muốn sử dụng AI dự đoán để xác định mức phí bảo hiểm tại bang Maryland, Mỹ – với mục tiêu tối đa hóa lợi nhuận mà không làm mất nhiều khách hàng. Kết quả là sự ra đời của một "danh sách nạn nhân", những người đột ngột bị tăng phí bảo hiểm so với mức phí trước đây[18]. Đáng chú ý, những người trên sáu mươi hai tuổi chiếm tỷ lệ cao bất thường trong danh sách này, thể hiện sự phân biệt đối xử tự động hóa. Có thể vì nhận thấy người lớn tuổi ít khi khảo giá thị trường, AI đã chọn ra xu hướng đó từ dữ liệu. Mức giá mới có thể giúp công ty tăng doanh thu, nhưng rõ ràng đây là một hành vi phi đạo đức. Bang Maryland đã bác bỏ đề xuất sử dụng công cụ này do tính chất phân biệt đối xử, nhưng Allstate vẫn áp dụng nó ở ít nhất mười bang khác tại Hoa Kỳ[i].

Nếu một cá nhân phản đối việc sử dụng AI trong quy trình tuyển dụng, họ có thể chọn không nộp đơn vào những công ty dùng AI để đánh giá hồ sơ ứng viên. Nhưng khi các chính phủ bắt đầu ứng dụng AI dự đoán, người dân sẽ không có lựa chọn

[i] Nhiều ví dụ trong sách, bao gồm ví dụ trên, xuất phát từ nước Mỹ đơn giản vì đó là nơi chúng tôi sinh sống và làm việc. Tuy nhiên, những bài học rút ra từ các ví dụ này được thiết kế để có tính ứng dụng rộng rãi.

nào ngoài tuân thủ. (Dù vậy, người ta cũng sẽ dấy lên những lo ngại tương tự nếu nhiều công ty cùng sử dụng một AI để quyết định tuyển dụng.) Nhiều nơi trên thế giới hiện đã sử dụng công cụ ước lượng rủi ro để quyết định có nên cho phép bị cáo tại ngoại trước phiên xét xử hay không. Các hệ thống này không chỉ bị ghi nhận là có thiên kiến – như thiên kiến về chủng tộc, giới tính hoặc độ tuổi – mà còn có vấn đề nghiêm trọng: bằng chứng cho thấy chúng chỉ chính xác hơn một chút so với việc đoán ngẫu nhiên xem một bị cáo có "rủi ro" hay không.

Một lý do khiến AI dự đoán hoạt động kém chính xác là dữ liệu không phản ánh hết các yếu tố quan trọng. Hãy xem xét ba bị cáo có cùng đặc điểm mà AI sử dụng để đánh giá: tuổi tác, số lần phạm tội trước đây và số thành viên trong gia đình có tiền án. Cả ba bị cáo này sẽ được gán cho một điểm số rủi ro như nhau. Tuy nhiên, trong ví dụ này có một người thực sự hối lỗi, một người bị cảnh sát bắt nhầm, còn người thứ ba đang nuôi ý định tiếp tục phạm tội. Không có cách nào hiệu quả để một công cụ AI cân nhắc được những khác biệt ấy.

Một nhược điểm khác của AI dự đoán là nó không hiệu quả khi đối tượng có động cơ mạnh mẽ để "lách luật". Ví dụ, AI từng được sử dụng để ước tính bệnh nhân suy thận có thể sống bao lâu sau phẫu thuật ghép thận[19]. Logic ở đây là những người có khả năng sống lâu nhất sau cấy ghép sẽ được ưu tiên nhận tạng. Trớ trêu thay, điều này lại dẫn đến một hệ quả bất ngờ: một số bệnh nhân *không muốn* duy trì chức năng thận của mình một cách tích cực, vì nếu thận của họ bị suy khi tuổi còn trẻ, họ sẽ có cơ hội được ghép tạng cao hơn. Rất may, quá trình

xây dựng hệ thống này diễn ra trong một khuôn khổ phản biện nghiêm túc, có sự tham gia của bệnh nhân, bác sĩ và các bên liên quan. Nhờ vậy, vấn đề về sự sai lệch trong động cơ hành vi đã được nhận diện kịp thời, và việc sử dụng AI dự đoán để phân bổ thận đã bị loại bỏ.

Chúng ta sẽ còn thấy nhiều ví dụ khác về thất bại của AI dự đoán ở Chương 2 và Chương 3. Liệu tình hình có thể cải thiện theo thời gian không? Đáng tiếc, chúng tôi không nghĩ vậy. Nhiều sai sót của AI dự đoán vốn là bản chất của nó. Ví dụ, AI dự đoán hấp dẫn vì sự tự động hóa giúp quá trình ra quyết định trở nên hiệu quả hơn, nhưng chính sự "hiệu quả" đó lại dẫn đến thiếu vắng trách nhiệm giải trình. Vì vậy, chúng ta cần tỉnh táo trước những tuyên bố đầy tham vọng của các công ty AI dự đoán, trừ phi chúng đi kèm với bằng chứng vững chắc.

Nhìn AI dưới một lăng kính duy nhất: dễ nhưng không chính xác

AI tạo sinh và AI dự đoán là hai loại AI chính được nhắc đến nhiều nhất hiện nay. Nhưng còn có bao nhiêu loại AI khác nữa? Không có câu trả lời rõ ràng, vì cho đến nay vẫn chưa có sự đồng thuận về điều gì được coi là AI và điều gì không.

Dưới đây là ba câu hỏi có thể giúp chúng ta xác định một hệ thống máy tính thực hiện tác vụ có được gọi là AI hay không. Mỗi câu hỏi chạm đến một khía cạnh của khái niệm AI, nhưng không câu nào cung cấp một định nghĩa đầy đủ.

Câu hỏi thứ nhất: công việc đó có đòi hỏi con người phải sáng tạo hoặc học hỏi mới thực hiện được không? Nếu có, và máy tính cũng làm được, thì có thể xem nó là AI. Điều này giải thích tại sao công cụ tạo hình ảnh được xếp vào nhóm AI. Để tạo ra một hình ảnh, con người cần có kỹ năng và quá trình rèn luyện nhất định, có thể là trong mỹ thuật hoặc thiết kế đồ họa. Ngay cả nhiệm vụ tưởng chừng đơn giản như nhận diện một con mèo hay một ấm trà trong hình – điều mà con người làm rất dễ dàng – cũng từng là một thách thức lớn trong lĩnh vực tự động hóa trước năm 2010. Việc nhận diện vật thể từ lâu đã được coi là một ứng dụng của AI, cho thấy rằng so sánh với trí thông minh con người không phải là tiêu chí duy nhất để xác định điều gì là "trí tuệ nhân tạo".

Câu hỏi thứ hai: hành vi của hệ thống đó có do mã của nhà phát triển lập trình trực tiếp không, hay nó xuất hiện gián tiếp, chẳng hạn bằng cách học từ các ví dụ hay tìm kiếm trong cơ sở dữ liệu? Nếu hành vi xuất hiện gián tiếp, hệ thống đó có thể được xem là AI. Việc học từ dữ liệu được gọi là "học máy" (machine learning), và đây là một hình thức chính của AI hiện đại. Tiêu chí này giải thích tại sao có thể coi một công thức tính giá bảo hiểm là AI nếu nó được xây dựng bằng cách cho máy tính phân tích dữ liệu yêu cầu bồi thường trước đó, nhưng không phải là AI nếu nó đơn giản là kết quả trực tiếp từ kiến thức của một chuyên gia, ngay cả khi quy tắc tính toán thực tế giống nhau trong cả hai trường hợp. Tuy nhiên, nhiều hệ thống được lập trình thủ công vẫn được coi là AI, chẳng hạn một số robot hút bụi có khả năng nhận diện và né tránh chướng ngại vật.

Câu hỏi thứ ba: hệ thống đó có khả năng đưa ra quyết định ít nhiều tự động và điều chỉnh, thích nghi ở một mức độ nào đó theo môi trường xung quanh hay không? Nếu có, nó có thể được coi là AI. Lái xe tự động là một ví dụ điển hình – nó thường được xếp vào danh mục AI vì đòi hỏi hệ thống phải xử lý linh hoạt các tình huống thay đổi liên tục trong thế giới thực. Nhưng cũng giống như các tiêu chí trước, tiêu chí này không thể tạo thành một định nghĩa trọn vẹn. Ví dụ, chúng ta không coi một chiếc nhiệt kế truyền thống không chứa các điện tử là AI, vì nó phản ứng với sự thay đổi nhiệt độ và điều chỉnh dòng điện dựa trên nguyên lý giãn nở hay co lại của kim loại.

Cuối cùng, việc một ứng dụng có được coi là AI hay không còn phụ thuộc nhiều vào lịch sử cách dùng, chiến lược tiếp thị và các yếu tố khác. Chúng ta không cần phải quá lo lắng vì thiếu một định nghĩa thống nhất. Điều này có thể khiến bạn ngạc nhiên khi đọc một quyển sách về AI. Nhưng hãy nhớ thông điệp chính của chúng tôi: không một lời khẳng định nào có thể áp dụng cho tất cả các loại AI. Trong quyển sách này, chúng tôi sẽ tập trung vào từng loại AI cụ thể, và chỉ cần mỗi loại được định nghĩa rõ ràng, chúng ta vẫn sẽ hiểu nhau.

Có một định nghĩa vui về AI rất đáng lưu ý, vì nó tiết lộ một điều quan trọng: "AI là bất cứ thứ gì chưa hoàn tất". Nói cách khác, khi một ứng dụng bắt đầu hoạt động ổn định và xác thực, nó sẽ dần trở nên bình thường và không còn được coi là AI nữa. Có rất nhiều ví dụ: robot hút bụi như Roomba, hệ thống lái tự động trên máy bay, tính năng tự động hoàn thành từ ngữ trên điện thoại, nhận diện chữ viết tay, nhận diện giọng nói,

lọc thư rác, kiểm tra chính tả,... Vâng, từng có lúc kiểm tra chính tả cũng là cả một vấn đề nan giải!

Chúng tôi nghĩ rằng những công cụ như vậy thật tuyệt vời. Chúng âm thầm cải thiện cuộc sống hằng ngày của chúng ta. Đây chính là loại AI mà chúng ta muốn có nhiều hơn nữa. Tuy nhiên, quyển sách này tập trung vào những loại AI có vấn đề theo một phương diện nào đó, bởi vì bạn có lẽ sẽ không muốn đọc vài trăm trang sách ca ngợi tính năng kiểm tra chính tả. Điều quan trọng cần hiểu là không phải tất cả AI đều có vấn đề, còn lâu mới vậy.

Hy vọng một ngày nào đó, các công nghệ AI mới sẽ dần trở nên bình thường. Ngày nay, xe tự lái thường xuất hiện trên các bản tin vì gây tai nạn và tử vong[20], nhưng việc lái xe tự động an toàn rốt cuộc cũng sẽ giải quyết được, dù mức độ khó khăn của nó thường bị đánh giá thấp. Tuy nhiên, thách thức lớn đối với xã hội không nằm ở vấn đề công nghệ, mà ở hậu quả về việc làm: nếu xe tự lái trở nên phổ biến, hàng triệu tài xế xe tải, taxi, xe công nghệ có thể bị ảnh hưởng nặng nề. Dù vậy, nếu vấn đề an toàn được giải quyết và những điều chỉnh cần thiết về mặt xã hội, chính trị được thực hiện, có thể một ngày nào đó xe tự lái sẽ trở nên bình thường như thang máy.

Ngược lại, chúng tôi cho rằng một số loại AI khác, đặc biệt là AI dự đoán, có thể sẽ không bao giờ trở nên bình thường. Dự đoán chính xác hành vi xã hội của con người không phải là vấn đề mà công nghệ có thể giải quyết, và việc định đoạt số phận của con người dựa trên những dự đoán có bản chất sai sót sẽ luôn là một vấn đề đạo đức.

Để hiểu rõ hơn tại sao chúng ta phải tránh đưa ra những kết luận chung về AI, hãy xem xét công nghệ nhận diện khuôn mặt, một công nghệ AI khiến những người ủng hộ nhân quyền lo ngại. Công nghệ này đã dẫn đến nhiều vụ bắt giữ nhầm ở Mỹ – khi quyển sách này được viết, đã có ít nhất sáu trường hợp – và tất cả nạn nhân đều là người da đen. Vậy cảnh sát có nên ngừng sử dụng nhận diện khuôn mặt vì nó dễ gây nhầm lẫn và thường xác định sai về người da đen không?

Một thực tế dễ bị bỏ qua trong cuộc tranh luận này là: tất cả các vụ bắt giữ sai đều liên quan đến hàng loạt lỗi khác của con người trong quá trình điều tra, chứ không chỉ vì công nghệ. Ví dụ, Robert Williams bị bắt vì ăn cắp ở cửa hàng dựa trên lời khai của một nhân viên bảo vệ không có mặt tại thời điểm diễn ra vụ trộm[21]. Randall Reid bị bắt ở Georgia vì một vụ trộm xảy ra ở Louisiana, nơi anh chưa từng đặt chân đến[22]. Porcha Woodruff bị bắt dựa trên một bức ảnh chụp từ năm 2015, mặc dù cảnh sát hoàn toàn có thể sử dụng ảnh trên bằng lái năm 2021 của cô để đối chiếu, nhưng đã không làm vậy[23]. Và còn nhiều vụ khác.

Những sai sót trong việc thực thi pháp luật dẫn đến bắt giữ nhầm người vẫn xảy ra mỗi ngày, và chắc chắn sẽ còn tiếp diễn bất kể có sử dụng công nghệ nhận diện khuôn mặt hay không.

Cảnh sát đã thực hiện hàng trăm ngàn lượt truy vấn bằng nhận diện khuôn mặt, vì vậy tỷ lệ sai sót của công nghệ này là rất nhỏ[24]. Thực tế, theo nghiên cứu của Viện Tiêu chuẩn và Công nghệ Quốc gia Mỹ (National Institute of Standards and

Technology - NIST), từ năm 2014 đến 2020, tỷ lệ lỗi đã giảm xuống chỉ còn 0,08% – tức giảm tới 50 lần[25].

Nếu được sử dụng đúng cách, AI nhận diện khuôn mặt thường chính xác vì nhiệm vụ này có rất ít yếu tố bất định hoặc mơ hồ. AI được huấn luyện bằng các cơ sở dữ liệu khổng lồ gồm hàng triệu bức ảnh và nhãn, giúp nó xác định xem hai bức ảnh có thuộc về cùng một người hay không. Khi đủ dữ liệu và tài nguyên tính toán, AI sẽ học được các đặc điểm giúp phân biệt khuôn mặt một cách hiệu quả. Điều này khiến nhận diện khuôn mặt khác với các tác vụ phân tích khuôn mặt khác, chẳng hạn như nhận diện giới tính hay cảm xúc, vốn thường có xu hướng mắc lỗi cao hơn nhiều[26,27]. Khác biệt quan trọng nằm ở chỗ: thông tin cần để nhận diện khuôn mặt đã thể hiện hết trên hình ảnh, còn việc nhận diện giới tính hay cảm xúc lại đòi hỏi hệ thống phải dựa trên khuôn mặt để phỏng đoán về một đặc điểm bên trong của con người, như bản dạng giới hoặc trạng thái cảm xúc, một việc vốn dĩ có bản chất khó xác định rõ ràng.

Các nhà nhân quyền thường gộp nhận diện khuôn mặt chung với các công nghệ khác có tỷ lệ sai sót cao trong hệ thống tư pháp hình sự, chẳng hạn như các mô hình dự đoán nguy cơ phạm tội, dù thật ra hai công nghệ này hoàn toàn khác nhau và có tỷ lệ mắc lỗi chênh lệch rất lớn. (Hầu hết những người bị AI dự đoán gắn nhãn là "có nguy cơ cao" thực tế không phạm tội thêm lần nào.)

Nguy cơ lớn nhất của nhận diện khuôn mặt nằm ở chỗ *nó hoạt động quá tốt*. Chính vì vậy, nếu nó rơi vào tay những kẻ

có ý đồ xấu, hậu quả có thể rất nghiêm trọng. Trong tác phẩm *Your Face Belongs to Us* (tạm dịch: Khuôn mặt bạn thuộc về chúng tôi), tác giả Kashmir Hill đã trình bày nhiều trường hợp công nghệ này bị sử dụng theo những cách cực kỳ nguy hiểm[28]. Ví dụ, một số chính phủ độc tài đã sử dụng nhận diện khuôn mặt để xác định danh tính người biểu tình ôn hòa, sau đó trả đũa họ[29].

Không chỉ chính phủ, các công ty tư nhân cũng có thể lạm dụng công nghệ này. Madison Square Garden – một địa điểm nổi tiếng chuyên tổ chức các sự kiện thể thao và hòa nhạc ở New York – từng sử dụng công nghệ này để ngăn một số luật sư vào cửa. Năm 2022, luật sư Nicolette Landi không được vào buổi hòa nhạc của Mariah Carey diễn ra tại đây[30], dù bạn trai cô đã mua vé trị giá gần 400 đô-la để tặng sinh nhật cô. Lý do? Công ty điều hành Madison Square Garden cấm tất cả luật sư làm việc cho các hãng luật từng kiện họ, kể cả những người không trực tiếp liên quan đến vụ kiện. Thậm chí, một số người trong nhóm bị cấm còn là khách hàng trung thành sở hữu vé cả mùa. Việc xác định và ngăn chặn họ được thực hiện bằng công nghệ nhận diện khuôn mặt.

Những người chỉ trích công nghệ này đôi khi phản đối với lập luận rằng nó không có hiệu quả nên phải bị cấm hoàn toàn. Họ còn lên án cả những nhà nghiên cứu đang phát triển công nghệ nhận diện khuôn mặt. Cách tiếp cận đó đã bỏ qua những lợi ích thực tế mà nó có thể mang lại. Chẳng hạn, Bộ An ninh Nội địa Hoa Kỳ từng sử dụng nhận diện khuôn mặt trong một chiến dịch kéo dài ba tuần để điều tra các vụ án

lạm dụng trẻ em chưa có lời giải. Bằng cách phân tích hình ảnh và video do tội phạm đăng tải trên mạng xã hội, công nghệ này đã giúp xác định được danh tính của hàng trăm nạn nhân và thủ phạm[31]. Tất nhiên, nó còn có nhiều ứng dụng hữu ích trong đời sống thường ngày, như mở khóa điện thoại hoặc sắp xếp ảnh vào album theo khuôn mặt.

Dù vậy, cần thừa nhận rằng tuy có độ chính xác cao trong điều kiện lý tưởng, công nghệ này vẫn dễ dàng thất bại trong thực tế. Ví dụ, khi sử dụng hình ảnh mờ nhòe từ camera giám sát thay vì ảnh rõ nét, nguy cơ nhận diện sai sẽ tăng lên đáng kể. Chuỗi nhà thuốc Rite Aid tại Hoa Kỳ từng sử dụng một hệ thống nhận diện khuôn mặt kém chất lượng, khiến nhân viên liên tục buộc tội sai khách hàng vì nghi trộm cắp. Hệ thống này đã tạo ra hàng ngàn kết quả sai và công ty cũng cố giữ kín việc đang sử dụng công nghệ đó. Rất may, các cơ quan chức năng đã vào cuộc, Ủy ban Thương mại Liên bang (Federal Trade Commission - FTC) đã cấm Rite Aid sử dụng nhận diện khuôn mặt vào mục đích giám sát trong vòng năm năm[32].

Tóm lại, cách tiếp cận hợp lý đối với công nghệ có bản chất vừa lợi vừa hại này là thúc đẩy các cuộc tranh luận dân chủ để xác định xem ứng dụng nào phù hợp, phản đối những cách sử dụng không đúng đắn, và xây dựng các cơ chế bảo vệ cần thiết để ngăn chặn việc lạm dụng – dù từ chính phủ hay các tổ chức tư nhân.

Một chuỗi sự kiện kỳ lạ đã dẫn đến quyển sách này

Cuối năm 2019, một cựu nghiên cứu viên tại một công ty AI bất ngờ liên hệ với Arvind. Công ty này đang kiếm lời trong lĩnh vực tự động hóa tuyển dụng – một ngành kinh doanh béo bở nhưng cũng đầy những trò bịp, như chúng tôi đã đề cập ở trên. Nhà nghiên cứu này tiết lộ rằng dù nội bộ công ty biết rõ công cụ tuyển dụng của họ không hiệu quả như quảng cáo, ban lãnh đạo vẫn cố tình cản trở mọi nỗ lực kiểm định độ chính xác của hệ thống.

Trùng hợp thay, ngay sau cuộc gặp đó, Arvind được mời diễn thuyết tại MIT. Anh quyết định nói về những trò bịp trong AI, đặc biệt là sự thiếu minh bạch trong việc tự động hóa tuyển dụng, và nhận được phản hồi tích cực. Arvind quyết định chia sẻ bản trình chiếu lên mạng, ngỡ rằng chỉ một số học giả và nhà hoạt động quan tâm đến chủ đề này, nhưng thật bất ngờ, bài thuyết trình nhanh chóng lan truyền mạnh mẽ với hàng chục ngàn lượt tải xuống và hai triệu lượt xem trên Twitter.

Khi đã hết cảm giác sốc, Arvind nhận ra lý do khiến chủ đề này gây được tiếng vang: rất nhiều người vốn đã nghi ngờ rằng công nghệ AI xung quanh họ có gì đó không ổn, nhưng họ thiếu từ vựng và thẩm quyền để đặt ra nghi vấn[33]. Suy cho cùng, những công nghệ đó được quảng bá bởi các thiên tài tự xưng và các tập đoàn ngàn tỷ đô-la mà. Nhưng khi một giáo sư khoa học máy tính lên tiếng vạch trần những điều vô lý này, sự hoài nghi của công chúng bỗng trở nên chính đáng và được lan tỏa rộng rãi hơn.

Chỉ trong vòng hai ngày, hộp thư của Arvind đã nhận được khoảng bốn mươi đến năm mươi lời mời chuyển bài thuyết trình đó thành một bài báo, hay thậm chí một quyển sách. Nhưng anh cảm thấy mình chưa hiểu đủ sâu về chủ đề này để viết thành sách. Anh không muốn viết nếu không có điều gì đáng để nói, và càng không muốn tận dụng sự nổi tiếng tạm thời để kiếm lợi.

Ở môi trường đại học, có hai cách để hiểu sâu một vấn đề: đăng ký học về nó, hoặc tốt nhất là giảng dạy về nó. Vì vậy, Arvind đã hợp tác với giáo sư xã hội học Matthew Salganik của Princeton để mở một khóa học có tên là Những giới hạn của dự đoán (Limits to Prediction). Matt đã công bố nhiều nghiên cứu nền tảng chỉ ra lý do tại sao lại khó có thể dự đoán tương lai bằng AI. Hai nghiên cứu quan trọng của ông sẽ được đề cập ở Chương 3. Trong khóa học, Arvind và Matt khuyến khích sinh viên tiến hành nghiên cứu thực tế. Một trong số những sinh viên đó là Sayash.

Sayash trước đây từng làm việc cho Facebook và vừa nhập học Princeton. Anh quyết định rời Facebook để theo đuổi bằng tiến sĩ, với mong muốn nghiên cứu công nghệ vì lợi ích cộng đồng thay vì phục vụ cho doanh nghiệp. Anh được nhận vào một số chương trình tiến sĩ khoa học máy tính, và như thường lệ, các nghiên cứu sinh sẽ đến thăm khoa, gặp gỡ các giáo sư tiềm năng và đặt câu hỏi để tìm người cố vấn phù hợp.

Thông thường, các nghiên cứu sinh thường được khuyến khích hỏi những câu như "Thầy/cô có phong cách hướng dẫn thế nào?", "Học viên của thầy/cô có thường tạm nghỉ không?",

hay "Thầy/cô nghĩ gì về việc cân bằng giữa công việc với cuộc sống?". Những câu hỏi này hữu ích, nhưng chỉ nói lên cách một giáo sư làm việc. Còn để hiểu giá trị và tư duy của họ, cần hỏi điều khác.

Sayash chọn một câu hỏi không ai lường trước: "Thầy/cô sẽ làm gì nếu bị một công ty công nghệ kiện?". Câu hỏi này vừa đủ bất ngờ để không ai có câu trả lời sẵn, nhưng cũng không hề phi thực tế. Câu trả lời có thể cho biết quan điểm của giáo sư đó về các tập đoàn công nghệ lớn, cách họ nhìn nhận tác động từ nghiên cứu của mình và cách họ phản ứng khi gặp khủng hoảng. Khi Arvind đáp "Tôi sẽ rất vui nếu bị một công ty đe dọa kiện vì nghiên cứu của mình. Điều đó chứng tỏ công việc của tôi đang tạo ra tác động thực sự", Sayash biết anh đã tìm được chương trình phù hợp.

Trong khóa học về những giới hạn của dự đoán, sinh viên rất hứng thú với AI dự đoán – tức mọi nỗ lực sử dụng dữ liệu để dự đoán tương lai, đặc biệt là trong bối cảnh xã hội, từ các nền văn minh đến mạng xã hội. Một số câu hỏi họ tìm hiểu là: chúng ta có dự đoán được sự kiện địa chính trị nào sẽ xảy ra không, chẳng hạn kết quả bầu cử, suy thoái kinh tế, phong trào xã hội? Có dự đoán được video nào sẽ lên xu hướng không?

Những gì họ tìm thấy là một "nghĩa trang" của những tham vọng dự đoán tương lai. Một nhóm rào cản cơ bản dường như cứ lặp lại hết lần này đến lần khác, nhưng vì giới nghiên cứu trong các lĩnh vực khác nhau hiếm khi trao đổi với nhau, nên nhiều ngành đã độc lập phát hiện ra những giới hạn giống nhau. Chúng ta hoảng sợ trước sự trái ngược giữa một bên là

bằng chứng rõ ràng và một bên là nhận thức rộng rãi rằng học máy là một công cụ tốt để dự đoán tương lai.

Một trong những nghiên cứu điển hình đáng chú ý mà khóa học đề cập đến là Google Flu Trends. Dự án này của Google ra mắt vào năm 2008, dự đoán dịch cúm bằng cách phân tích các truy vấn tìm kiếm mà hàng triệu người dùng của họ thực hiện mỗi ngày. Thông tin liên quan đến triệu chứng cúm được tìm kiếm nhiều hơn có thể là dấu hiệu cho thấy dịch cúm sắp bùng phát. Google quảng bá mạnh mẽ dự án này như một ví dụ điển hình của việc sử dụng AI và dữ liệu lớn để phục vụ xã hội. Nhưng chỉ vài năm sau, độ chính xác của dự đoán sụt giảm nghiêm trọng với lý do chính: không thể phân biệt giữa người tìm kiếm vì bị truyền thông làm cho hoảng sợ với người thực sự mắc bệnh. Ngoài ra, những thay đổi trong thuật toán của chính Google cũng vô tình làm thay đổi mô hình tìm kiếm của người dùng, khiến AI dự đoán chệch hướng. Cuối cùng, Google Flu Trends trở thành một câu chuyện cảnh báo[34]. Bài học ở đây là ngay cả khi tưởng chừng có thể dự đoán được, chỉ một chi tiết sai trong cách tiếp cận cũng có thể khiến toàn bộ kết quả sai lệch.

Sayash nhận thấy khóa học đã xác nhận lại những trải nghiệm trước đây của anh tại Facebook, nơi anh chứng kiến việc xây dựng AI dễ mắc lỗi ra sao và việc lạc quan quá mức về hiệu quả của nó nguy hiểm thế nào. Lỗi có thể phát sinh do nhiều nguyên nhân tinh vi và thường không bị bắt lại trong quá trình kiểm thử, mà chỉ lộ ra khi AI được đưa vào sử dụng thực tế[35]. Sayash quyết định chọn "giới hạn của AI" làm đề tài nghiên cứu.

Sau bốn năm đào sâu – cả độc lập và hợp tác – giờ đây chúng tôi đã sẵn sàng chia sẻ những điều mình học được. Nhưng quyển sách này không dừng lại ở việc chia sẻ kiến thức đơn thuần. AI đang ngày càng được sử dụng để đưa ra các quyết định quan trọng liên quan đến cuộc sống của mỗi người. Nếu nó mắc lỗi, hậu quả có thể là hủy hoại sự nghiệp hoặc cuộc sống của con người. Tất nhiên, không phải tất cả AI đều là "trò bịp", còn lâu mới vậy. Nhưng cũng vì thế, việc phân biệt giữa tiến bộ thực sự và quảng cáo thổi phồng là điều tối quan trọng. Có lẽ quyển sách này sẽ giúp bạn làm điều đó.

Vòng xoáy thổi phồng về AI

Kể từ khi bắt đầu làm việc cùng nhau, chúng tôi đã dần hiểu rõ hơn vì sao lại có quá nhiều thông tin sai lệch, hiểu nhầm và lầm tưởng xoay quanh AI. Nói ngắn gọn, vấn đề này tồn tại dai dẳng là vì cả giới nghiên cứu, các công ty và truyền thông đều góp phần vào việc duy trì nó.

Hãy bắt đầu với một ví dụ từ giới học thuật. Một bài báo khoa học năm 2023 tuyên bố rằng học máy có thể dự đoán chính xác đến 97% bài hát nào sẽ trở thành hit (bài hát nổi tiếng)[36]. Với ngành công nghiệp âm nhạc luôn khát khao tìm ra bản hit tiếp theo, phát hiện này chắc hẳn nghe rất êm tai. Nhiều hãng truyền thông, bao gồm cả *Scientific American* và Axios, đã nhanh chóng đưa tin rằng "độ chính xác đến mức đáng sợ" này sẽ cách mạng hóa ngành âm nhạc[37,38].

Tiếc là sự thật không được như thế. Phương pháp được trình bày trong bài nghiên cứu này để lộ một trong những sai lầm cơ bản nhất của học máy: rò rỉ dữ liệu. Tức là mô hình đánh giá trên một tập dữ liệu trùng hoặc gần giống với tập đã dùng để huấn luyện, khiến độ chính xác bị phóng đại. Nói thẳng ra, điều này giống như bạn cho học sinh học thuộc câu hỏi rồi dùng chính bộ đề đó để thi, hoặc tệ hơn, đưa trước luôn đáp án. Khi chúng tôi sửa lỗi này và cho chạy lại phân tích, mô hình không dự đoán tốt hơn so với việc đoán ngẫu nhiên.

Đây không phải trường hợp cá biệt. Sai sót cơ bản trong các nghiên cứu học máy xảy ra phổ biến đến mức đáng ngại, nhất là khi các nhà nghiên cứu không chuyên về khoa học máy tính sử dụng AI như một công cụ sẵn có. Ví dụ, giới y học dùng AI để dự đoán bệnh, giới khoa học xã hội dùng để dự đoán cuộc đời con người, giới khoa học chính trị dùng để dự đoán nội chiến.

Các đánh giá có hệ thống về những nghiên cứu đã công bố trong nhiều lĩnh vực chỉ ra rằng *phần lớn* nghiên cứu dựa trên học máy hóa ra đều chứa sai sót khi được kiểm tra lại[39]. Nguyên nhân không phải lúc nào cũng do cố ý, mà đơn giản vì học máy là một lĩnh vực vốn ẩn chứa rủi ro, rất dễ khiến người làm nghiên cứu tự đánh lừa mình. Nhiều nhóm nghiên cứu trong cả chục lĩnh vực khác nhau đã thu thập bằng chứng về những sai lệch phổ biến trong lĩnh vực của họ, nhưng lại không nhận ra họ đang cùng góp phần vào một cuộc khủng hoảng sâu rộng về độ tin cậy của học máy.

Một nghịch lý đáng chú ý: chủ đề càng được chú ý nhiều, chất lượng nghiên cứu lại càng kém. Lúc trước, có hàng ngàn

nghiên cứu tuyên bố có thể phát hiện COVID-19 từ ảnh chụp X-quang ngực và các dữ liệu hình ảnh khác. Một đánh giá có hệ thống đã xem xét hơn 400 bài báo và kết luận rằng *tất cả các nghiên cứu đó* đều không đủ giá trị lâm sàng vì mắc sai sót nghiêm trọng về phương pháp[40]. Trong nhiều trường hợp, toàn bộ ảnh chụp bệnh nhân nhiễm COVID-19 là người lớn, còn ảnh người không nhiễm lại là trẻ em. Hệ quả là AI chỉ học được cách phân biệt trẻ em với người lớn, nhưng các nhà nghiên cứu tưởng rằng mình đã tạo ra được công cụ chẩn đoán COVID-19.

Chúng tôi cũng đã phát hiện ra nhiều sai sót tương tự trong nhiều nghiên cứu khác, nhất là trong lĩnh vực dự đoán nội chiến – nói ngắn gọn là không hiệu quả. Khi chúng tôi cố gắng công bố một bài báo chỉ ra lỗi hệ thống trong lĩnh vực này, không một tạp chí nào quan tâm. Chỉnh sửa những sai lầm trong lịch sử khoa học là một việc vô cùng gian nan. Cuối cùng, chúng tôi cũng đăng được bài viết, nhưng phải chỉnh sửa theo hướng "thân thiện" hơn, biến nó thành một cẩm nang hướng dẫn để các nhà nghiên cứu tương lai không phạm phải lỗi cũ.

Thời điểm này, nếu phát hiện một nghiên cứu học máy có vấn đề, chúng tôi không còn cố gắng sửa nữa. Hệ thống hiện tại không hỗ trợ điều đó. Thật trớ trêu, trong nhiều lĩnh vực, những nghiên cứu không thể tái lập còn được trích dẫn nhiều hơn cả những nghiên cứu đã tái lập thành công[41]. Người ta vẫn nói rằng khoa học có khả năng "tự điều chỉnh", nhưng trải nghiệm của chúng tôi cho thấy điều ngược lại.

Cần phải nói rõ rằng những tuyên bố sai lệch trong nghiên cứu học máy không nhất thiết sẽ dẫn đến sản phẩm

AI thương mại kém chất lượng. Ví dụ, nếu một nhà sản xuất âm nhạc dùng mô hình dự đoán các bản hit sai, họ sẽ nhanh chóng nhận ra rằng nó không hiệu quả. (Các sản phẩm AI thương mại kém chất lượng thường do các công ty cố tình bán công nghệ AI không hiệu quả, chứ không phải họ tự đánh lừa mình.) Tuy nhiên, biển thông tin sai lệch trong khoa học làm nhận thức của công chúng về AI bị méo mó, đặc biệt khi truyền thông có xu hướng thổi phồng mọi tuyên bố về "đột phá công nghệ".

Dẫu vậy, vẫn có những tia hy vọng. Mùa hè năm 2022, chúng tôi đã tổ chức một hội thảo trực tuyến kéo dài một ngày để thảo luận về thực trạng rất nhiều nghiên cứu dựa trên học máy bị mắc lỗi, và bất ngờ thay, hàng trăm nhà khoa học đã tham gia. Từ sự kiện đó, chúng tôi thành lập một nhóm khoảng hai mươi nhà nghiên cứu đến từ nhiều lĩnh vực, cùng xây dựng một danh sách kiểm tra đơn giản giúp cải thiện cách dẫn chứng dữ liệu sử dụng học máy. Mục tiêu là giảm thiểu sai sót và giúp phát hiện lỗi dễ hơn[42]. Tuy nhiên, đây mới chỉ là bước đầu, chưa rõ liệu nỗ lực này có được phổ biến rộng rãi không. Dẫu sao thì những thay đổi trong thực hành khoa học cũng luôn diễn ra rất chậm, và mọi chuyện có thể còn tiếp tục tệ hơn rồi tình hình mới được cải thiện.

Còn ở phía doanh nghiệp, nếu các nghiên cứu bị thổi phồng gây hiểu nhầm cho công chúng, thì sản phẩm bị thổi phồng có thể gây hại trực tiếp. Chúng tôi đã hợp tác với các đồng nghiệp Angelina Wang và Solon Barocas để tìm hiểu việc sử dụng AI dự đoán trong lĩnh vực công nghiệp và hoạt động của chính quyền[43]. Nhóm chúng tôi khảo sát hơn năm mươi

ứng dụng trong các ngành tư pháp, y tế, phúc lợi, tài chính, giáo dục, quản trị nhân sự và tiếp thị. Phần lớn mới được triển khai gần đây. Trong thập niên 2010, AI dự đoán đã len lỏi vào mọi mặt đời sống – âm thầm đánh giá và định đoạt cơ hội của con người dựa trên những dữ liệu được thu thập lén lút về chúng ta.

Chúng tôi nhận thấy nhà cung cấp các công cụ AI này rất tích cực tiếp thị sản phẩm, nhưng lại hiếm khi minh bạch về độ hiệu quả của sản phẩm, hay thậm chí cho biết chúng có chút hiệu quả nào hay không. Đáng chú ý là, theo chúng tôi biết, chưa có công ty tự động hóa tuyển dụng nào từng công bố một nghiên cứu được thẩm định độc lập để kiểm chứng độ chính xác của AI dự đoán, cũng chưa có ai cho phép nhà nghiên cứu bên ngoài đánh giá AI dự đoán của họ. Hai công ty hàng đầu đã cố gắng thể hiện tính minh bạch bằng cách thực hiện kiểm toán bên ngoài: Pymetrics ký hợp đồng với một nhóm nghiên cứu hàng đầu từ Đại học Northeastern, còn HireVue mời một đơn vị kiểm toán độc lập có uy tín. Nhưng trong cả hai trường hợp này, các nhà nghiên cứu chỉ được phép phân tích yếu tố định kiến giới tính hay chủng tộc, chứ không thể kiểm tra xem nó có hiệu quả hay không. Các công ty đã khéo léo lợi dụng mối quan tâm về phân biệt đối xử để đánh lạc hướng vấn đề thực sự. Nếu sản phẩm của bạn thực chất chỉ là một bộ tạo số ngẫu nhiên phức tạp, hoạt động kém hiệu quả như nhau với tất cả mọi người, thì quá dễ để làm cho nó không định kiến!

Tuy nhiên, các cơ quan quản lý bắt đầu tỉnh táo hơn. Năm 2023, Ủy ban Thương mại Liên bang Hoa Kỳ (U.S. Federal Trade

Commission - FTC) đã cảnh báo các công ty rằng: "Chúng ta vẫn chưa sống trong thế giới khoa học viễn tưởng, nơi máy móc nhìn chung có thể đưa ra những dự đoán chính xác về hành vi con người. Những tuyên bố về hiệu suất của bạn sẽ bị coi là lừa dối nếu thiếu bằng chứng khoa học, hay chỉ áp dụng cho một số nhóm người dùng hoặc trong những hoàn cảnh nhất định"[44]. Từ khóa ở đây là "lừa dối", FTC được Quốc hội giao cho quyền giải quyết các hành vi lừa dối của doanh nghiệp. Chúng tôi hy vọng các công ty sẽ chú ý đến cảnh báo này.

Còn truyền thông thì sao? Nếu các nhà nghiên cứu và công ty khơi lên một đốm lửa thổi phồng, thì truyền thông lại thổi bùng nó thành ngọn lửa. Ngày nào cũng có bài viết dồn dập ca ngợi những đột phá của AI. Nhiều bài trong số đó chỉ là thông cáo báo chí được biên tập lại.

Tất nhiên, điều này không có gì lạ, bởi các tòa soạn phụ thuộc vào lượt nhấp chuột và họ cũng cần tiền. Nhưng ngoài yếu tố kinh tế, báo chí công nghệ còn gặp những vấn đề mang tính cấu trúc. Nhiều phóng viên AI thực hành cái gọi là "báo chí tiếp cận" – nghĩa là duy trì mối quan hệ tốt với các công ty công nghệ để được quyền phỏng vấn hoặc có thông tin sớm. Vì vậy họ không được đặt quá nhiều câu hỏi hoài nghi.

Những tuyên bố về "AI có tri giác" đặc biệt hấp dẫn với truyền thông. Khi một kỹ sư của Google tuyên bố vào tháng Sáu năm 2022 rằng chatbot nội bộ của công ty đã trở nên có tri giác (thậm chí còn phải chịu "định kiến"), hầu như mọi tờ báo đều giựt tít về việc đó[45]. Điều tương tự cũng xảy ra với chatbot của Bing vào đầu năm 2023. Mặc dù hầu hết giới nghiên cứu AI

cho rằng những tuyên bố đó không có chút cơ sở khoa học nào, chúng vẫn được lan truyền rộng rãi.

Vẫn có những nhà báo chuyên viết về AI làm việc cẩn trọng và rất giỏi tìm hiểu. Nhưng họ chỉ là số ít, và còn phải vật lộn không ngừng giữa biển thông tin sai lệch. Chúng tôi từng có dịp trao đổi về vấn đề thổi phồng AI với một số người trong số đó, đồng thời cũng từng phát biểu tại vài hội nghị báo chí, và chúng tôi biết nhiều sáng kiến cải thiện chất lượng báo chí công nghệ đang thành hình.

Ví dụ, Trung tâm Pulitzer tài trợ cho một mạng lưới nhà báo chuyên thực hiện "các phóng sự chuyên sâu về trách nhiệm giải trình khi dùng AI, xem xét cách chính phủ và doanh nghiệp sử dụng công nghệ dự đoán và giám sát để đưa ra quyết định trong các lĩnh vực trị an, y tế, phúc lợi xã hội, hệ thống tư pháp hình sự, tuyển dụng,..."[46]. Một trong những bài điều tra nổi bật từ chương trình này là của Ari Sen và Derêka K. Bennett, đăng trên tờ *Dallas Morning News*, phơi bày cách các trường học sử dụng phần mềm Social Sentinel – một sản phẩm AI được quảng cáo là công cụ giúp các trường trên khắp nước Mỹ quét bài đăng trên mạng của học sinh nhằm phát hiện các rủi ro về an ninh – nhưng thực tế lại được dùng để giám sát các cuộc biểu tình của học sinh[47].

Tuy nhiên, mỗi năm Trung tâm Pulitzer chỉ tài trợ cho mười nhà báo. Về lâu dài, việc báo chí có thể trở thành rào cản chống lại quyền lực của Big Tech (tập đoàn công nghệ lớn) hay không còn phụ thuộc vào việc các mô hình tài trợ không dựa vào lượt xem như thế này có được nhân rộng hay không.

Cuối cùng, các chuyên gia AI có trách nhiệm lên tiếng chống lại sự thổi phồng phi lý, dù khởi nguồn từ các nhà nghiên cứu, doanh nghiệp hay truyền thông. Chúng tôi đang cố gắng làm phần việc của mình. Trong bản tin trên trang AISnakeOil.com, chúng tôi sẽ bình luận về các xu hướng AI mới và giúp độc giả phân biệt đâu là thực chất, đâu chỉ là bong bóng xà phòng[48].

Trò bịp AI (AI snake oil) là gì?

Vào cuối thế kỷ 19 - đầu thế kỷ 20, các nhà buôn dầu rắn (snake oil) đi khắp nước Mỹ, chào mời các loại "thần dược" chữa bách bệnh với những lời khẳng định giả dối. Những tay buôn này đã lợi dụng niềm tin thiếu cơ sở khoa học rằng dầu chiết xuất từ rắn mang lại nhiều lợi ích cho sức khỏe và việc dân chúng thiếu khả năng phân biệt giữa phương pháp điều trị hiệu quả với sản phẩm vô dụng.

Thậm chí, phần lớn các loại dầu rắn không hề chứa một chút dầu rắn nào. Có thể nói, dầu rắn là một trò bịp bợm, lừa đảo. Một số loại có thể vô dụng nhưng vô hại, trong khi nhiều loại khác lại gây tổn hại nghiêm trọng đến sức khỏe, thậm chí tính mạng. Trước khi Cục Quản lý Thực phẩm và Dược phẩm Hoa Kỳ (Food and Drug Administration - FDA) ra đời vào năm 1906, gần như không có cách nào để buộc những người bán dầu rắn chịu trách nhiệm cho những hứa hẹn của họ về thành phần, độ an toàn hay hiệu quả của sản phẩm.

"Dầu rắn AI" là những hệ thống AI bịp bợm, không (và không thể) hoạt động đúng như đã quảng cáo, chẳng hạn

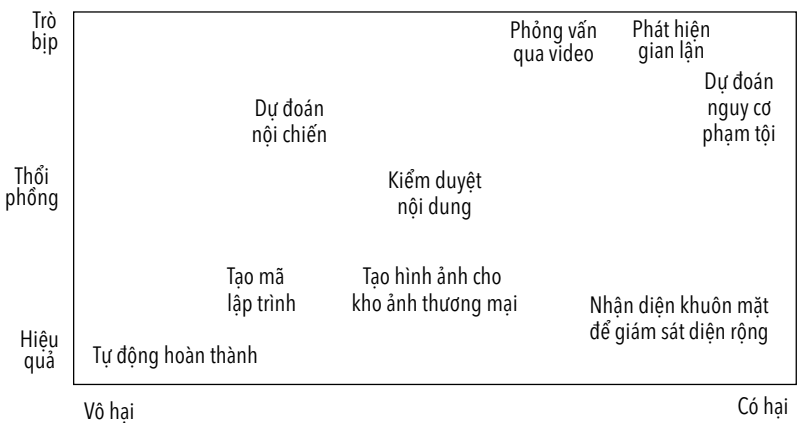
như phần mềm phân tích video trong tuyển dụng – một trong những động lực ban đầu thúc đẩy chúng tôi bắt tay vào nghiên cứu để viết nên quyển sách này. Mục tiêu của quyển sách là xác định rõ đâu là AI bịp bợm và đâu là những hệ thống có thể hoạt động hiệu quả nếu được triển khai đúng cách. Có những trường hợp AI bịp bợm rất rõ ràng, nhưng trong nhiều trường hợp khác, ranh giới lại khá mơ hồ. Có những công cụ hiệu quả ở một chừng mực nào đó, nhưng lại bị các công ty thổi phồng quá mức, khiến người dùng lầm tưởng rằng chúng có thể thay thế con người, thay vì chỉ là một cách bổ sung cho năng lực của con người.

Một điều quan trọng không kém: ngay cả khi hoạt động hiệu quả, AI vẫn có thể gây hại, như chúng ta đã thấy trong trường hợp công nghệ nhận diện khuôn mặt bị lạm dụng để giám sát hàng loạt. Để hiểu được bản chất và biết cách khắc phục những tác hại này, ta cần xác định được vấn đề nảy sinh là do AI không có tác dụng, do nó bị thổi phồng quá mức, hay thực ra là vì nó đang hoạt động đúng như ý định ban đầu. Hai yếu tố chính cần xem xét là tác hại và độ trung thực của AI như hai trục ở hình 1.1. Trong quyển sách này, chúng tôi sẽ phân tích tất cả các vùng của biểu đồ, trừ góc dưới bên trái (khi AI vừa hiệu quả vừa vô hại).

Với bức tranh tổng quan đã nêu, phần còn lại của quyển sách sẽ đi theo lộ trình như sau:

Chương 2 tập trung vào các hệ thống ra quyết định tự động, một lĩnh vực mà AI – đặc biệt là AI dự đoán – ngày càng được ứng dụng rộng rãi. Chúng ta sẽ xem xét các hệ thống được

thiết kế để dự đoán ai có thể phạm tội, ai sẽ bỏ học,… và phân tích nhiều trường hợp thất bại gây ra hậu quả nghiêm trọng. Trong quá trình nghiên cứu, chúng tôi đã xác định được những nguyên nhân lặp đi lặp lại dẫn đến các thất bại này, những nguyên nhân bắt nguồn từ logic dự đoán khi áp dụng vào các hệ thống có ảnh hưởng sâu rộng đến con người. Kết thúc chương, chúng tôi đặt ra câu hỏi: liệu có thể thiết kế lại quy trình ra quyết định mà không cần đến AI dự đoán? Và nếu có, chúng ta cần thay đổi những gì về tổ chức và văn hóa để chấp nhận tính bất định cố hữu của các quyết định quan trọng?



Hình 1.1. Tổng quan về trò bịp AI, sự thổi phồng và các tác hại, minh họa qua một số ứng dụng tiêu biểu.

Chương 3 đưa chúng ta lùi lại một bước để hiểu tại sao việc dự đoán tương lai lại khó khăn đến vậy. Thách thức này rốt cuộc không chỉ sinh ra từ giới hạn của công nghệ, mà còn bắt nguồn từ bản chất của các quá trình xã hội. Dự đoán

hành vi con người vốn dĩ là một việc khó khăn. Chúng ta sẽ xem xét nhiều nỗ lực dự đoán tương lai, từ nguy cơ phạm tội đến triển vọng nghề nghiệp, khả năng hoàn trả khoản vay, cho đến thành công trong việc xuất bản sách. Một số ví dụ rút ra từ nghiên cứu khoa học, một số khác là sản phẩm thương mại hiếm hoi được đánh giá nghiêm túc. Ngoài ra, chúng ta cũng sẽ phân tích các nhiệm vụ dễ lượng hóa hơn như dự đoán bài đăng nào trên mạng xã hội có thể lan truyền mạnh, hay diễn biến của đại dịch ở quy mô toàn cầu. Những trường hợp này cho thấy một điều rõ ràng: những giới hạn cố hữu của AI dự đoán sẽ không dễ dàng biến mất trong tương lai gần.

Chương 4 chuyển trọng tâm sang AI tạo sinh. Không khó để chỉ ra điểm giới hạn của AI dự đoán, nhưng với AI tạo sinh, mọi việc lại phức tạp hơn nhiều. Công nghệ này có những khả năng đáng kinh ngạc, nhưng lại thất bại ở nhiều tác vụ đơn giản mà ngay cả một đứa trẻ cũng làm được[49]. Vì vậy, để hiểu được giới hạn của AI tạo sinh, cũng như tiềm năng phát triển của nó, điều quan trọng là phải hiểu cách thức vận hành bên trong. Chúng tôi sẽ giải thích cơ chế hoạt động của AI tạo sinh, đồng thời phân tích những rủi ro nó gây ra. Một số rủi ro bắt nguồn từ việc công cụ hoạt động sai lệch, như phần mềm phát hiện gian lận học thuật tạo ra cáo buộc sai. Một số khác lại xuất hiện khi công cụ hoạt động quá tốt, chẳng hạn khi công ty AI sử dụng ảnh của các nhiếp ảnh gia mà không trả phí, gây mất việc hàng loạt trong ngành nhiếp ảnh. Chúng tôi cũng sẽ nêu ra những rủi ro âm thầm hơn, như lỗi trong mã do AI tạo ra có thể tạo ra lỗ hổng bảo mật. Dù quyển sách này không đi sâu

vào những ứng dụng tích cực, nhưng không có nghĩa là chúng không tồn tại hay không quan trọng.

Chương 5 bàn về những rủi ro với sự tồn vong, mối quan tâm ngày càng lớn trong dư luận. Một số người lo sợ rằng AI có thể trở nên quá thông minh và vượt khỏi tầm kiểm soát của con người. Tuy nhiên, chúng tôi cho rằng những lo ngại này thường dựa trên một cái nhìn nhị phân về AI – tức là cho rằng AI sẽ đạt tới một ngưỡng tự chủ nhất định rồi đột ngột "thức giấc". Nhưng lịch sử phát triển công nghệ AI lại cho thấy một xu hướng tăng trưởng tuyến tính: từng bước, chậm nhưng chắc. Chúng ta có nhiều cơ sở để tin rằng AI sẽ tiếp tục phát triển theo xu hướng này, có thể dựa vào lịch sử để định hướng thay vì suy đoán mơ hồ. Nhiều lo ngại về việc AI vượt tầm kiểm soát thực chất dựa trên hàng loạt giả định thiếu cơ sở. Dù vẫn cần thận trọng với các rủi ro từ AI siêu việt, chúng ta đã có công cụ để ứng phó một cách bình tĩnh và hợp lý.

Chương 6 đưa chúng ta đến lĩnh vực mạng xã hội, nơi các thuật toán AI đóng vai trò trung tâm trong việc lựa chọn nội dung mà người dùng nhìn thấy. AI còn được dùng để kiểm duyệt nội dung – xác định đâu là bài viết cần xóa vì vi phạm chính sách. Chúng tôi đặt ra câu hỏi cốt lõi: liệu AI có thể xác định và loại bỏ nội dung độc hại mà vẫn bảo vệ tự do ngôn luận như các công ty công nghệ hứa hẹn? Ngay cả khi lỗi báo cáo sai được khắc phục, vấn đề lớn hơn vẫn tồn tại: ai là người định nghĩa "độc hại" và "chính đáng"? Trong thực tế, các nền tảng công nghệ có khả năng điều chỉnh phát ngôn mà hầu như không phải chịu trách nhiệm giải trình. Chúng ta thiếu một

quy trình dân chủ để quyết định các quy tắc quản lý lời nói trên mạng và để tìm ra sự cân bằng giữa các giá trị như tự do ngôn luận và an toàn. Những thất bại trong kiểm duyệt, nếu có, là hệ quả tất yếu của thiếu sót đó, chứ không hẳn do AI yếu kém.

Chương 7 lý giải vì sao những huyền thoại về AI lại phổ biến. Các công ty không chỉ thổi phồng công nghệ của mình, mà còn dùng quyền lực và tiền bạc để làm suy yếu vai trò giám sát của giới học thuật và báo chí. Trớ trêu thay, nhiều nhà nghiên cứu cũng góp phần lan truyền sự phóng đại đó, thay vì phản biện nó. Trong nhiều lĩnh vực, họ đi đến đồng thuận sai lệch rằng AI rất chính xác, dựa trên các nghiên cứu đầy lỗi và không thể tái lập. Một ví dụ là các mô hình dự đoán nội chiến – chủ đề sẽ được phân tích trong chương này. Dù không phải lúc nào cũng tạo ra các sản phẩm AI sai hỏng, những nghiên cứu như vậy vẫn gây lãng phí, và tệ hơn nữa là làm lệch lạc nhận thức của công chúng. Bên cạnh đó, chúng tôi sẽ phân tích vai trò của truyền thông trong việc vô tình hoặc cố ý góp phần thổi bùng ngọn lửa phóng đại AI, đồng thời hướng dẫn cho bạn cách đọc tin tức tỉnh táo hơn.

Cuối cùng, **Chương 8**, chúng tôi đề xuất ba hướng thay đổi lớn. Thứ nhất, cần đặt ra quy tắc nền tảng cho các công ty trong việc phát triển và tiếp thị sản phẩm AI. Lưu ý, dù hoạt động quản lý là cần thiết, chúng tôi cho rằng không nên siết chặt quá mức. Thứ hai, cần xem xét cách AI được tích hợp vào xã hội: AI sẽ góp mặt ra sao trong giáo dục, đời sống trẻ em nói chung, hay môi trường làm việc? Những vấn đề này không chỉ phụ thuộc vào khả năng của công nghệ, mà còn là kết quả của

các quyết định mang tính xã hội và chính trị. Thứ ba, thay vì chỉ tập trung kiểm soát nguồn cung "AI bịp bợm", chúng ta cần nhìn vào nhu cầu đang thúc đẩy nó. Nhiều tổ chức như trường học tìm đến các công cụ AI không phải vì chúng hiệu quả, mà vì họ đang lúng túng trước một vấn đề lớn và cần giải pháp tức thời. Ví dụ, khi học sinh bắt đầu dùng AI để làm bài tập, nhiều giáo viên vốn đã quá tải phản ứng bằng cách sử dụng phần mềm phát hiện gian lận. Nhưng những phần mềm này thường không chính xác, dẫn đến các cáo buộc sai và gây hậu quả nghiêm trọng cho học sinh.

Chúng ta không thể giải quyết những vấn đề này chỉ bằng cách cải tiến công nghệ. Thậm chí, "AI bịp bợm" đôi khi lại hữu ích vì nó phơi bày những vấn đề sâu xa. Trên phạm vi rộng hơn, chúng tôi sẽ chỉ ra rằng thực chất những lo ngại về AI, đặc biệt là trong thị trường lao động, phản ánh những vấn đề của chủ nghĩa tư bản. Điều quan trọng là chúng ta cần khẩn trương tìm ra cách củng cố các mạng lưới an sinh hiện có và phát triển thêm các cơ chế bảo vệ mới để có thể thích nghi tốt hơn với những thay đổi nhanh chóng do tiến bộ công nghệ mang lại và tận dụng triệt để lợi ích của nó.

AI dự đoán, AI tạo sinh và AI kiểm duyệt nội dung là ba dạng AI chính mà chúng tôi sẽ phân tích. Danh sách này chưa toàn diện. Như đã đề cập ở trên, còn nhiều ứng dụng AI khác hoạt động khá hiệu quả nhưng thường ít được chú ý vì mức độ ảnh hưởng xã hội của chúng tương đối nhỏ, chẳng hạn như tính năng tự động điền hay kiểm tra chính tả. Ngược lại, những công nghệ như robot hay xe tự lái đáng được quan tâm và

thảo luận, song chúng tôi không đi sâu trong khuôn khổ quyển sách này, phần vì tác động xã hội của chúng hiện vẫn chưa phổ biến ở quy mô lớn.

Dù vậy, các nguyên tắc và cách tiếp cận được trình bày ở đây có thể giúp bạn tự đánh giá các ứng dụng khác: đâu là công nghệ có cơ sở vững chắc, và đâu có thể chỉ là một "trò bịp AI".

Quyển sách này dành cho ai?

Chúng tôi hy vọng quyển sách này sẽ hữu ích cho ba nhóm độc giả chính.

Trước hết, có thể bạn đơn giản chỉ muốn hiểu rõ hơn chuyện gì đang xảy ra. Bạn đã thấy những tiêu đề như "AI có thể dự đoán động đất" hay "AI vượt qua kỳ thi tuyển luật sư" và tự hỏi: Điều đó có thật không, và nếu có thì nó diễn ra thế nào? Những ngành nghề nào sẽ còn tồn tại trong hai mươi năm tới? Cuộc sống của con cái chúng ta sẽ ra sao? Để trả lời, chúng tôi không đi theo hướng triết lý về con người trong thời đại AI, mà sẽ mang đến một cái nhìn thiết thực, cốt lõi, giúp bạn hiểu được công nghệ này thực sự đang hoạt động như thế nào phía sau những lời quảng bá.

Tiếp theo, có thể bạn đang cần quyết định sẽ sử dụng AI như thế nào cho công việc. Chúng tôi hy vọng quyển sách sẽ cung cấp thông tin hữu ích để bạn đánh giá đâu là những loại AI có thể hoạt động hiệu quả, đâu là những công cụ bị thổi phồng quá mức, và thách thức nào cần lưu ý. Thay vì đưa ra các

tuyên bố bao quát về những gì là "dễ" hay "khó" đối với AI như các nhà khoa học máy tính vẫn làm, chúng tôi sẽ phân tích cụ thể từng loại.

Cuối cùng, nếu bạn quan tâm đến AI vì mong muốn hành động để bảo vệ lợi ích cộng đồng, quyển sách này cũng dành cho bạn. Trong lĩnh vực AI dự đoán, các nhà hoạt động đã thành công trong việc đặt ra giới hạn đối với những ứng dụng gây hại. Nhưng ở lĩnh vực AI tạo sinh, ranh giới vẫn đang hình thành. Công nghệ này không chỉ gây biến đổi nền kinh tế, làm mất công việc, mà còn dựa vào lao động vô hình của hàng triệu người để phát triển – từ những công việc lặp đi lặp lại, được trả lương thấp, cho đến việc thu thập dữ liệu huấn luyện mà không ghi nhận hay trả công cho các tác giả, nghệ sĩ, nhiếp ảnh gia.

Tình huống này gợi nhớ thời kỳ hậu Cách mạng Công nghiệp, khi hàng triệu người lao động bị cuốn vào các nhà máy, hầm mỏ với điều kiện làm việc tồi tệ. Phải mất nhiều thập niên đấu tranh mới có thể bảo vệ được quyền lợi lao động, cải thiện tiền lương và điều kiện làm việc. Ngày nay, một cuộc đấu tranh tương tự đang bắt đầu, mục tiêu là bảo vệ sự sáng tạo, phẩm giá và quyền lợi của con người trước sự xâm lấn của tự động hóa[50]. Phong trào này có thành công hay không vẫn là một câu hỏi còn bỏ ngỏ. Điều đó tùy thuộc vào tất cả chúng ta.

Cuối cùng, trước khi đi sâu vào nội dung từng chương, xin lưu ý rằng đối với các giảng viên và sinh viên sử dụng quyển sách này trong khóa học, chúng tôi đã chuẩn bị sẵn bài tập và tài liệu giảng dạy bổ sung trên trang web AISnakeOil.com[47].

Chương 2

AI dự đoán sai lầm như thế nào

Vào năm 2015, ban quản trị Mount St. Mary's University, một trường đại học tư thục ở Maryland, Hoa Kỳ, mong muốn nâng cao tỷ lệ giữ chân sinh viên, tức tỷ lệ sinh viên được tuyển vào và theo học đến khi tốt nghiệp. Để thực hiện mục tiêu này, nhà trường đã tiến hành một cuộc khảo sát nhằm xác định những sinh viên đang gặp khó khăn. Nghe có vẻ là một nỗ lực đáng trân trọng; nếu biết ai đang gặp trở ngại, ban quản trị có thể hỗ trợ thêm để giúp họ thích nghi với môi trường đại học. Thế nhưng, thay vì hỗ trợ, hiệu trưởng lại đề xuất loại bỏ những sinh viên có kết quả yếu. Theo lý lẽ của ông, nếu các sinh viên này rút lui trong vài tuần đầu học kỳ, họ sẽ không bị tính là "đã nhập học", như vậy sẽ không ảnh hưởng đến tỷ lệ giữ chân của nhà trường.

Trong một cuộc họp với giảng viên, hiệu trưởng nói: "Mục tiêu ngắn hạn của tôi là cho 20-25 người rời trường trước ngày 25 [tháng Chín]. Việc này sẽ giúp nâng tỷ lệ giữ chân sinh viên của chúng ta thêm 4-5%"[1]. Các giảng viên lập tức phản đối,

họ cho rằng rất khó để đánh giá chính xác khả năng thành công của sinh viên chỉ trong vài tuần đầu ở trường đại học. Hiệu trưởng vặc lại: "Điều đó khó với các vị, vì các vị xem sinh viên như những chú thỏ con đáng yêu. Nhưng không thể để mãi như vậy được. Các vị phải học cách... nhấn chìm những chú thỏ đó… Cứ tưởng tượng như đang kê súng Glock vào đầu chúng".

Đây là một ví dụ gây sốc, nhưng thực sự có nhiều trường muốn dự đoán sinh viên nào có nguy cơ bỏ học, một số trường làm điều đó vì lợi ích của sinh viên. EAB Navigate, một sản phẩm dựa trên AI, ra đời để tự động hóa quy trình này. Trong các tài liệu tiếp thị, EAB cam kết rằng: "Mô hình này sẽ mang lại cho nhà trường và các cố vấn những hiểu biết vô giá về khả năng thành công trên con đường học tập của sinh viên, mà không thể có được bằng cách nào khác". Dù một số trường có thể dùng thông tin này để buộc sinh viên ra đi, nhưng những trường khác lại có thể tận dụng nó để thiết kế các biện pháp can thiệp, giúp sinh viên tiếp tục theo học. Tuy nhiên, ngay cả những biện pháp có vẻ tích cực cũng có thể gây tranh cãi. Chẳng hạn, công cụ này có thể đề xuất những chuyên ngành thay thế mà sinh viên được cho là có nhiều khả năng thành công hơn. Điều đó sẽ vô tình đẩy các sinh viên có hoàn cảnh khó khăn hoặc da màu – những người dễ bị công cụ này gắn cờ – ra khỏi các chuyên ngành STEM vốn nhiều tiềm năng nhưng cũng đòi hỏi cao hơn[2]. Và trong toàn bộ quá trình, sinh viên có thể hoàn toàn không hề biết rằng họ đang bị AI âm thầm đánh giá.

ARVIND NARAYANAN & SAYASH KAPOOR

AI - LỢI & HẠI

First News
Where there's a will - there's a way

NHÀ XUẤT BẢN DÂN TRÍ
Số 9 - Ngõ 26 - Phố Hoàng Cầu - Phường Ô Chợ Dừa - TP. Hà Nội
VPGD: Tầng 1, Dãy nhà A trong khuôn viên Công ty Cổ phần
vận tải biển và thương mại Phương Đông,
số 278 Tôn Đức Thắng - Phường Ô Chợ Dừa - TP. Hà Nội
ĐT: (024) 66860751 - (024) 66860752
Email: nxbdantri@gmail.com
Website: nxbdantri.com.vn

Chịu trách nhiệm xuất bản:
BÙI THỊ HƯƠNG
Chịu trách nhiệm nội dung:
LÊ QUANG KHÔI

Biên tập : Phạm Trần Việt Anh
Trình bày : Văn Đông
Bìa : Vũ Thành

Thực hiện liên kết:
CÔNG TY TNHH VĂN HÓA SÁNG TẠO TRÍ VIỆT (First News)
Địa chỉ: 11 I-H Nguyễn Thị Minh Khai, Phường Sài Gòn, TP. HCM

In 3.000 cuốn, khổ 14,5 x 20,5 cm tại Công ty Cổ phần In Khuyến Học Phía Nam
(Lô B5-8 Đường D4, Khu Công Nghiệp Tân Phú Trung, Xã Củ Chi, TP. Hồ Chí Minh)
XNĐKXB số 9-2026/CXBIPH/11-01/DT -
QĐXB số 144/QĐXB-NXBDT cấp ngày 14/01/2026.
In xong và nộp lưu chiểu quý I/2026. ISBN: 978-632-00-2829-0.