# SIÊU TRÍ TUỆ

---

**Biên mục trên xuất bản phẩm của Thư viện Quốc gia Việt Nam**
Bostrom, Nick
Siêu trí tuệ = Superintelligence : AI trỗi dậy và chiến lược ứng phó của loài người trong Kỷ nguyên số / Nick Bostrom ; Dịch: Phạm Hồng Anh, Nguyễn Duy Anh. - H. : Thế giới ; Công ty Sách Alpha, 2025. - 436tr. ; 24cm
ISBN 9786047783397

1. Trí tuệ nhân tạo 2. Kỉ nguyên số
006.3 - dc23
TGK0161p-CIP

---

Góp ý về sách, liên hệ về bản thảo và bản dịch: publication@alphabooks.vn
Liên hệ hợp tác về nội dung số: ebook@alphabooks.vn
Liên hệ hợp tác xuất bản & truyền thông trên sách: publication@alphabooks.vn
Liên hệ dịch vụ tư vấn, đại diện & giao dịch bản quyền: copyright@alphabooks.vn

NICK BOSTROM

# SIÊU TRÍ TUỆ

## SUPERINTELLIGENCE

AI trỗi dậy và chiến lược ứng phó của loài người trong Kỷ nguyên số

Phạm Hồng Anh - Nguyễn Duy Anh *dịch*

alpha books
NHÀ XUẤT BẢN THẾ GIỚI

# LỜI GIỚI THIỆU

Trí tuệ nhân tạo, hay AI (artificial intelligence), là khoa học nhằm làm cho máy tính nói riêng và máy móc nói chung biết hoạt động như có trí thông minh của con người. Với khát vọng này, từ lúc bắt đầu vào giữa những năm 1950 và nhiều thập kỷ tiếp theo, mục tiêu trung tâm của AI là làm sao cho máy biết lập luận và suy diễn (theo logic của con người) và có tri thức (nhờ đưa tri thức con người vào máy). Nhiều năng lực khác của trí thông minh cũng là những mục tiêu cụ thể của AI, đó là làm cho máy có thể nhận biết thế giới bên ngoài như với giác quan con người: nghe (nhận dạng tiếng nói), nhìn (thị giác máy), hiểu (ngôn ngữ tự nhiên), học (để có kiến thức mới), giải quyết vấn đề (như khám chữa bệnh), lập kế hoạch (như lập thời gian biểu thông minh và định giá vé tối ưu của các chuyến bay), hay các robot thông minh…

Những nghiên cứu suốt hơn bảy thập kỷ qua để làm cho máy có trí thông minh và trí tuệ ở mức cao của con người đã tạo nên hướng phát triển "AI bắt chước con người" hay còn gọi "AI ở cấp độ con người". Hướng phát triển này được gọi là "AI tổng quát" (general AI) hoặc "AI mạnh" (strong AI). Một ví dụ là từ ba thập kỷ trước, giới nghiên cứu AI đã đặt mục tiêu vào năm 2050 sẽ chế ra một đội robots đá bóng và thắng đội vô địch World Cup. Đây là mục tiêu tạo một đội bóng gồm các robot di chuyển nhanh trong một môi trường động và chơi bóng hay. Mục tiêu này đòi hỏi tích hợp các kỹ thuật AI và nhiều loại công nghệ với robot thông minh, công nghệ đa tác tử tự trị và hợp

tác, lập luận thời gian thực với dữ liệu từ nhiều cảm biến… Dù đã đạt được những kết quả khích lệ, vẫn còn một chặng đường rất dài để có thể đạt được khát vọng về AI ở cấp độ con người.

Cũng trong quá trình phát triển bảy thập kỷ qua, lĩnh vực học máy (machine learning) của AI – với mục tiêu làm cho máy có thể học như con người và với các phương pháp học quy nạp từ dữ liệu – đã phát triển rất nhanh, giải quyết được nhiều bài toán thực tế, và là phương pháp hiệu quả để hầu hết các lĩnh vực khác của AI như nhận dạng tiếng nói và hình ảnh, vượt qua  hạn chế của các cách tiếp cận truyền thống. Cùng lúc AI "bắt chước con người" được theo đuổi, lĩnh vực về các hệ thống thông minh với tên gọi "cybernetics" (điều khiển học) – nổi tiếng từ cuốn sách cùng tên của Norbert Wiener xuất bản năm 1948 với cảm hứng từ trí thông minh của con người (và của động vật) – cũng liên tục phát triển. Khi các thuật toán học máy từ dựa vào heuristic chuyển dần sang dựa vào các nền tảng toán học vững chắc cũng là lúc các phương pháp học máy trộn với các phương pháp của điều khiển học.

Khi nhận xét rằng các phương pháp của điều khiển học tập trung vào xử lý tín hiệu và ra quyết định cho những công việc có thể phức tạp nhưng không đòi hỏi quá nhiều năng lực trí tuệ cao của con người. Ví dụ như một chiếc xe tự lái chủ yếu cần nhận ra mọi thứ trên con đường, biết đi nhanh đi chậm, biết tránh các xe khác để đi an toàn. Nhiều thành tựu nổi bật khoảng hai mươi năm qua dưới tên "AI" đạt được chủ yếu trong các lĩnh vực như nhận dạng và điều khiển chuyển động, hay thống kê và học máy thống kê nhằm tìm kiếm các mẫu dạng trong dữ liệu và đưa ra các dự đoán có căn cứ, kiểm tra các giả thuyết và đề xuất các quyết định. Các hệ thống tìm kiếm tài liệu, phân loại văn bản, phát hiện gian lận, khuyến nghị hành động, cá thể hoá việc chẩn đoán và chữa bệnh, phân tích mạng xã hội… là những thành công lớn của AI và được theo đuổi bởi các công ty khổng lồ như Google, Netflix, Facebook và Amazon... Những thành công này cho ta chứng kiến trong hai thập kỷ qua nhiều tiến bộ về học thuật và phát triển

sản phẩm cho một hướng phát triển nữa của AI với tên gọi "AI tăng cường trí tuệ". Hướng nghiên cứu này của AI giúp con người giải quyết những bài toán, những nhiệm vụ cụ thể cần đến trí thông minh và sự sáng tạo. Những nhiệm vụ này có thể có ở mọi lĩnh vực hoạt động của con người, từ nông nghiệp, du lịch, thương mại đến y học, giáo dục… Hướng phát triển này được gọi là "AI chuyên dụng" (narrow AI) hoặc "AI yếu" (weak AI). "Yếu" ở đây vì không làm việc vạn năng, nhưng rất "mạnh" ở việc cụ thể phải làm.

Chiến lược phát triển AI của các quốc gia hiện nay chủ yếu tập trung vào AI chuyên dụng. Với lượng dữ liệu ngày càng nhiều và phong phú, các máy tính ngày càng mạnh, và các thuật toán ngày càng hiệu quả, AI chuyên dụng đang đem AI vào cuộc sống con người nhiều hơn bao giờ hết, và là công nghệ số chủ yếu giúp con người sống và làm việc trên môi trường số.

*Siêu trí tuệ* gần hơn với hướng phát triển của AI ở cấp độ con người và mở cho ta những gợi ý, hiểu biết và suy nghĩ về chặng đường dài AI đang tiếp tục đi.

*- GS.TSKH. Hồ Tú Bảo, Viện Nghiên cứu Cao cấp về Toán (VIASM) và Viện John von Neumann (JVN), ĐHQG Tp. HCM*

## Câu chuyện ngụ ngôn dở dang về đàn chim sẻ

Câu chuyện xảy ra vào mùa làm tổ. Sau nhiều ngày lao động vất vả, đàn chim sẻ ngồi trong ánh nắng chiều, thư giãn và chuyện trò ríu rít.

"Chúng ta thật nhỏ bé và yếu đuối. Nghĩ mà xem, cuộc sống sẽ dễ dàng làm sao nếu chúng ta có một cậu cú mèo để giúp xây tổ nhỉ!"

"Phải rồi!" Một con khác nói. "Và mình có thể nhờ cậu ấy chăm sóc các cụ sẻ già và cả lũ sẻ non nữa chứ."

"Cậu ấy sẽ khuyên bảo chúng ta và canh chừng con mèo hàng xóm." Con thứ ba nói.

Lúc này, ông sẻ già Pastus lên tiếng: "Hãy cử trinh sát đi khắp nơi và tìm một cái tổ cú bỏ hoang, một quả trứng, một con quạ con hay một con chồn nhỏ. Đây có thể trở thành điều tuyệt vời nhất mà chúng ta từng có, ít nhất là từ ngày mở cửa Vạn Cốc Lâu ở khu sân sau."

Cả đàn chim phấn chấn, và những chú chim sẻ ở khắp nơi cũng bắt đầu hò reo.

Chỉ có Scronkfinkle, gã chim sẻ chột mắt luôn cáu kỉnh là không mấy tin tưởng vào sự thông thái của việc làm này. Gã nói: "Đó sẽ là một việc làm ngu ngốc. Tại sao chúng ta không nghĩ đến chuyện thuần hóa cú trước khi đưa một thằng cha như thế về đàn?"

Pastus trả lời: "Việc luyện cú này xem ra khó lắm đấy. Tìm được một quả trứng cú cũng đủ khó rồi. Chúng ta cứ bắt đầu như vậy đã. Chúng ta có thể nghĩ về thử thách đó sau khi đã nuôi dạy được một cậu cú."

"Có một lỗ hổng trong kế hoạch này!" Scronkfinkle rít lên; nhưng phản ứng của gã trở nên vô nghĩa khi cả đàn đã cất cánh để bắt đầu thực hiện những chỉ thị của Pastus.

Chỉ còn lại hai ba chú sẻ. Chúng cùng nhau nghĩ cách luyện cú, và gần như ngay lập tức nhận ra rằng Pastus đã đúng: chuyện này thực sự quá khó khăn, nhất là khi không có một con cú thật để thực hành. Mặc dù vậy chúng cũng cố gắng hết sức, vì sợ rằng đàn sẻ sẽ trở về với một quả trứng cú trước khi chúng kịp tìm ra giải pháp cho vấn đề kiểm soát này.

Câu chuyện này không biết sẽ kết thúc ra sao, nhưng tác giả dành tặng cuốn sách này cho gã chim sẻ Scronkfinkle và những chú sẻ đồng hành.

# LỜI NÓI ĐẦU

Bên trong đầu bạn có một thứ đang đọc. Thứ đó, hay chính là bộ não con người, có một vài năng lực mà não bộ của những con vật khác không có, và chính nhờ những năng lực khác biệt ấy mà chúng ta có thể thống trị hành tinh này. Những loài vật khác có cơ bắp khỏe mạnh và móng vuốt sắc nhọn, còn chúng ta lại có một bộ não thông minh hơn. Ưu thế nhỏ này về trí tuệ đã dẫn đến việc chúng ta phát triển ngôn ngữ, công nghệ và tổ chức xã hội phức tạp. Ưu thế đó tích tụ dần theo thời gian, vì mỗi thế hệ lại kế thừa thành quả của những người đi trước.

Nếu ngày nào đó chúng ta chế tạo ra những bộ não máy vượt qua được não người về trí tuệ tổng quát (general intelligence), thì thứ siêu trí tuệ này có thể trở nên vô cùng mạnh mẽ. Và cũng giống như số phận của lũ gorila hiện tại phụ thuộc nhiều vào con người hơn là chính bản thân chúng, thì số phận của chúng ta cũng sẽ phụ thuộc y như vậy vào hành động của những siêu trí tuệ máy.

Thật ra, con người có một ưu thế: chúng ta chế tạo ra những thứ như vậy. Về nguyên tắc chúng ta có thể chế tạo ra một loại siêu trí tuệ có ý thức bảo vệ các giá trị nhân bản, và chắc chắn chúng ta có lý do để làm việc đó. Trên thực tế, vấn đề kiểm soát – phương thức kiểm soát những gì siêu trí tuệ sẽ làm – xem ra khá khó khăn. Chúng ta dường như  chỉ có một cơ hội. Một khi siêu trí tuệ không thân thiện đã tồn tại, nó sẽ ngăn cản chúng ta thay thế nó hay thay đổi những ưu tiên của nó, và khi đó, số phận nhân loại sẽ bị định đoạt.

Trong cuốn sách này, tôi sẽ tìm hiểu thách thức tồn tại trong viễn cảnh siêu trí tuệ, và cách tốt nhất để chúng ta ứng phó với nó. Rất có thể đó sẽ là thách thức quan trọng và khó khăn nhất mà nhân loại từng phải đối mặt, và cho dù thành công hay thất bại, đó có thể cũng sẽ là thách thức cuối cùng dành cho chúng ta.

Cuốn sách này không đưa ra bất kỳ luận điểm nào cho rằng chúng ta đang ở ngưỡng của một đột phá lớn về trí tuệ nhân tạo, hay có thể dự báo thời điểm chuyện này sẽ xảy ra với bất kỳ mức độ chính xác nào. Chuyện đó dường như sẽ diễn ra vào một lúc nào đó trong thế kỷ này, nhưng chúng ta không biết chắc. Các chương đầu thảo luận về những lộ trình có thể và đề cập đôi chút về thời điểm, nhưng phần lớn cuốn sách nói về những gì xảy ra sau thời điểm đó. Chúng tôi nghiên cứu động lực của một cuộc bùng nổ trí tuệ, những dạng thức và quyền năng của siêu trí tuệ, và những lựa chọn chiến lược sẵn có cho một tác tử siêu trí tuệ chiếm được ưu thế quyết định. Sau đó chúng tôi chuyển tiêu điểm sang vấn đề kiểm soát và đặt câu hỏi về việc chúng ta có thể làm để thiết lập những điều kiện ban đầu nhằm sống sót và có được lợi ích vào thời điểm sau cùng. Tới cuối cuốn sách, chúng tôi sẽ nghiền ngẫm bức tranh toàn cảnh được vẽ ra từ những nghiên cứu của mình. Chúng tôi cũng sẽ đưa ra một số khuyến nghị về những việc cần làm nhằm tăng khả năng tránh được một thảm họa mang tính sống còn sau này.

Đây không phải là một cuốn sách dễ viết. Tôi hy vọng rằng con đường đã vạch ra sẽ cho phép những nhà thám hiểm khác vươn tới những ranh giới mới nhanh chóng và dễ dàng hơn, để khi đến đó, họ vẫn còn khỏe khoắn và sẵn sàng tham gia vào việc tiếp tục mở rộng tầm hiểu biết của chúng ta. (Và nếu con đường đã mở có chút gập ghềnh quanh co, tôi hy vọng rằng những người quan sát, khi đánh giá kết quả, sẽ không coi nhẹ những gian truân sẽ gặp phải trên con đường đó!)

Đây cũng không phải là một cuốn sách dễ viết: tôi đã cố làm cho nó dễ đọc, nhưng có lẽ không thành công cho lắm. Khi viết, tôi luôn tâm niệm rằng bản thân mình trong quá khứ chính là một độc giả mục

tiêu và cố gắng viết sao cho chính mình khi đó cũng thích đọc. Giả định như vậy có thể là quá hẹp. Mặc dù vậy, tôi nghĩ rằng nội dung cuốn sách có thể phù hợp cho nhiều người, nếu như họ để tâm một chút và cố gắng đừng để bị cuốn theo việc hiểu sai ngay lập tức từng ý tưởng mới bằng cách đánh đồng nó với một lời khuôn sáo tương tự nào đó trong kho tàng văn hóa của cá nhân mình. Những độc giả "ngoại đạo" cũng không nên nản chí bởi những từ vựng chuyên môn hoặc toán học thi thoảng xuất hiện, vì lúc nào cũng có thể chắt lọc được ý tưởng chính từ những lời giải thích xung quanh. (Ngược lại, những độc giả muốn tìm hiểu cụ thể hơn có thể tìm được nhiều điều trong phần chú thích cuối sách.[1])

Nhiều điểm đưa ra trong cuốn sách này có thể là không đúng.[2] Cũng có thể tôi chưa cân nhắc đến một số điểm trọng yếu, và do đó một số hay toàn bộ kết luận đã trở nên vô giá trị. Tôi đã khá cố gắng để chỉ ra nhiều mức độ không chắc chắn trong suốt cuốn sách và rào đón chúng bằng những từ như "dường như", "có lẽ", "có khả năng", "gần như", "có thể", "rất có thể", "gần như chắc chắn". Các từ hạn định được sử dụng vô cùng thận trọng và dè dặt, mặc dù vậy, sự khiêm tốn về nhận thức này vẫn chưa đủ mà còn cần được bổ trợ bằng việc chấp nhận một cách có hệ thống sự bất định và khả năng xảy ra sai lầm. Đây không phải là sự khiêm tốn giả tạo: trong khi tin rằng cuốn sách của mình có thể đã sai và lạc lối nghiêm trọng, tôi đã nghĩ rằng những quan điểm khác được trình bày trong đó còn tệ hại hơn – bao gồm cả quan điểm mặc định, hay "giả thuyết không" (null hypothesis), mà theo đó, chúng ta tạm thời có thể lờ đi viễn cảnh về siêu trí tuệ một cách an toàn và hợp lý.

# MỤC LỤC

# DANH MỤC HÌNH, BẢNG & KHUNG

**Danh mục hình minh họa**



**Danh mục bảng**

**Danh mục khung**

CHƯƠNG 1

# Sự phát triển trong quá khứ và những năng lực hiện có

**Chúng tôi bắt đầu bằng cách xem xét lại quá khứ. Lịch sử, ở quy mô lớn nhất, có vẻ như là một chuỗi phương thức tăng trưởng khác nhau, trong đó cái sau nhanh hơn nhiều so với cái trước. Theo quy luật này, ta suy ra rằng một hình thức tăng trưởng khác (còn nhanh hơn nữa) hoàn toàn có thể xuất hiện. Tuy nhiên, chúng tôi không quá coi trọng quan sát này – đây không phải là một cuốn sách nói về "sự tăng tốc của công nghệ" hay "sự tăng trưởng theo cấp số nhân" hoặc những ghi nhận tản mạn khác đôi khi được tập hợp lại dưới đề mục "điểm kỳ dị". Sau khi nhìn lại, chúng tôi rà soát lịch sử của trí tuệ nhân tạo, tiếp đó khảo sát những năng lực hiện có của lĩnh vực này. Cuối cùng, chúng tôi tìm hiểu sơ lược kết quả của một số khảo sát ý kiến chuyên gia mới nhất và suy ngẫm về sự thiếu hiểu biết của chúng ta về trục thời gian của những tiến bộ trong tương lai.**

## Các phương thức tăng trưởng và lịch sử nói chung

Vài triệu năm về trước, tổ tiên của chúng ta vẫn còn đang chuyền cành dưới tán rừng châu Phi. Theo thang thời gian địa chất hoặc thậm chí là tiến hóa, sự trỗi dậy của *Homo sapiens* từ tổ tiên chung cuối cùng của chúng ta với các loài linh trưởng lớn đã xảy ra nhanh chóng. Chúng ta phát triển dáng đứng thẳng, ngón cái có khả năng đối trọng khi cầm

nắm, và quan trọng nhất là một vài thay đổi tương đối nhỏ về kích thước bộ não cùng tổ chức thần kinh, tạo ra bước nhảy vọt về năng lực tư duy. Hệ quả là loài người có thể tư duy trừu tượng, bày tỏ những ý nghĩ phức tạp và tích góp thông tin qua nhiều thế hệ tốt hơn rất nhiều so với bất kỳ loài nào khác trên Trái đất.

Những năng lực này cho phép loài người phát triển các "công nghệ" ngày càng hiệu quả, tạo điều kiện cho tổ tiên ta sinh sống ở khu vực cách xa các xa-van và rừng mưa nhiệt đới. Đặc biệt sau khi có nông nghiệp, dân số và mật độ dân cư của loài người đã nhanh chóng tăng lên. Dân số tăng đồng nghĩa với việc có nhiều ý tưởng hơn; và mật độ dân cư cao hơn có nghĩa là ý tưởng có thể lan truyền dễ dàng hơn, một số cá nhân có thể chuyên tâm phát triển những kỹ năng chuyên biệt. Sự phát triển này đẩy nhanh *tốc độ tăng trưởng* về năng suất kinh tế và năng lực công nghệ. Thêm vào đó, những phát triển sau này, có liên quan đến cuộc Cách mạng Công nghiệp, cũng đã mang lại bước tiến thứ hai không kém phần nhanh chóng cho tốc độ tăng trưởng này.

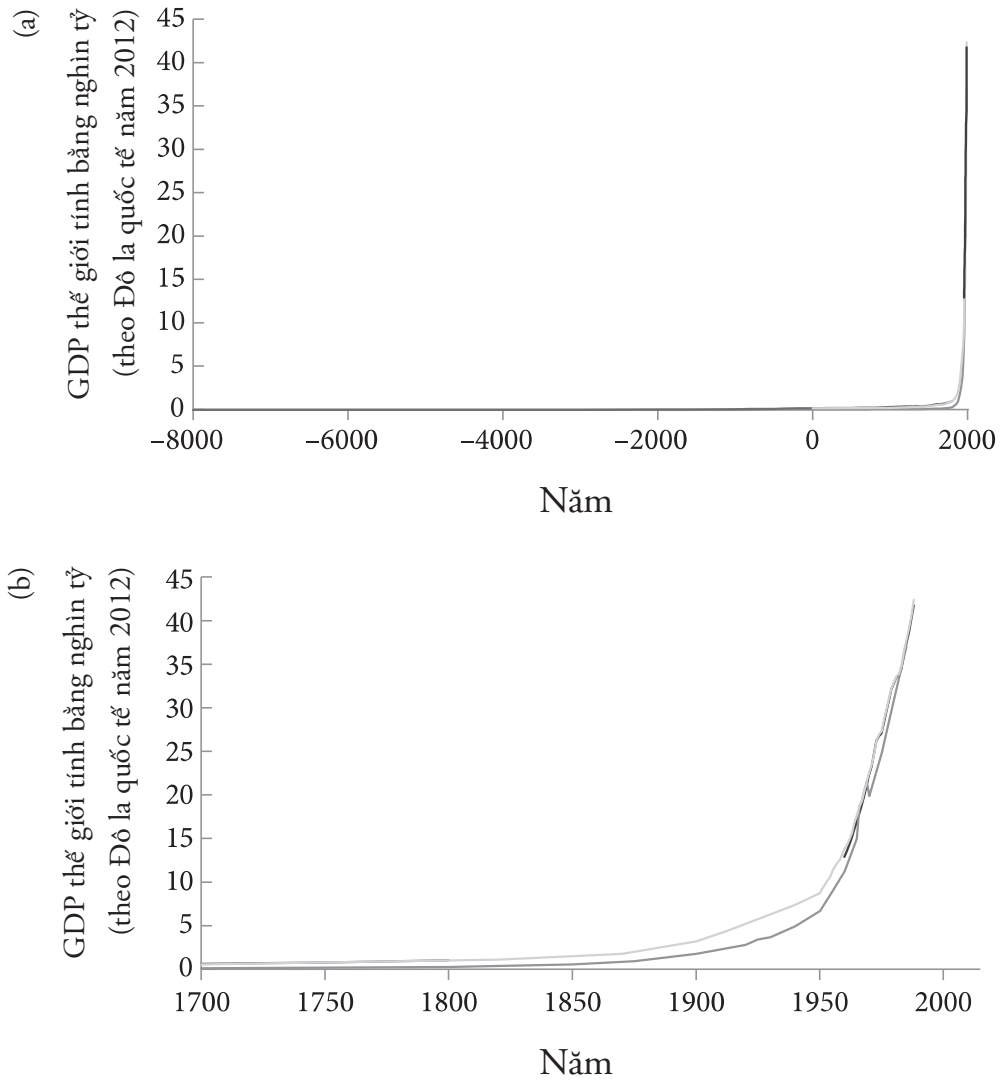
Những thay đổi như vậy về tốc độ tăng trưởng có nhiều hệ quả quan trọng. Trong vài trăm nghìn năm trước, vào đầu thời kỳ tiền sử của loài người (hoặc *hominid* – họ người), tốc độ tăng trưởng chậm đến nỗi cần tới một triệu năm để năng lực sản xuất của loài người tăng lên tới mức đủ nuôi sống thêm một triệu cá thể nữa ở mức độ tối thiểu. Tới năm 5000 TCN, sau Cách mạng Nông nghiệp, tốc độ tăng trưởng tăng vượt bậc, và chỉ cần hai thế kỷ là đã có thể đạt được khối lượng tăng trưởng tương tự. Ngày nay, sau Cách mạng Công nghiệp, nền kinh tế thế giới có thể đạt được khối lượng tăng trưởng như vậy trung bình chỉ sau mỗi 19 phút.[1]

Thậm chí tốc độ tăng trưởng hiện tại cũng sẽ cho những kết quả ấn tượng nếu được duy trì trong một thời gian dài. Nếu nền kinh tế tiếp tục phát triển với tốc độ giống như 50 năm vừa qua thì thế giới sẽ giàu hơn hiện tại 4,8 lần vào năm 2050 và 34 lần vào năm 2100.[2]

Mặc dù vậy, viễn cảnh của việc tiếp tục phát triển bền vững theo cấp số nhân sẽ chẳng thấm vào đâu so với những gì có thể xảy ra nếu thế giới lại một lần nữa trải qua một bước nhảy vọt về *tốc độ tăng trưởng* giống như sau Cách mạng Nông nghiệp và Cách mạng Công nghiệp. Dựa trên cơ sở dữ liệu kinh tế và dân số, nhà kinh tế Robin Hanson đã ước tính khoảng thời gian điển hình để nền kinh tế tăng trưởng gấp đôi trong xã hội săn bắt hái lượm của thế Pleitoscene là 224.000 năm; trong xã hội nông nghiệp là 909 năm; và trong xã hội công nghiệp là 6,3 năm.[3] (Trong mô hình của Hanson, thời kỳ hiện tại chứng kiến sự pha trộn giữa hai phương thức tăng trưởng nông nghiệp và công nghiệp – nền kinh tế thế giới nói chung vẫn chưa tăng trưởng gấp đôi sau 6,3 năm.) Nếu xảy ra sự chuyển dịch tương tự sang một phương thức tăng trưởng khác, và nếu phương thức này có tầm vóc tương đương với hai phương thức trước đó, nó sẽ tạo ra một chế độ tăng trưởng mới mà trong đó quy mô nền kinh tế thế giới sẽ tăng gấp đôi sau mỗi hai tuần.

Theo góc nhìn hiện tại, tốc độ tăng trưởng này có vẻ là một thứ gì đó không tưởng. Những người quan sát trong các thời kỳ trước cũng sẽ cảm thấy phi lý khi nghĩ về việc nền kinh tế thế giới một ngày nào đó sẽ tăng gấp đôi nhiều lần trong cuộc đời một con người, nhưng ngày nay chúng ta lại không coi đó là một điều kỳ lạ.

Ý tưởng về một điểm kỳ dị công nghệ sắp xuất hiện đang được phổ biến rất rộng rãi, khởi đầu bằng tiểu luận quan trọng của Vernor Vinge và tiếp tục với các tác phẩm của Ray Kurweil và những người khác.[4] Tuy nhiên, thuật ngữ "điểm kỳ dị" đang được sử dụng lẫn lộn với nhiều ý nghĩa khác nhau, tạo ra một mớ nghĩa bóng hỗn độn nhưng vẫn ra vẻ hàn lâm và mang đầy tính ảo tưởng công nghệ.[5] Do hầu hết các nghĩa này này không liên quan đến luận điểm của chúng tôi nên thuật ngữ "điểm kỳ dị" đã được loại bỏ để sử dụng thuật ngữ khác chính xác hơn.




**Hình 1** Lịch sử dài hạn của GDP thế giới. Khi được vẽ trên thang đo tuyến tính, lịch sử của nền kinh tế toàn cầu trông giống như một đường thẳng bám sát trục hoành, cho đến khi bất ngờ tăng vọt lên. (a) Ngay cả khi nhìn vào 10.000 năm gần đây nhất, mô hình này về cơ bản vẫn là một góc vuông 90°. (b) Chỉ trong khoảng 100 năm trở lại đây, đường cong mới nhích lên rõ rệt khỏi mức bằng không. (Các đường khác nhau trên biểu đồ tương ứng với các bộ dữ liệu khác nhau, mang lại những ước tính hơi khác nhau.[6])

Ý tưởng liên quan đến điểm kỳ dị mà chúng tôi quan tâm ở đây là khả năng xảy ra một *sự bùng nổ trí tuệ,* cụ thể là trong viễn cảnh của siêu trí tuệ máy. Nhiều người có thể bị thuyết phục bởi những sơ đồ

tăng trưởng như Hình 1 rằng một thay đổi lớn khác trong phương thức tăng trưởng sẽ có thể xảy ra, ngang tầm với Cách mạng Nông nghiệp hoặc Công nghiệp. Sau đó, họ có thể sẽ nghĩ lại rằng kịch bản mà trong đó thời gian để nền kinh tế thế giới tăng trưởng gấp đôi chỉ còn được tính bằng tuần mà không hề liên quan đến việc tạo ra những bộ não nhanh hơn và hiệu quả hơn nhiều so với não sinh học là rất khó tưởng tượng. Tuy nhiên, việc suy nghĩ nghiêm túc về viễn cảnh siêu trí tuệ máy không nhất thiết phải dựa vào các hoạt động đối chiếu biểu đồ hay ngoại suy từ lịch sử phát triển kinh tế. Chúng ta sẽ thấy rằng mình có nhiều lý lẽ thuyết phục hơn cho viễn cảnh này.

## Những kỳ vọng lớn lao

Việc máy móc sánh được với con người về trí tuệ tổng quát (hiểu biết về lẽ thường), khả năng hiệu quả trong việc học tập, lý luận và lên kế hoạch để giải quyết những thách thức về xử lý thông tin phức tạp trong nhiều lĩnh vực tự nhiên và trừu tượng) đã được kỳ vọng từ khi phát minh ra máy tính vào những năm 1940. Khi đó, sự xuất hiện của những cỗ máy như vậy từng được cho rằng chỉ có thể xảy ra vào khoảng 20 năm sau đó.[7] Từ đó trở đi, thời điểm xuất hiện được kỳ vọng này qua mỗi năm lại kéo dài thêm đúng một năm nữa, nên cho đến nay, các nhà tương lai học trong lĩnh vực trí tuệ nhân tạo tổng quát (artificial general intelligence) vẫn tin rằng máy móc thông minh còn cách chúng ta khoảng hai thập kỷ nữa.[8]

Hai thập kỷ là khoảng thời gian ưa thích của những "nhà tiên tri" về các thay đổi toàn diện: đủ gần để thu hút sự chú ý và có tính liên quan, nhưng lại đủ xa để giả định rằng một chuỗi đột phá mà hiện tại chỉ có thể hình dung một cách mờ nhạt có thể sẽ xảy ra. Hãy đối chiếu nó với các thang thời gian ngắn hơn: hầu hết các công nghệ sẽ có tác động lớn trên thế giới trong 5 hay 10 năm tới hiện đều đang được sử dụng một cách hạn chế, còn những công nghệ sẽ định hình lại thế giới trong thời gian dưới 15 năm có thể cũng đã tồn tại dưới dạng

các nguyên mẫu trong phòng thí nghiệm. 20 năm có thể cũng gần với khoảng thời gian còn lại điển hình trong sự nghiệp của các nhà dự báo, gắn liền với rủi ro bị mất uy tín khi đưa ra một dự đoán táo bạo.

Tuy nhiên, việc ai đó đã dự báo quá mức về trí tuệ nhân tạo trong quá khứ không đồng nghĩa với việc AI không thể hoặc sẽ không bao giờ được phát triển.[9] Lý do chính khiến cho quá trình này tiến bộ chậm hơn dự kiến là những khó khăn kỹ thuật trong việc chế tạo máy móc thông minh đã tỏ ra vượt xa dự đoán của những người tiên phong. Tuy nhiên, mức độ chính xác của những khó khăn này và khoảng cách giữa chúng ta hiện nay tới thời điểm giải quyết được chúng vẫn đang là câu hỏi mở. Đôi khi một vấn đề ban đầu có vẻ phức tạp tới mức vô vọng hóa ra lại có một giải pháp đơn giản đến ngạc nhiên (cho dù điều ngược lại có lẽ xảy ra thường xuyên hơn.)

Trong chương tiếp theo, chúng tôi sẽ xem xét những con đường khác nhau có thể dẫn đến một thứ trí tuệ máy ở cấp độ con người. Nhưng xin lưu ý ngay từ đầu rằng cho dù có bao nhiêu điểm dừng từ nay đến khi có được thứ trí tuệ máy đó thì cũng vẫn chưa phải là đích đến cuối cùng. Điểm dừng tiếp theo, chỉ một quãng đường ngắn nữa, là trí tuệ máy cấp độ siêu nhân. Chuyến tàu này có thể không dừng lại, thậm chí là không giảm tốc khi đến Ga Người mà sẽ lao nhanh qua nó.

Nhà toán học I.J. Good, cựu chuyên viên thống kê chính trong nhóm giải mã của Alan Turing trong Thế chiến II, có thể là người đầu tiên nói lên những khía cạnh chính yếu của kịch bản này. Trong một đoạn văn thường được trích dẫn từ năm 1965, ông viết:

> Hãy định nghĩa một cỗ máy siêu trí tuệ là cỗ máy có thể vượt qua mọi hoạt động trí tuệ của bất cứ người nào dù thông minh đến đâu. Vì việc thiết kế máy móc là một trong những hoạt động trí tuệ đó, nên một cỗ máy siêu trí tuệ có thể thiết kế ra những cỗ máy ưu việt hơn; khi đó chắc chắn sẽ có một "sự bùng nổ trí tuệ", và trí tuệ con người sẽ bị tụt hậu rất xa. Vì vậy, cỗ máy siêu trí tuệ đầu tiên là phát minh cuối cùng mà con người cần tạo ra, với điều kiện cỗ máy đủ dễ bảo để nói với chúng ta cách thức kiểm soát chính nó.[10]

Xem ra rõ ràng chúng ta đang có những rủi ro tồn vong đáng kể gắn liền với một sự bùng nổ trí tuệ như vậy, nên viễn cảnh đó cần được nghiên cứu một cách nghiêm túc nhất, ngay cả nếu nó có xác suất xảy ra tương đối nhỏ (mà điều này là không đúng). Tuy nhiên, những người tiên phong về trí tuệ nhân tạo, mặc dù có lòng tin vào AI cấp độ con người, lại hầu như không hề dự tính đến xác suất xuất hiện của AI cao-hơn-con-người. Sức mạnh tư duy của họ như thể đã bị vắt kiệt bởi việc hình dung về khả năng xuất hiện của những cỗ máy có trí tuệ cấp độ con người, đến mức không thể nắm bắt được hậu quả của điều này – rằng máy móc sẽ dần trở thành siêu trí tuệ.

Các nhà tiên phong AI hầu hết không chấp nhận khả năng lĩnh vực của họ có thể có rủi ro.[11] Họ thậm chí không hề bàn tán (chứ chưa nói đến suy nghĩ nghiêm túc) về bất cứ mối quan ngại an toàn hay cảm giác bất ổn về đạo đức nào liên quan đến việc tạo ra những trí não nhân tạo và những cỗ máy tính cai trị tiềm tàng: đây là một khoảng trống đáng kinh ngạc ngay cả trên các tiêu chuẩn đánh giá công nghệ thiết yếu không thực sự ấn tượng của kỷ nguyên này.[12] Chúng ta cần hy vọng rằng đến khi điều này trở thành hiện thực, chúng ta không chỉ có được năng lực công nghệ để châm ngòi cho sự bùng nổ trí tuệ mà cả sự thành thục ở mức độ cao, cần thiết để sống sót qua vụ nổ đó.

Nhưng trước khi hướng tới những gì sẽ xảy ra phía trước, một cái nhìn sơ bộ vào lịch sử của trí tuệ máy cũng sẽ hữu ích cho chúng ta.

## Mùa của hy vọng và tuyệt vọng

Mùa hè năm 1956 tại Đại học Dartmouth, 10 nhà khoa học cùng quan tâm đến mạng lưới thần kinh, lý thuyết Automata và nghiên cứu trí tuệ cùng tổ chức một hội thảo sáu tuần. Dự án Mùa hè Dartmouth (Dartmouth Summer Project) này thường được coi như buổi bình minh của trí tuệ nhân tạo với tư cách là một ngành nghiên cứu. Nhiều người tham gia dự án sau này được công nhận là những nhân vật sáng

lập. Viễn cảnh lạc quan của các đại biểu được thể hiện trong đề xuất trình lên Quỹ Rockefeller, đơn vị tài trợ cho sự kiện:

> Chúng tôi đề nghị thực hiện một nghiên cứu về trí tuệ nhân tạo kéo dài hai tháng với 10 người... Nghiên cứu này được thực hiện trên cơ sở phỏng đoán rằng mọi khía cạnh của tính năng học tập, hay bất cứ tính năng nào khác của trí tuệ, về nguyên tắc đều có thể được mô tả chính xác đến mức tạo ra được một cỗ máy để mô phỏng nó. Chúng tôi nỗ lực tập trung vào việc tìm cách chế tạo những cỗ máy có khả năng sử dụng ngôn ngữ, tạo dựng các khái niệm trừu tượng và cụ thể, giải quyết các vấn đề hiện tại chỉ thuộc về con người và tự cải thiện bản thân. Chúng tôi cho rằng có thể có tiến bộ đáng kể ở một hoặc nhiều vấn đề trong số đó nếu một nhóm các nhà khoa học được lựa chọn cẩn thận cùng làm việc trong một mùa hè.

Trong sáu thập kỷ kể từ sau khởi đầu ồn ào này, lĩnh vực trí tuệ nhân tạo đã trải qua những giai đoạn thăng hoa và đầy kỳ vọng, đan xen với những giai đoạn tụt hậu và thất vọng.

Giai đoạn phấn khích đầu tiên, bắt đầu với hội nghị Dartmouth, sau này được mô John McCarthy (nhà tổ chức hội nghị) mô tả như kỷ nguyên "Mẹ nhìn này, không cần dùng tay nhé!". Trong những ngày tháng ban đầu đó, các nhà nghiên cứu đã chế tạo ra những hệ thống được thiết kế để phản bác lại những tuyên bố kiểu như "Chẳng máy móc nào làm được việc *X* đâu!" Khi đó, những tuyên bố hoài nghi này khá phổ biến. Để phản bác, các nhà nghiên cứu AI tạo ra những hệ thống nhỏ có thể làm được *X* trong một "tiểu thế giới" (một miền hạn chế, được xác định rõ ràng cho phép một phiên bản rút gọn của công việc được thực hiện), qua đó chứng minh được ý tưởng và cho thấy rằng *X*, về nguyên tắc, có thể được máy móc thực hiện. Một trong những hệ thống ban đầu là Logic Theorist (Nhà lý thuyết logic), có thể chứng minh hầu hết các định lý trong chương 2 của cuốn *Principia Mathematica* (Nguyên lý toán học) của Whitehead và Russel, thậm chí còn chứng minh mãn nhãn hơn nhiều so với bản gốc, nhờ đó phê phán luận điểm cho rằng máy móc chỉ có thể "tư duy bằng con số" và cho

thấy máy móc có khả năng suy luận và tạo ra các phép chứng minh logic.[13] Một chương trình theo sau đó là General Problem Solver (Trình giải toán tổng quát) về nguyên tắc có thể giải một số lượng lớn các bài toán được xác định bằng công thức.[14] Những chương trình có thể giải các bài vi-tích phân tiêu biểu trong chương trình đại học năm thứ nhất, các bài toán suy luận trực quan (giống các bài toán xuất hiện trong một số bài kiểm tra IQ) và những bài toán đại số bằng lời đơn giản cũng đã được viết.[15] Robot Shakey (được đặt tên như vậy vì nó thường rung lắc trong lúc vận hành) đã trình diễn cách tích hợp suy luận logic với nhận thức và sử dụng để lập kế hoạch và điều khiển hoạt động vật lý.[16] Chương trình ELIZA cho thấy máy tính có thể bắt chước một nhà trị liệu tâm lý trường phái Roger[a] như thế nào;[17] và vào giữa những năm 1970, chương trình SHRDLU cũng đã cho thấy một cánh tay robot mô phỏng trong một thế giới mô phỏng của các khối hình học có thể làm theo lệnh và trả lời các câu hỏi tiếng Anh do người dùng nhập từ bàn phím.[18] Trong những thập kỷ tiếp theo, nhiều hệ thống tiếp tục được tạo ra và chứng minh việc máy móc có thể soạn nhạc theo phong cách của các nhạc sĩ cổ điển khác nhau, làm việc tốt hơn các bác sĩ trẻ trong một số chẩn đoán lâm sàng nhất định, tự động lái xe và có những sáng tạo có thể đăng ký bản quyền.[19] Thậm chí, một AI còn có thể kể các câu chuyện cười độc đáo.[20] (Điều này không có nghĩa là nó có khiếu hài hước – "Bạn có gì khi kết hợp *nhãn khoa* với *tâm thần học*? Một *nhãn tưởng*[b]"– nhưng bọn trẻ lại thấy những câu chuyện chơi chữ này rất buồn cười).

Những phương pháp thành công trong các hệ thống trình diễn sơ khai này thường rất khó mở rộng cho nhiều dạng bài toán hoặc cho các bài toán khó hơn. Một lý do cho việc này là "sự bùng nổ tổ hợp" trong số các khả năng cần khảo sát bằng phương pháp dựa trên các thao tác như tìm kiếm vét cạn. Các phương pháp này phù hợp với những bài

a. Trị liệu tâm lý hướng cá nhân, do Carl Roger phát triển trong những năm 1940. (ND)

b. Nguyên văn: "eye-dea", phát âm giống "idea" – ý tưởng. (BTV)

toán đơn giản, nhưng thất bại khi tình huống trở nên phức tạp hơn. Ví dụ, để chứng minh một định lý có phép chứng minh dài năm dòng trong một hệ thống diễn dịch với một quy tắc suy luận và năm tiên đề, chỉ cần liệt kê 3.125 tổ hợp có thể có và kiểm tra từng tổ hợp xem nó có cho kết luận mong muốn hay không. Tìm kiếm vét cạn vẫn có thể dùng được cho các chứng minh sáu hoặc bảy dòng, nhưng khi nhiệm vụ trở nên khó khăn hơn, phương pháp này sẽ gặp rắc rối. Chứng minh một định lý với phép chứng minh dài 50 dòng không chỉ cần nhiều thời gian gấp 10 lần chứng minh định lý có 5 dòng, mà nếu sử dụng tìm kiếm vét cạn, ta sẽ phải rà soát $5^{50} \approx 8{,}9 \times 10^{34}$ chuỗi khả dĩ. Đây là một công việc bất khả về mặt tính toán, ngay cả với các siêu máy tính nhanh nhất.

Để vượt qua được sự bùng nổ tổ hợp, người ta cần các thuật toán khai thác được cấu trúc của miền đích và tận dụng kiến thức trước đó bằng phương pháp tìm kiếm tự nghiệm (heuristic), lập kế hoạch và các phép biểu diễn trừu tượng linh hoạt – những năng lực rất kém phát triển trong các hệ thống AI thời kỳ đầu. Hiệu năng của những hệ thống ban đầu này cũng bị ảnh hưởng bởi các phương pháp tệ hại dùng để xử lý bất ổn, sự phụ thuộc vào những phép biểu diễn bằng biểu tượng cứng nhắc và thiếu nền tảng, sự khan hiếm dữ liệu và hạn chế quá lớn của phần cứng về dung lượng bộ nhớ và tốc độ bộ xử lý. Tới giữa những năm 1970, nhận thức về những vấn đề này đã tăng lên. Việc nhận ra rằng nhiều dự án AI có thể không bao giờ thực hiện được những lời hứa ban đầu của mình đã dẫn đến "mùa đông AI" đầu tiên: một giai đoạn thắt lưng buộc bụng, khi mà ngân sách giảm, sự nghi hoặc tăng lên, và AI trở thành lĩnh vực thoái trào.

Một mùa xuân mới lại đến vào đầu những năm 1980, khi Nhật Bản khởi động Fifth-Generation Computer Systems Project (Dự án Hệ thống Máy tính Thế hệ thứ Năm), một dự án hợp tác công-tư với ngân sách lớn nhằm đi tắt đón đầu công nghệ cao bằng cách phát triển một kiến trúc tính toán song song hàng loạt có thể làm nền tảng cho trí tuệ

nhân tạo. Điều này xảy ra một cách đầy màu nhiệm cùng với "phép màu kinh tế hậu chiến" Nhật Bản, giai đoạn mà các chính phủ và các chủ doanh nghiệp phương Tây đang lo lắng kiếm tìm công thức tạo nên sự thành công về kinh tế của Nhật Bản với hy vọng lặp lại phép màu này ở nước mình. Khi Nhật Bản quyết định đầu tư lớn vào AI, một vài quốc gia khác đã lần lượt theo chân.

Những năm tháng sau đó chứng kiến sự sinh sôi nảy nở nhanh chóng của các *hệ chuyên gia* (expert systems). Được thiết kế như các công cụ hỗ trợ cho người ra quyết định, chúng là những chương trình hoạt động theo quy tắc thực hiện những suy luận đơn giản dựa trên cơ sở kiến thức được cung cấp bởi các chuyên gia thuộc nhiều lĩnh vực và mã hóa thủ công bằng một ngôn ngữ hình thức. Hàng trăm hệ chuyên gia như thế đã được chế tạo. Tuy nhiên, những hệ thống nhỏ chỉ đem lại lợi ích không đáng kể, còn những hệ thống lớn hơn cho thấy việc phát triển, kiểm định và cập nhật chúng khá tốn kém, và về cơ bản, chúng tương đối cồng kềnh khi sử dụng. Sẽ thật thiếu thực tiễn nếu mua cả một chiếc máy tính chỉ để chạy một chương trình nào đó. Mùa tăng trưởng này cuối cùng cũng kết thúc vào cuối những năm 1980.

Dự án Thế hệ thứ Năm không đạt được mục tiêu, cũng như những dự án tương tự ở Mỹ và châu Âu. Mùa đông AI thứ hai lại đến. Tới lúc này, một nhà phê bình có thể than thở đầy chính đáng rằng "lịch sử nghiên cứu trí tuệ nhân tạo cho tới lúc này luôn chỉ có những thành công rất hạn chế trong những lĩnh vực hẹp, và ngay lập tức được nối tiếp bằng thất bại trong việc đạt được những mục tiêu rộng lớn hơn mà những thành công ban đầu tưởng như đã khơi dậy".[21] Các nhà đầu tư tư nhân bắt đầu lảng tránh mọi công ty mang theo thương hiệu "trí tuệ nhân tạo". Thậm chí trong giới hàn lâm và những nhà tài trợ, "AI" đã trở thành một biệt danh không mong muốn.[22]

Tuy nhiên, việc nghiên cứu kỹ thuật đã tiếp diễn nhanh chóng, và đến những năm 1990, mùa đông AI thứ hai đã dần ấm lên. Sự lạc quan được nhóm lên nhờ những kỹ thuật mới, có vẻ như đem lại các

phương án thay thế cho mô hình logic truyền thống (thường được gọi là "Good Old-Fashioned Artificial Intelligence – Trí tuệ Nhân tạo Kiểu cũ", hay "GOFAI"), vốn tập trung vào việc vận dụng ký hiệu cấp độ cao và đã đạt đến đỉnh cao trong các hệ chuyên gia của những năm 1980. Những kỹ thuật mới được phổ biến, bao gồm mạng lưới thần kinh và các thuật toán di truyền, hứa hẹn bù đắp được một số thiếu hụt của cách tiếp cận GOFAI, cụ thể là "tính dễ đổ vỡ" (brittleness) đặc trưng của các chương trình AI (thường tạo ra những thứ hoàn toàn vô nghĩa nếu các lập trình viên chỉ cần có một giả định hơi sai lệch). Những kỹ thuật mới có phương pháp vận hành mang tính tổ chức hơn. Ví dụ, các mạng lưới thần kinh có thuộc tính "suy giảm thanh nhã" (hay thoái hóa không đáng kể – graceful degradation): một hư hại nhỏ của mạng lưới thần kinh thường chỉ gây ra một suy giảm nhỏ về hiệu năng của nó thay vì làm sập hoàn toàn. Quan trọng hơn nữa, các mạng lưới thần kinh có thể học hỏi từ kinh nghiệm, tìm kiếm những cách thức tự nhiên để khái quát hóa từ các ví dụ và tìm ra những quy luật thống kê ẩn giấu trong dữ liệu đầu vào.[23] Điều này khiến chúng  nhận ra quy luật và phân loại vấn đề một cách hiệu quả. Ví dụ, thông qua đào tạo bằng một tập dữ liệu tín hiệu siêu âm, một mạng lưới thần kinh có thể được dạy để phân biệt các đặc điểm âm thanh của tàu ngầm, mìn và động thực vật biển tốt hơn các chuyên gia – và điều đó có thể thực hiện được mà không cần có người tìm ra trước cách thức xác định các phạm trù phân loại và đánh giá trọng số của các đặc tính khác nhau.

Trong khi các mô hình mạng lưới thần kinh đơn giản đã được biết đến từ cuối những năm 1950, lĩnh vực này được phục hưng sau khi thuật toán lan truyền ngược được giới thiệu, nhờ đó các mạng lưới thần kinh đa lớp có thể được đào tạo.[24] Các mạng đa lớp (có một hay nhiều lớp neuron trung gian "ẩn" giữa các lớp đầu vào và đầu ra) này có thể học nhiều loại chức năng hơn các "tiền bối" đơn giản hơn của chúng.[25] Kết hợp với các máy tính đang ngày càng mạnh hơn và sẵn có hơn, những cải thiện về thuật toán này cho phép các kỹ sư chế tạo ra các mạng lưới thần kinh đủ tốt để có thể hữu ích trong nhiều ứng dụng.

Những phẩm chất tương tự não bộ của mạng lưới thần kinh vượt trội hơn so với các hệ thống GOFAI truyền thống dựa trên quy tắc vốn có phương pháp hoạt động logic cứng nhắc nhưng dễ đổ vỡ. Nó đủ tốt để truyền cảm hứng cho một trường phái mới, *kết nối luận* (connectionism), nhấn mạnh vào tầm quan trọng của phương pháp xử lý biểu tượng thứ cấp (sub-symbolic) song song. Hơn 150.000 bài báo khoa học đã được công bố về mạng lưới thần kinh nhân tạo, và đây tiếp tục là cách tiếp cận quan trọng trong học máy (machine learning).

Các phương pháp dựa trên tiến hóa như thuật toán di truyền và lập trình di truyền tạo nên một cách tiếp cận khác đã giúp chấm dứt mùa đông AI thứ hai khi xuất hiện. Cách tiếp cận này có thể có ít ảnh hưởng về mặt hàn lâm hơn mạng lưới thần kinh nhưng lại được phổ biến rộng rãi. Trong các mô hình tiến hóa, một tập hợp các giải pháp ứng viên (có thể là các cấu trúc dữ liệu hay chương trình) được duy trì, và những giải pháp ứng viên mới được ngẫu nhiên tạo ra bằng cách đột biến hoặc tái tổ hợp các biến thể trong quần thể giải pháp hiện có. Định kỳ, quần thể này được xén tỉa bằng cách áp dụng một tiêu chí lựa chọn (một hàm thích nghi – fitness function) chỉ cho phép những ứng viên tốt hơn sống sót sang thế hệ sau. Được lặp lại qua hàng ngàn thế hệ, chất lượng trung bình trong nhóm ứng viên dần tăng lên. Khi có tác dụng, loại thuật toán này có thể tạo ra những giải pháp hiệu quả cho rất nhiều loại vấn đề. Những giải pháp này có thể rất mới và phi trực giác, thường trông giống những cấu trúc tự nhiên hơn bất cứ thứ gì mà các kỹ sư có thể thiết kế ra. Và trên nguyên tắc, điều đó có thể diễn ra mà không cần có thông tin đầu vào của con người vượt quá các tham số ban đầu của hàm thích nghi, thường là rất đơn giản. Tuy nhiên, trên thực tế, để làm cho các phương pháp tiến hóa có tác dụng tốt, cần phải có kỹ năng và sự khéo léo, cụ thể là khi đưa ra một định dạng biểu diễn tối ưu. Nếu không có một cách thức hiệu quả để mã hóa các giải pháp ứng viên (một ngôn ngữ di truyền khớp với cấu trúc tiềm tàng trong miền mục tiêu), việc tìm kiếm tiến hóa thường vòng vèo bất tận trong

một không gian tìm kiếm khổng lồ hoặc bị kẹt với một giá trị tối ưu cục bộ. Thậm chí nếu tìm được một định dạng biểu diễn tốt, tiến hóa cũng đòi hỏi nhiều công suất tính toán và thường bị đánh bại bởi sự bùng nổ tổ hợp.

Các mạng lưới thần kinh và thuật toán di truyền là những ví dụ về các phương pháp đã tạo nên sự phấn khích trong những năng 1990 chỉ bằng việc (có vẻ như) đưa ra được những lựa chọn thay thế mô hình GOFAI trì trệ. Nhưng ý định ở đây không phải là ca ngợi hai phương pháp này hay nâng chúng lên cao hơn nhiều kỹ thuật khác trong học máy. Thực ra, một trong những sự phát triển lớn về lý thuyết của 20 năm qua là sự nhận thức rõ ràng hơn về cách mà các kỹ thuật bề ngoài riêng rẽ được hiểu như những trường hợp đặc biệt trong một khuôn khổ toán học chung. Ví dụ, nhiều loại mạng lưới thần kinh nhân tạo có thể được xem như các bộ phân loại (classifier) thực hiện một loại hình tính toán thống kê cụ thể (ước lượng xác suất tối đa).[26] Cách nhìn này cho phép mạng lưới thần kinh sánh ngang với một lớp thuật toán lớn hơn cho các bộ phân loại quá trình học từ các ví dụ, như "cây quyết định", "mô hình hồi quy logistic", "máy vector hỗ trợ", "thuật toán naive Bayes", "hồi quy k-nearest-neighbor," và các loại khác.[27] Tương tự, các thuật toán di truyền có thể được xem xét như đang thực hiện việc "leo đồi ngẫu nhiên" (stochastic hill-climbing), và chúng một lần nữa lại là một tập nhỏ trong một lớp rộng hơn của các thuật toán tối ưu. Mỗi thuật toán dùng cho việc xây dựng các bộ phân loại hay tìm kiếm một không gian giải pháp này có các điểm mạnh, điểm yếu riêng có thể được nghiên cứu bằng toán học. Các thuật toán khác nhau về yêu cầu thời gian của bộ xử lý và không gian bộ nhớ với những thiên kiến quy nạp (inductive bias), về sự dễ dàng trong việc tích hợp nội dung được tạo ra từ bên ngoài, và về sự khả kiến trong hoạt động nội bộ của chúng đối với một người phân tích.

Do đó, đứng sau những ồn ào của học máy và giải quyết vấn đề một cách sáng tạo là một tập hợp những đánh đổi được xác định rõ

ràng về mặt toán học. Tập hợp lý tưởng là tập hợp của tác tử Bayes (Bayesian agent) hoàn hảo, sử dụng thông tin sẵn có một cách tối ưu về mặt xác suất. Tập hợp này là không thể có được vì nó đòi hỏi quá cao về mặt tính toán để có thể triển khai được trong bất cứ máy tính vật lý nào (xem Khung 1). Theo đó, ta có thể coi trí tuệ nhân tạo như một cuộc tìm kiếm lối tắt: những cách thức để mô phỏng một cách dễ kiểm soát một tác tử Bayes lý tưởng bằng cách hy sinh phần nào tính tối ưu hoặc tính tổng thể trong khi vẫn bảo tồn đủ để có được hiệu năng cao trong những miền thực sự cần quan tâm.

Các công trình được thực hiện trong một vài thế kỷ qua về những mô hình đồ họa xác suất, chẳng hạn như mạng Bayes là một sự phản ảnh về bức tranh này. Các mạng Bayes cho ta một phương pháp súc tích để biểu diễn các mối quan hệ độc lập có điều kiện và xác suất đúng trong một số miền nhất định. (Khai thác những mối quan hệ độc lập này là rất quan trọng để vượt qua được sự bùng nổ tổ hợp, vốn là vấn đề của cả uy luận xác suất cũng như diễn dịch logic). Chúng cũng cung cấp những hiểu biết quan trọng về khái niệm tính nhân quả.[34]

**Khung 1 Tác tử Bayes tối ưu**

Một tác tử Bayes lý tưởng khởi đầu với một "phân bổ xác suất tiên nghiệm", một hàm gán xác suất cho từng "thế giới khả thể" (nghĩa là cho từng cách thức cụ thể mà thế giới có thể trở thành).[28] Phân bố tiên nghiệm này tạo ra một thiên kiến quy nạp cho rằng những "thế giới khả thể" đơn giản hơn được ấn định xác suất cao hơn. (Một cách để định nghĩa chính thức sự đơn giản của một thế giới khả thể là nói về "độ phức tạp Kolmogorov" của nó, một số đo dựa trên độ dài của chương trình máy tính ngắn nhất có thể tạo ra một mô tả đầy đủ về thế giới đó.[29]) Tiên nghiệm còn tích hợp mọi kiến thức nền tảng mà lập trình viên muốn cung cấp cho tác tử.

Khi tác tử nhận thông tin mới từ các cảm biến của mình, nó cập nhật phân bố xác suất bằng cách điều chỉnh lại phân bố này theo thông tin mới theo định lý Bayes.[30] Điều kiện hóa (conditionalization) là phép toán thiết lập xác suất mới của những thế giới không nhất quán với thông tin nhận được để

xóa và tái chuẩn hóa phân bố xác suất của những thế giới khả thể còn lại. Kết quả của phép toán này là một "phân bố xác suất hậu nghiệm" (mà tác tử có thể dùng như tiên nghiệm của nó trong bước tiếp theo). Khi tác tử quan sát, khối lượng xác suất của nó tập trung vào một tập đang thu hẹp lại của các thế giới khả thể vẫn còn phù hợp với bằng chứng; và trong tập hợp đó, những thế giới đơn giản hơn luôn có xác suất lớn hơn.

Để hình dung, ta có thể so sánh xác suất với cát trên một tờ giấy lớn. Tờ giấy được chia thành các vùng có kích thước khác nhau, mỗi vùng tương ứng với một thế giới khả thể, vùng có diện tích lớn hơn tương ứng với những thế giới đơn giản hơn. Hãy hình dung thêm một lớp cát có độ dày đồng đều được rải lên cả tờ giấy: đó chính là phân bố xác suất tiên nghiệm của chúng ta. Mỗi khi một quan sát được thực hiện và loại bỏ một số thế giới, chúng ta gạt cát ra khỏi vùng tương ứng của tờ giấy và phân phối lại cát đồng đều trên những vùng còn lại. Như vậy tổng lượng cát trên tờ giấy không thay đổi, mà chỉ tập trung vào ít vùng hơn khi bằng chứng quan sát được tích luỹ. Đây là bức tranh về quá trình học ở dạng thức thuần khiết nhất. (Để tính toán xác suất của một *giả thuyết*, ta chỉ cần đo lượng cát ở tất cả các vùng tương ứng với những thế giới khả thể mà trong đó giả thuyết là đúng).

Cho tới nay, chúng ta đã định nghĩa được một quy tắc học tập. Để có một tác tử, ta còn cần một quy tắc quyết định. Để có quy tắc này, ta trao cho tác tử một "hàm thỏa dụng" có chức năng gán một con số cho mỗi thế giới khả thể. Con số này biểu diễn mức độ đáng mong đợi (desirability) của thế giới đó theo những ưu tiên cơ bản của tác tử. Bây giờ, trong từng bước, tác tử lựa chọn hành động với độ thỏa dụng kỳ vọng cao nhất.[31] (Để tìm hành động với độ thỏa dụng kỳ vọng cao nhất, tác tử có thể liệt kê tất cả các hành động khả thể. Sau đó, nó có thể tính toán phân bố xác suất có điều kiện khi đã xảy ra hành động đó – là phân bố xác suất được tạo ra từ việc điều kiện hóa phân bố xác suất hiện tại dựa trên quan sát rằng hành động đã được thực hiện. Cuối cùng nó có thể tính toán giá trị kỳ vọng của hành động là tổng giá trị của từng thế giới khả thể nhân với xác suất có điều kiện của thế giới đó khi đã có hành động.[32])

Quy tắc học tập và quy tắc quyết định cùng nhau xác định một "khái niệm về tính tối ưu" cho tác tử. (Về bản chất, khái niệm về tính tối ưu này có thể được sử dụng rộng rãi trong trí tuệ nhân tạo, nhận thức luận, triết lý khoa học, kinh tế và thống kê.[33]) Trên thực tế, không thể xây dựng được một

tác tử như vậy vì về mặt tính toán, ta không thể thực hiện được những phép toán cần thiết. Bất cứ nỗ lực nào làm việc đó đều không chống lại được sự bùng nổ tổ hợp đã được mô tả trong phần thảo luận về GOFAI. Để thấy được nguyên nhân của điều này, hãy xem xét một tập hợp rất nhỏ của tất cả các thế giới khả thể: những thế giới gồm có một màn hình máy tính duy nhất trôi nổi trong khoảng chân không vô hạn. Màn hình này có độ phân giải 1.000 x 1.000 điểm ảnh, và mỗi điểm ảnh luôn có thể bật hoặc tắt. Ngay cả tập hợp nhỏ này của những thế giới khả thể cũng đã vô cùng lớn: $2^{(1.000 \times 1.000)}$ trạng thái khả thể của màn hình, vượt qua mọi tính toán có thể xảy ra trong vũ trụ quan sát được. Do vậy, chúng ta thậm chí còn không thể đếm hết tất cả các thế giới khả thể trong tập hợp nhỏ của toàn bộ các thế giới khả thể này, chứ chưa nói đến việc thực hiện những tính toán chi tiết hơn trong từng thế giới riêng biệt.

Các khái niệm về tính tối ưu có thể trở nên hấp dẫn về lý thuyết ngay cả nếu chúng không thể hiện thực hóa được. Các khái niệm này cho chúng ta một tiêu chuẩn để đánh giá những phép xấp xỉ tự nghiệm, và đôi khi, ta có thể lý luận về những gì một tác tử tối ưu có thể làm trong một số trường hợp đặc biệt. Chúng ta sẽ gặp một số khái niệm về tính tối ưu khác cho các tác tử nhân tạo trong Chương 12.

Một ưu điểm của việc liên kết vấn đề học tập từ những miền cụ thể với vấn đề chung của suy luận Bayes là những thuật toán mới làm cho suy luận Bayes hiệu quả hơn sau đó sẽ tạo nên những cải thiện tức thời trong nhiều lĩnh vực khác nhau. Ví dụ, những tiến bộ trong các kỹ thuật tính xấp xỉ của Monte Carlo đang được áp dụng trực tiếp trong thị giác máy, robotic và di truyền học điện toán. Một ưu điểm khác là nó cho phép các nhà nghiên cứu từ các lĩnh vực khác nhau hợp nhất kết quả nghiên cứu của mình dễ dàng hơn. Những mô hình đồ họa và thống kê Bayes trở thành tiêu điểm nghiên cứu chung trong nhiều lĩnh vực, bao gồm học máy, vật lý thống kê, tin học sinh học, tối ưu hóa tổ hợp và lý thuyết truyền thông.[35] Những tiến bộ đáng kể gần đây trong học máy có được từ việc tổ hợp các kết quả chính thống khởi phát trong các lĩnh vực học thuật khác. (Các ứng dụng học máy cũng có

được lợi ích rất lớn từ những máy tính nhanh hơn và tính sẵn sàng cao hơn của các tập dữ liệu lớn.)

## Những công nghệ tiên tiến

Trí tuệ nhân tạo đã vượt trí tuệ con người trong nhiều lĩnh vực. Bảng 1 khảo sát tình trạng của các máy tính chơi trò chơi và cho thấy hiện nay AI đã thắng được các nhà vô địch trong nhiều trò.[36]

Ngày nay những thành tích này có vẻ không thực sự ấn tượng. Nhưng đó chỉ là vì tiêu chuẩn của chúng ta về những thứ ấn tượng vẫn thường xuyên thích ứng với những tiến bộ mới. Ví dụ, chơi cờ vua cấp kiện tướng từng được cho là hình ảnh thu nhỏ của trí tuệ con người. Theo quan điểm của một số kiện tướng cuối những năm 1950: "Nếu ai đó có thể chế tạo thành công một chiếc máy chơi cờ, người đó có thể được xem như đã thâm nhập được vào cốt lõi của trí tuệ nhân loại."[37] Mọi việc hiện giờ xem ra không còn như vậy nữa. Chúng ta có thể đồng cảm với John McCarthy, người từng cảm thán rằng: "Ngay sau khi nó làm được việc, sẽ không ai còn gọi nó là AI nữa."[38]

Tuy nhiên, trong việc này còn có một ý nghĩa quan trọng khác. Đó là việc AI chơi cờ vua hóa ra lại là một "chiến thắng" khiêm tốn so với những gì nhiều người vẫn tưởng. Người ta từng giả định, có thể cũng không hoàn toàn vô căn cứ, rằng để máy tính có thể chơi được cờ vua ở cấp đại kiện tướng, nó phải được trang bị trí tuệ tổng quát ở cấp độ cao.[39] Ví dụ, người ta có thể nghĩ rằng để chơi cờ giỏi cần có khả năng học những khái niệm trừu tượng, suy nghĩ thông minh về chiến lược, lập những kế hoạch linh hoạt, thực hiện nhiều phép suy luận logic độc đáo và thậm chí có thể là lập mô hình phương thức tư duy của đối thủ. Tuy nhiên mọi thứ không phải như vậy. Ta có thể xây dựng một công cụ chơi cờ vua giỏi hoàn hảo từ một thuật toán có mục đích riêng biệt.[40] Khi được thực thi trên các bộ xử lý nhanh xuất hiện trong những năm cuối thế kỷ 20, nó đã tạo ra một năng lực chơi cờ mạnh

mẽ. Nhưng một AI được chế tạo như vậy lại chỉ có khả năng hạn hẹp: Nó biết chơi cờ nhưng không làm được gì khác.[41]

Trong những lĩnh vực khác, các giải pháp hóa ra lại phức tạp *hơn* so với kỳ vọng ban đầu và cũng tiến triển chậm hơn. Nhà khoa học máy tính Donald Knuth đã ngạc nhiên với việc

**Bảng 1** ***AI chơi trò chơi***

| | | |
|---|---|---|
| Checkers | Nhà vô địch | Chương trình checkers của Arthur Samuel được viết lần đầu tiên năm 1952 và sau đó được cải tiến (phiên bản 1955 đã được kết hợp học máy), trở thành chương trình đầu tiên học được cách chơi giỏi hơn người tạo ra nó.[42] Năm 1994, chương trình CHINOOK đánh bại nhà đương kim vô địch, đánh dấu lần đầu tiên một chương trình chiến thắng một giải vô địch thế giới chính thức trong một trò chơi kỹ năng. Năm 2002, Jonathan Schaeffer và nhóm của anh ta "giải" được checkers, nghĩa là tạo ra được một chương trình luôn đưa ra được nước đi tốt nhất có thể (kết hợp tìm kiếm alpha-beta với cơ sở dữ liệu với 39 ngàn tỷ thế tàn cục có thể). Ván cờ hoàn hảo của cả hai bên dẫn đến tỷ số hòa.[43] |

*(Xem tiếp trang sau)*

| | | |
|---|---|---|
| **Backgammon** | Nhà vô địch | 1979: chương trình backgammon BKG của Hans Berliner đánh bại nhà vô địch thế giới – chương trình máy tính đầu tiên đánh bại một nhà vô địch thế giới trong tất cả các ván (trong một trận đấu giao hữu) – mặc dù sau này, Berliner cho rằng chiến thắng là nhờ may mắn khi ném xúc xắc.[44]<br>1992: chương trình chơi backgammon là TD-Gammon của Gerry Tesauro đạt trình độ vô địch, sử dụng học tập chênh lệch thời gian (temporal difference learning, một dạng học tăng cường) và nhiều ván chơi lặp lại với chính nó để nâng cao trình độ.[45]<br>Trong những năm sau đó, các chương trình chơi backgammon đã vượt xa những người chơi giỏi nhất.[46] |
| **Traveller TCS** | Nhà vô địch kết hợp với người bình thường[47] | Trong cả hai năm 1981 và 1982, chương trình Eurisko của Douglas Lenat đã giành chức vô địch trong giải đấu Traveller TCS (một trò chơi hải chiến tương lai) của Mỹ, dẫn đến việc thay đổi luật chơi để ngăn chặn những chiến lược khác thường của nó.[48] Eurisko có tự nghiệm để thiết kế hạm đội của riêng mình và có cả tự nghiệm để điều chỉnh tự nghiệm của chính mình. |
| **Othello** | Nhà vô địch | 1997: Chương trình Logistello thắng tất cả các ván trong một trận đấu sáu ván với nhà vô định thế giới Takeshi Murakami.[49] |
| **Cờ vua** | Nhà vô địch | 1997: Deep Blue đánh bại nhà vô địch cờ vua thế giới Garry Kasparov. Kasparov nói rằng đã thoáng nhìn thấy trí tuệ và sự sáng tạo thực sự trong một số nước đi của máy tính này.[50] Từ đó trở đi, các chương trình đánh cờ tiếp tục được cải thiện.[51] |

| | | |
|---|---|---|
| **Giải ô chữ** | Chuyên gia | 1999: chương trình giải ô chữ Proverb đã chơi tốt hơn những người chơi giải ô chữ trung bình.[52]<br>2012: Chương trình Dr. Fill do Matt Ginsberg tạo ra đã ghi được điểm số nằm trong top 25% dẫn đầu trước những đối thủ là con người trong Giải đầu Giải ô chữ của Mỹ. (Phong độ của Dr. Fill không ổn định. Nó giải hoàn hảo ô chữ được con người đánh giá là khó nhất, nhưng lại gặp khó khăn với một vài ô chữ không tiêu chuẩn theo lối đánh vần ngược hoặc viết câu trả lời theo đường chéo.)[53] |
| **Scrabble** | Nhà vô địch | Tính đến 2002, phần mềm chơi Srabble đã vượt qua những người chơi giỏi nhất.[54] |
| **Bài bridge** | Tương đương người chơi giỏi nhất | Tính đến 2005, phần mềm chơi bài bridge đạt đến trình độ của những người chơi giỏi nhất.[55] |
| **Jeopardy!** | Nhà vô địch | 2010: Watson của IBM đánh bại hai nhà vô địch trò chơi *Jeopardy!* vĩ đại nhất mọi thời đại là Ken Jennings và Brad Rutter.[56] *Jeopardy!* là trò chơi truyền hình với những câu hỏi nhỏ về lịch sử, văn học, thể thao, địa lý, văn hóa đại chúng, khoa học và những chủ đề khác. Những câu hỏi được đặt dưới dạng gợi ý và thường hay chơi chữ. |
| **Poker** | Trình độ đa dạng | Máy tính chơi poker vẫn yếu hơn đôi chút so với những người chơi giỏi nhất trong trò full-ring Texas hold 'em nhưng chơi ở trình độ vô địch trong một vài kiểu chơi poker khác.[57] |
| **FreeCell** | Nhà vô địch | Tự nghiệm sử dụng các thuật toán di truyền đã tạo ra một chương trình chơi trò solitaire FreeCell (dạng tổng quát là NP-complete) có thể đánh bại những người chơi có thứ hạng cao.[58] |

| | | |
|---|---|---|
| **Cờ vây** | Cao thủ nghiệp dư | Tính đến năm 2012, các chương trình chơi cờ vây trong series Zen đã đạt đến trình độ lục đẳng trong trò cờ nhanh (trình độ của những cao thủ cờ vây nghiệp dư), sử dụng các kỹ thuật tìm kiếm cây (tree search) Monte Carlo và học máy.[59] Các chương trình chơi cờ vây được cải thiện với tốc độ một đẳng/năm trong những năm gần đây. Nếu tốc độ tiến bộ này tiếp tục được duy trì, chúng sẽ thắng nhà vô địch thế giới trong khoảng một thập kỷ nữa. |

"AI cho tới nay đã thành công khi làm hầu hết mọi thứ đòi hỏi 'tư duy' nhưng không làm được hầu hết những thứ mà người và động vật có thể làm mà 'không cần suy nghĩ' – điều này, bằng cách nào đó, hóa ra lại khó hơn nhiều!"[60] Việc phân tích những khung cảnh trực quan, nhận diện đối tượng hoặc kiểm soát hành vi của một robot khi tương tác với môi trường tự nhiên tỏ ra là những việc vô cùng thách thức. Mặc dù vậy, chúng ta đã và đang đạt được nhiều tiến bộ, và những tiến bộ này được bổ trợ bởi những cải thiện phần cứng đáng kể.

Hiểu lẽ thường và hiểu ngôn ngữ tự nhiên cũng là những vấn đề khó khăn. Giờ đây, người ta thường nghĩ rằng việc đạt trình độ hoàn toàn ngang với con người khi thực thi những nhiệm vụ này là vấn đề "AI toàn diện" (AI-complete), có nghĩa là độ khó của việc giải quyết những vấn đề này về cơ bản tương đương với độ khó của việc chế tạo ra những chiếc máy có trí tuệ tổng quát cấp độ con người.[61] Nói cách khác, nếu ai đó thành công trong việc tạo ra một AI có thể hiểu ngôn ngữ tự nhiên tốt như một người lớn thì họ hoàn toàn có thể tạo ra một AI làm được mọi việc khác mà con người có thể, hoặc họ đã đến được rất gần với một năng lực tổng thể như vậy.[62]

Trình độ cờ vua vô địch có thể đạt được bằng những thuật toán đơn giản đến ngạc nhiên. Thật thú vị khi nghĩ rằng những năng lực

khác – như khả năng lý luận chung hay một số khả năng liên quan đến lập trình – cũng có thể đạt được một cách tương tự nhờ một thuật toán vô cùng đơn giản nào đó. Việc có được hiệu năng cao nhất trong một nhiệm vụ nhất định tại một thời điểm nhất định bằng một cơ chế phức tạp không đồng nghĩa với việc không có cơ chế đơn giản nào có thể thực hiện nhiệm vụ đó với hiệu năng ngang bằng hay thậm chí cao hơn. Có thể chỉ đơn giản là chưa ai tìm ra cách thay thế tốt hơn mà thôi. Hệ Ptolemy (với Trái đất ở trung tâm, xung quanh là Mặt trời, Mặt trăng, các hành tinh và những ngôi sao) đại diện cho trình độ phát triển thiên văn học suốt hơn 1.000 năm, và độ chính xác dự báo của nó được dần cải thiện bằng cách phức tạp hóa mô hình: Hết đường ngoại luân này tới đường ngoại luân khác được bổ sung để dự đoán chuyển động thiên thể. Sau đó, toàn bộ hệ thống đã bị lật đổ bởi thuyết nhật tâm của Copernicus. Thuyết này đơn giản hơn và chính xác hơn về mặt dự báo (tuy chỉ sau khi được Kepler phát triển thêm).[63]

Các phương pháp trí tuệ nhân tạo hiện đang được sử dụng trong quá nhiều lĩnh vực để có thể xem xét tất cả ở đây, nhưng một ví dụ sẽ cho ta ý niệm về sự trải rộng của các ứng dụng. Ngoài những AI chơi trò chơi được liệt kê trong Bảng 1 còn có những thiết bị trợ thính với các thuật toán lọc tiếng ồn xung quanh; những bộ tìm đường hiển thị bản đồ và chỉ đường cho lái xe; các hệ thống đề xuất sách hoặc album nhạc trên cơ sở những lần mua và đánh giá trước đó của một người dùng; và những hệ thống ra quyết định y tế giúp các bác sĩ chẩn đoán ung thư vú, khuyến nghị phác đồ điều trị và hỗ trợ đọc điện tâm đồ. Chúng ta có những robot thú cưng, robot vệ sinh, robot cắt cỏ, robot giải cứu, robot phẫu thuật và hơn một triệu robot công nghiệp.[64] Số lượng robot thế giới đã vượt quá 10 triệu con.[65]

Nhận dạng giọng nói hiện đại dựa trên những kỹ thuật thống kê như các mô hình Markov ẩn đã đủ chính xác để áp dụng trong thực tiễn (một số nội dung trong cuốn sách này được soạn thảo với sự trợ giúp của một chương trình nhận dạng giọng nói). Các trợ lý cá nhân

kỹ thuật số, như Siri của Apple, phản hồi với mệnh lệnh bằng giọng nói và có thể trả lời những câu hỏi đơn giản và thực hiện các lệnh. Kỹ thuật nhận dạng ký tự quang học đối với các văn bản viết tay và đánh máy thường xuyên được sử dụng trong các ứng dụng như phân loại thư tín và số hóa các tài liệu cũ.[66]

Dịch máy vẫn chưa được hoàn hảo nhưng đã đủ tốt để có nhiều ứng dụng. Các hệ thống ban đầu sử dụng cách tiếp cận GOFAI với quy tắc ngữ pháp được mã hóa cứng, và những quy tắc này phải do các nhà ngôn ngữ học có kỹ năng phát triển ngay từ đầu cho từng ngôn ngữ. Các hệ thống mới hơn sử dụng những kỹ thuật học máy thống kê có thể tự động xây dựng các mô hình thống kê từ những thói quen sử dụng quan sát được. Máy luận ra các tham số cho những mô hình này bằng cách phân tích ngữ liệu song ngữ. Cách tiếp cận này không cần sự trợ giúp của các nhà ngôn ngữ học: các lập trình viên xây dựng những hệ thống này thậm chí không nói thứ ngôn ngữ mà họ đang xử lý.[67]

Trong những năm qua, công nghệ nhận dạng khuôn mặt đã cải thiện đến mức có thể sử dụng để phục vụ việc nhập cảnh tự động ở châu Âu và Australia. Bộ Ngoại giao Mỹ vận hành một hệ thống nhận dạng khuôn mặt với hơn 75 triệu tấm ảnh để cấp thị thực. Các hệ thống giám sát sử dụng các công nghệ AI và khai thác dữ liệu ngày càng phức tạp để phân tích giọng nói, video hoặc văn bản. Một lượng lớn dữ liệu này được thu thập từ các phương tiện truyền thông thế giới và được lưu trữ trong những trung tâm dữ liệu khổng lồ.

Kỹ thuật chứng minh định lý và giải phương trình giờ đây đã phát triển đến mức gần như không được coi như AI nữa. Các bộ giải phương trình đã được đưa vào các chương trình tính toán khoa học như Mathematica. Các phương pháp xác minh chính thống, bao gồm các bộ chứng minh định lý tự động thường xuyên được các nhà sản xuất chip sử dụng để xác minh hành vi của các thiết kế mạch trước khi đưa vào sản xuất.

Các cơ sở tình báo và quân đội Mỹ đã đi đầu trong việc triển khai quy mô lớn các robot rà phá bom mìn, thiết bị bay giám sát và tấn công không người lái, cùng các phương tiện không người lái khác. Những phương tiện này chủ yếu vẫn dựa vào điều khiển từ xa, tuy nhiên, người ta cũng đang nghiên cứu để tăng cường năng lực tự chủ của chúng.

Lập thời gian biểu thông minh là một trong những lĩnh vực thành công quan trọng. Công cụ lập kế hoạch và thời gian biểu tự động DART phục vụ hậu cần được sử dụng trong Chiến dịch Bão sa mạc năm 1991 hiệu quả tới mức DARPA (Cơ quan Dự án nghiên cứu quốc phòng tiên tiến của Mỹ) tuyên bố rằng chỉ ứng dụng duy nhất này đã đủ hoàn vốn toàn bộ đầu tư trong suốt 30 năm của họ vào AI.[68] Các hệ thống đặt vé máy bay sử dụng các hệ thống lập thời gian biểu và định giá phức tạp. Các doanh nghiệp cũng ứng dụng rộng rãi kỹ thuật AI vào các hệ thống kiểm soát kho hàng. Họ cũng dùng các hệ thống đặt hàng tự động và trợ giúp qua điện thoại, được kết nối với phần mềm nhận dạng giọng nói để hướng dẫn những khách hàng kém may mắn vượt qua một mê lộ các menu đan lồng phức tạp.

Công nghệ AI đứng sau nhiều dịch vụ Internet. Các phần mềm kiểm soát lưu lượng email toàn cầu, và mặc dù những kẻ phát tán email spam liên tục thích ứng để tránh né các biện pháp phòng chống, nhưng các bộ lọc spam Bayes cơ bản đã có thể cầm chân được các cơn bão spam. Phần mềm sử dụng các cấu phần AI chịu trách nhiệm tự động phê duyệt hay từ chối các giao dịch thẻ tín dụng và thường xuyên giám sát hoạt động của tài khoản để phát hiện dấu hiệu rửa tiền. Công nghệ học máy cũng được các hệ thống truy hồi thông tin sử dụng rộng rãi. Chúng ta có thể nói rằng công cụ tìm kiếm Google là hệ thống AI vĩ đại nhất từng được xây dựng.

Giờ đây, ta phải nhấn mạnh rằng ranh giới giữa trí tuệ nhân tạo và phần mềm nói chung đã không còn rõ nét. Một số ứng dụng được liệt kê ở trên có thể được xem như các ứng dụng phần mềm chung thay vì AI cụ thể, mặc dù điều đó đưa chúng ta quay lại với tuyên bố của

McCarthy, rằng khi thứ gì đó làm được việc thì nó sẽ không còn là AI nữa. Một cách phân biệt khác phù hợp hơn với mục tiêu của chúng ta là giữa những hệ thống có năng lực nhận thức hẹp (bất kể chúng có được gọi là "AI" hay không) và những hệ thống giải quyết vấn đề có ứng dụng tổng quát hơn. Về cơ bản tất cả các hệ thống hiện đang được sử dụng thuộc về loại thứ nhất: loại hẹp. Tuy nhiên nhiều hệ thống trong số đó có những cấu phần có thể đóng vai trò nhất định trong trí tuệ nhân tạo tương lai hoặc có thể phục vụ cho sự phát triển của ngành này, chẳng hạn như bộ phân loại, các thuật toán tìm kiếm, bộ lập kế hoạch, bộ giải toán và các khung biểu diễn.

Một môi trường vô cùng quan trọng và siêu cạnh tranh mà trong đó, AI hiện đang vận hành là thị trường tài chính toàn cầu. Những hệ thống giao dịch chứng khoán tự động đang được các công ty đầu tư lớn sử dụng rộng rãi. Trong khi một số chỉ là những phương cách đơn giản để tự động hóa việc thực thi các lệnh mua bán cụ thể, thì những hệ thống khác lại có những chiến lược giao dịch phức tạp thích ứng theo điều kiện thay đổi của thị trường. Các hệ thống phân tích sử dụng một số kỹ thật khai thác dữ liệu và phân tích chuỗi thời gian để tìm kiếm quy luật và xu hướng thị trường chứng khoán, hoặc liên kết những biến động giá trong quá khứ với các biến bên ngoài, như những từ khóa trong các bản tin thời sự. Các nhà cung cấp tin tức tài chính bán các bản tin được định dạng đặc biệt để những chương trình AI như vậy sử dụng. Những hệ thống khác thì chuyên tìm kiếm các cơ hội lướt sóng bên trong một thị trường hoặc giữa nhiều thị trường để tìm kiếm lợi nhuận từ những biến động giá nhất thời chỉ xảy ra trong vài mili giây. (Trong khoảng thời gian đó, độ trễ liên lạc trở nên đáng kể, kể cả đối với tín hiệu có tốc độ ánh sáng trong cáp quang, khiến cho việc đặt máy tính gần sàn giao dịch là một lợi thế.) Những chương trình giao dịch thuật toán cao tần chiếm tới hơn một nửa lượng cổ phiếu được giao dịch trên thị trường Mỹ.[69] Hoạt động giao dịch thuật toán cũng có dính líu tới vụ sụp đổ chớp nhoáng thị trường năm 2010 (2010 Flash Crash) (xem Khung 2).

**Khung 2 Vụ Flash Crash năm 2010**

Đến chiều ngày 6/5/2010, thị trường chứng khoán Mỹ đã giảm 4% do những lo lắng về khủng hoảng nợ châu Âu. Vào 2h32 chiều, một người bán lớn (một tổ hợp quỹ tương hỗ) đã khởi động một thuật toán bán để bán đi một số lượng lớn hợp đồng tương lai E-Mini S&P 500 với giá bán được liên kết với một chỉ số đo thanh khoản từng phút trên thị trường chứng khoán. Những hợp đồng này được mua thông qua chương trình giao dịch thuật toán cao tần, được lập trình để nhanh chóng loại bỏ thế giá lên (long position) tạm thời bằng cách bán các hợp đồng cho các chương trình giao dịch thuật toán khác. Khi nhu cầu từ của những người mua cơ bản suy yếu, trình giao dịch thuật toán bắt đầu bán E-Mini cho các trình giao dịch thuật toán khác (sơ cấp), và các trình này lại tiếp tục bán cho các trình giao dịch thuật toán khác nữa (thứ cấp), tạo ra một hiệu ứng "củ khoai nóng" (hot potato) đẩy khối lượng giao dịch lên cao – điều đó được thuật toán bán diễn dịch như một chỉ số cho thấy tính thanh khoản cao, thúc đẩy nó tăng tốc độ đưa các hợp đồng E-Mini ra thị trường, tạo ra một xoáy ốc đi xuống. Đến một thời điểm nào đó, các chương trình giao dịch thuật toán cao tần bắt đầu rút khỏi thị trường, làm giảm tính thanh khoản trong lúc giá tiếp tục giảm. Đến 2h45, giao dịch E-Mini bị tạm dừng bởi một bộ ngắt mạch tự động, chức năng chặn logic (stop logic functionality) của sàn giao dịch. Khi giao dịch được khôi phục, chỉ năm giây sau đó, giá ổn định lại và sau đó bắt đầu bù đắp được phần lớn thiệt hại. Nhưng trong một khoảng thời gian, dưới đáy của cuộc khủng hoảng, thị trường mất một ngàn tỷ đô la, và hiệu ứng "tràn" đã dẫn đến một số lượng đáng kể giao dịch chứng khoán riêng lẻ được thực hiện với mức giá "lố bịch", như 1 cent hay 100.000 đô la. Sau khi thị trường đóng cửa giao dịch ngày hôm đó, đại diện của các sàn giao dịch đã gặp các nhà quản lý và quyết định hủy mọi giao dịch được thực hiện với mức giá chênh lệch 60% hoặc cao hơn so với mức trước khủng hoảng (coi những giao dịch này là "sai lỗi rõ ràng" và do đó sẽ bị huỷ bỏ *hồi tố* (post facto cancellation) theo các quy tắc giao dịch hiện hành.[70]

Ở đây, câu chuyện này có đôi chút lạc đề, vì những chương trình máy tính có liên quan trong Flash Crash không thực sự thông minh hay phức tạp, và

kiểu đe dọa chúng tạo ra cơ bản khác với những quan ngại mà chúng tôi nêu ra trong phần sau của cuốn sách này về tiềm năng của siêu trí tuệ máy. Mặc dù vậy, những sự kiện này cho ta một số bài học hữu ích. Thứ nhất, nó nhắc nhở rằng các tương tác giữa những cấu phần riêng rẽ đơn giản (như thuật toán bán và các chương trình giao dịch thuật toán cao tần) có thể tạo ra những hiệu ứng phức tạp và bất ngờ. Rủi ro hệ thống có thể tích tụ trong một hệ thống khi những phần tử mới được đưa vào, những rủi ro chưa thực sự rõ ràng cho tới khi có chuyện không đúng xảy ra (và đôi khi không cần đến lúc đó).[71]

Một bài học khác là các chuyên gia thông minh có thể ra lệnh cho một chương trình trên cơ sở một giả định có vẻ có nghĩa và thông thường là đúng (ví dụ, khối lượng giao dịch là một số đo tốt của tính thanh khoản thị trường), và điều đó có thể tạo ra những kết quả thảm họa khi chương trình tiếp tục tuân lệnh với sự nhất quán logic cứng nhắc ngay cả trong tình huống bất thường khi giả định không còn đúng. Thuật toán chỉ làm những gì nó thường làm; và trừ khi có một loại thuật toán rất đặc biệt, nó sẽ không quan tâm chúng ta có ôm đầu tuyệt vọng hay hổn hển vì hoảng sợ vì sự không thích hợp tới mức lố bịch của những hành động của nó. Đó là một chủ đề mà chúng ta sẽ còn gặp lại.

Quan sát thứ ba liên quan đến Flash Crash là trong khi tự động hóa có đóng góp vào biến cố thì nó cũng đóng góp vào việc giải quyết. Logic lệnh dừng được lập trình trước – tạm ngừng giao dịch khi giá thay đổi quá nhiều – đã được thiết lập để tự thực thi, bởi lẽ người ta đã dự đoán đúng rằng các sự kiện kích hoạt có thể xảy ra với tỷ lệ thời gian quá nhanh để con người có thể phản hồi. Nhu cầu đối với chức năng an toàn được cài đặt sẵn và tự động thực thi – thay vì dựa vào sự giám sát của con người – một lần nữa lại báo trước về một chủ đề quan trọng trong cuộc thảo luận của chúng ra về siêu trí tuệ máy.[72]

## Những quan điểm về tương lai của trí tuệ máy

Tiến bộ trên hai mặt trận chính – một mặt hướng tới một nền tảng thống kê và lý thuyết thông tin vững chắc hơn cho học máy, và mặt khác hướng tới thành công thực tiễn và thương mại của các ứng dụng đặc thù ứng dụng hoặc đặc thù miền – đã trả lại cho nghiên cứu AI một số đặc quyền đã mất. Tuy nhiên, hiệu ứng văn hóa của lịch sử trước đó lên cộng đồng AI có thể vẫn còn tồn dư và khiến nhiều nhà nghiên cứu

chính thống không muốn điều chỉnh lại mình theo những tham vọng quá mức. Do đó Nils Nilsson, một trong những người kỳ cựu trong lĩnh vực này, than phiền rằng những đồng nghiệp hiện nay thiếu sự dũng cảm về tinh thần đã từng thúc đẩy thế hệ của ông:

> Tôi nghĩ rằng những quan ngại về "uy tín" đã có hiệu ứng gây trì trệ lên một số nhà nghiên cứu AI. Tôi nghe họ nói những điều như, "AI từng bị phê phán vì sự màu mè. Giờ đây khi chúng ta có được những tiến bộ chắc chắn, đừng liều lĩnh đánh mất uy tín của mình." Một trong những kết quả của sự bảo thủ này là sự tập trung nhiều hơn vào "AI yếu" (weak AI) – lối tiếp cận chuyên cung cấp trợ giúp cho tư duy con người – và tránh xa "AI mạnh" (strong AI) – lối tiếp cận tìm cách cơ khí hóa trí tuệ cấp độ con người.[73]

Cảm nghĩ của Nilsson được một số nhà sáng lập khác hưởng ứng, trong đó có Marvin Minsky, John McCarthy, và Patrick Winston.[74]

Những năm vừa qua chứng kiến sự hồi sinh của mối quan tâm đối với AI, điều có thể làm nảy sinh những nỗ lực mới hướng tới trí tuệ nhân tạo tổng thể (cái mà Nilsson gọi là "AI mạnh"). Ngoài phần cứng nhanh hơn, một dự án đương thời cũng sẽ hưởng lợi từ những tiến bộ lớn trong nhiều lĩnh vực phụ trợ của AI, trong thiết kế phần mềm nói chung và trong các lĩnh vực lân cận như khoa học thần kinh điện toán. Một yếu tố cho thấy nhu cầu bức thiết về chất lượng thông tin và giáo dục được thể hiện trong phản hồi đối với thông tin liên quan đến một khóa học trực tuyến miễn phí về đại cương trí tuệ nhân tạo tại Đại học Standford mùa thu năm 2011 do Sebastian Thrun và Peter Norvig tổ chức. Khoảng 160.000 sinh viên trên khắp thế giới đã đăng ký tham gia (và 23.000 đã hoàn thành khóa học).[75]

Ý kiến của các chuyên gia về tương lai của AI rất đa chiều. Có một sự không đồng thuận về thang thời gian cũng như dạng thức cuối cùng của AI. Theo một nghiên cứu mới đây, những dự báo về phát triển tương lai của trí tuệ nhân tạo "đa dạng đến đâu thì đáng tin cậy đến đó".[76]

Mặc dù sự phân bổ của niềm tin hiện thời chưa được đo đếm cẩn thận, chúng ta vẫn có thể ước lượng thô từ những khảo sát nhỏ lẻ và

quan sát không chính thức. Cụ thể, một loạt khảo sát gần đây đã hỏi ý kiến thành viên của một vài cộng đồng chuyên gia về thời điểm mà họ kỳ vọng rằng "trí tuệ máy cấp độ con người" (HLMI) sẽ được phát triển. Trí tuệ máy này được định nghĩa là "thứ có thể thực hiện được hầu hết công việc của con người, với hiệu năng ít nhất cũng ngang bằng một người điển hình."[77] Kết quả của các khảo sát này được đưa ra trong Bảng 2. Phân tích mẫu tổng hợp cho ra đánh giá (trung bình) như sau: xác suất có HLMI vào năm 2022 là 10%, xác suất có HLMI năm 2040 là 50% và xác suất có HLMI năm 2040 là 90%. (Những người trả lời được yêu cầu đặt ra ước tính dựa trên giả định rằng "hoạt động khoa học của con người tiếp tục diễn ra mà không gặp phải gián đoạn tiêu cực lớn".)

Những con số này cần được chấp nhận với đôi chút hoài nghi: kích thước mẫu tương đối nhỏ và không hẳn đại diện cho tầng lớp chuyên gia nói chung. Tuy nhiên, chúng nhất quán với kết quả từ những khảo sát khác.[78]

Các kết quả khảo sát cũng tương đồng với một số cuộc phỏng vấn mới được công bố gần đây với khoảng hơn 20 nhà nghiên cứu trong các lĩnh vực liên quan đến AI. Chẳng hạn, Nils Nilsson đã có một sự nghiệp lâu dài và hiệu quả dành cho những vấn đề về tìm kiếm, lập kế hoạch, biểu diễn tri thức và người máy. Ông đã viết sách giáo khoa về trí tuệ nhân tạo, và gần đây đã hoàn thành bộ sử toàn diện nhất từng được viết cho tới nay về lĩnh vực này.[79] Khi được hỏi về ngày "đổ bộ" của HLMI, ông đưa ra những ý kiến sau:[80]

10% khả năng: 2030

50% khả năng: 2050

90% khả năng: 2100

**Bảng 2** ***Khi nào chúng ta có được trí tuệ máy cấp độ con người?***[81]

| | 10% | 50% | 90% |
|---|---|---|---|
| PT-AI | 2023 | 2048 | 2080 |
| AGI | 2022 | 2040 | 2065 |
| EETN | 2020 | 2050 | 2093 |
| TOP100 | 2024 | 2050 | 2070 |
| Tổng hợp | 2022 | 2040 | 2075 |

Dựa trên các bản ghi phỏng vấn đã công bố, phân bố xác suất của giáo sư Nilsson xem ra mang tính đại diện cho khá nhiều chuyên gia trong lĩnh vực, mặc dù vẫn cần phải nhấn mạnh rằng có sự khác biệt rất lớn về quan điểm: trong giới chuyên môn có nhiều người cuồng nhiệt hơn và kỳ vọng một cách tự tin rằng HLMI sẽ xuất hiện từ 2020 – 2040; cũng có những người khác tự tin rằng chúng ta sẽ còn lâu hoặc chẳng bao giờ có được thứ trí tuệ máy đó.[82] Ngoài ra, một số người được phỏng vấn cảm thấy khái niệm trí tuệ nhân tạo "cấp độ con người" được định nghĩa tệ hại hoặc sai lầm, hoặc vì những lý do khác không muốn đưa ra một dự báo định lượng.

Theo quan điểm của riêng tôi, những con số trung bình đưa ra trong khảo sát chuyên gia chưa có đủ khối lượng xác suất về thời gian xuất hiện muộn hơn. Xác suất 10% HLMI chưa được phát triển vào những năm 2075 hoặc thậm chí 2100 (sau khi đã điều chỉnh theo điều kiện "hoạt động khoa học của loài người tiếp tục diễn ra mà không gặp phải gián đoạn tiêu cực lớn") xem ra quá thấp.

Về mặt lịch sử, các nhà nghiên cứu AI chưa từng có thành tựu nổi bật về khả năng dự đoán tốc độ phát triển trong chính lĩnh vực của họ hay bóng dáng tương lai của những phát triển đó. Một mặt, một số nhiệm vụ như chơi cờ vua hóa ra lại có thể thực hiện được bằng những chương trình đơn giản đến ngạc nhiên; và những kẻ ác khẩu từng nói rằng máy móc sẽ "không bao giờ" làm được việc này hay việc kia đã

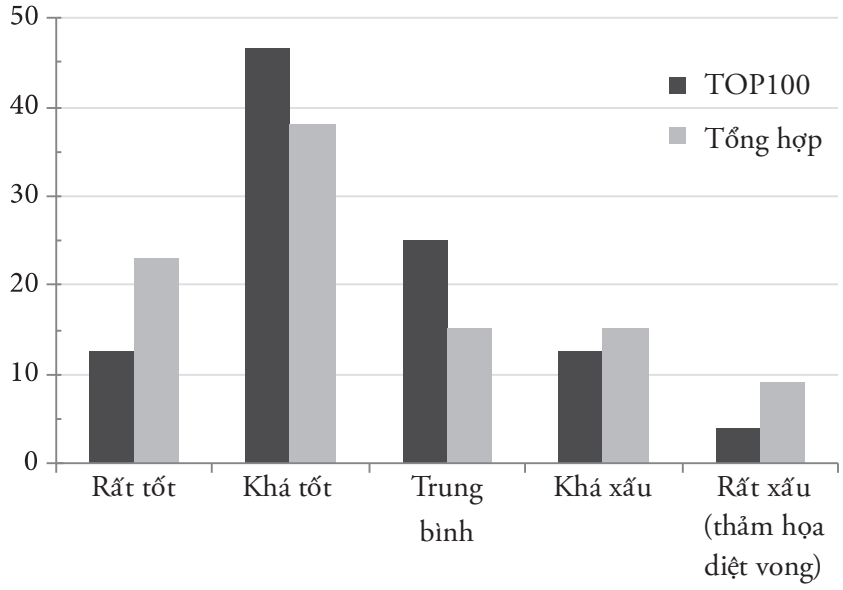
thường xuyên bị chứng minh là sai lầm. Mặt khác, những lỗi lầm tiêu biểu hơn của giới chuyên môn cho thấy khó khăn trong việc làm cho một hệ thống thực hiện được một cách chắc chắn những nhiệm vụ thực tế vẫn đang bị đánh giá thấp, và những tiến bộ của một vài dự án hay kỹ thuật ưa thích cụ thể của họ đang được đánh giá quá cao.

Khảo sát trên cũng đưa ra hai câu hỏi khác có liên quan đến nghiên cứu của chúng tôi. Một trong số đó yêu cầu người trả lời đưa ra suy nghĩ của mình về việc còn bao lâu nữa cho tới khi có được siêu trí tuệ, giả sử máy móc có trí tuệ cấp độ con người đã được chế tạo. Kết quả được đưa ra trong Bảng 3.

Một câu khác hỏi về suy nghĩ của người trả lời đối với tác động dài hạn lên loài người khi chế tạo ra trí tuệ máy cấp độ con người. Tổng hợp các câu trả lời được đưa ra trên Hình 2.

**Bảng 3** ***Mất bao lâu để đi từ trí tuệ cấp độ con người tới siêu trí tuệ?***

| | Dưới hai năm từ khi có HLMI | Dưới 30 năm từ khi có HLMI |
|---|---|---|
| TOP100 | 5% | 50% |
| Tổng hợp | 10% | 75% |




**Hình 2** Tác động tổng thể lâu dài của HLMI.[83]

Quan điểm riêng của tôi về mặt nào đó có sự khác biệt với những quan điểm được đưa ra trong khảo sát. Tôi gắn xác suất cao hơn cho việc siêu trí tuệ được tạo ra ngay sau khi có trí tuệ máy cấp độ con người. Tôi cũng có tầm nhìn phân cực hơn về hậu quả, cho rằng kết quả rất tốt hoặc rất xấu có thể dễ xảy ra hơn so với một kết quả tương đối hơn. Lý do sẽ được đưa ra sau trong cuốn sách này.

Kích thước mẫu bé, thiên kiến lựa chọn và trên hết là sự không đáng tin cậy cố hữu của các quan điểm chủ quan đồng nghĩa với việc ta không nên quá trông cậy vào những khảo sát và phỏng vấn chuyên gia này. Chúng không cho phép chúng ta đưa ra bất cứ kết luận mạnh mẽ nào, nhưng cũng đã gợi ý về một kết luận chưa chắc chắn. Chúng khuyến nghị (ít nhất thay cho những dữ liệu hay phân tích tốt hơn) rằng có thể có lý do để tin vào việc trí tuệ máy cấp độ con người may ra thì sẽ xuất hiện vào giữa thế kỷ này, và có rất ít khả năng để nó được phát triển sớm hay muộn hơn; rằng có thể không lâu sau đó nó sẽ dẫn đến siêu trí tuệ; và rằng nhiều kết quả khác nhau có khả năng xảy ra đáng kể, bao gồm cả những kết quả rất tốt hoặc rất xấu, xấu đến mức có thể loài người sẽ tuyệt chủng.[84] Ít nhất chúng cũng khuyến nghị rằng chủ đề này xứng đáng được xem xét kỹ lưỡng hơn.